# Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

# THE INTERPRETATION OF INFRARED AND RAMAN SPECTRA USING PATTERN RECOGNITION

by

John Martin Comerford

A dissertation

presented in partial fulfillment of

the requirements for the degree

of

DOCTOR OF ENGINEERING SCIENCE IN CHEMICAL ENGINEERING

AT

NEWARK COLLEGE OF ENGINEERING



Newark, New Jersey
1974

## ABSTRACT

An automatic classification of chemical compounds by computer processing of digitized spectral data is presented. The classification system is based on a branch of artificial intelligence known as supervised learning, and used binary linear classifiers to identify compounds as alcohols, esters, ethers, ketones or compounds containing double bonds.

Each of the 1117 spectra in volume one of Sadtler's Standard Raman Spectra was coded using a scale from 0 to 9 in the range from 4000 to 200 $cm^{-1}$. One hundred and twelve readings were taken on each spectrum. These data were then examined using pattern recognition techniques, and several methods of combining infrared and Raman data from the same compound were tested.

The classification techniques were most successful when applied to concatenated infrared and Raman data. By taking the infrared data in the range from 500 to 1900 $cm^{-1}$ and concatenating them with Raman data from the same range, a vector representing each of the 400 compounds in the data set was obtained. The parallel polarized Raman was used in preference to the perpendicularly polarized spectrum when both were available, otherwise the nonpolarized spectrum was used. Using an iterative pattern recognition technique a

vector was then calculated which would recognize compounds as members of a class or not members based on the sign of the dot product of the calculated vector and the vector representing the compound.

When vectors were calculated using only half the data set, then tested for their ability to correctly classify the remaining compounds; it was found that they could correctly classify compounds more than 90% of the time.

Vectors which were trained using the entire data set were helpful in determining characteristic group frequencies and each class treated is discussed.

Several compounds for which observed frequencies have been assigned in the literature were tested to determine if the assignments could be supported by the trained vectors. In nearly every case supporting evidence for the assignment was available from the appropriate trained vector.

APPROVAL OF DISSERTATION

THE INTERPRETATION OF INFRARED AND RAMAN SPECTRA

USING

PATTERN RECOGNITION

BY

JOHN MARTIN COMERFORD

FOR

DEPARTMENT OF CHEMICAL ENGINEERING AND CHEMISTRY

NEWARK COLLEGE OF ENGINEERING

BY

FACULTY COMMITTEE

APPROVED:

Howard S. Kimmel, Chairman

Peter A. Anderson

Howard D. Perlmutter

Edward C. Roche

William H. Snyder

Newark, New Jersey
September, 1974

ACKNOWLEDGEMENT

The author wishes to express his sincere appreciation to Dr. Howard Kimmel for his guidance and support throughout the work and through the years of graduate study.

I also wish to thank Dr. William H. Snyder for the many useful discussions and particularly for his encouragement through my graduate work.

Dr. Edward C. Roche, Dr. Howard D. Perlmuther and Dr. Peter G. Anderson are each thanked for serving on the advisory committee and for helping to shape this work.

Finally, I find it difficult to pen an adequate expression of my love and appreciation for my wife, Marie, who typed the manuscript, took care of our four children and loved and encouraged when I needed it most.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

## LIST OF TABLES

## LIST OF FIGURES

LIST OF FIGURES (continued)

# INTRODUCTION

## 1. Background

Artificial intelligence is the name given to the developing body of knowledge concerning computer techniques for handling large bodies of data or attacking very complex problems. The techniques are as varied as the problems and the body of knowledge is ill-defined. Generally the techniques are ad hoc to the problem at hand. One branch of artifical intelligence, however, which has been developing at a rapid pace and which does have universal application is the area known as pattern recognition. In Nagy's[1] review of the state of the art in pattern recognition (1968) he conceded at the outset that pattern recognition was hardly a discipline in its own right, and at best was a collection of highly varied problems. In his review, however, he managed to outline the skeleton of what he viewed as a highly amorphous structure. In 1972 in an effort to introduce pattern recognition to the chemical literature Kowalski and Bender[2] offered the functional breakdown of pattern recognition which has been reproduced in Figure 1.

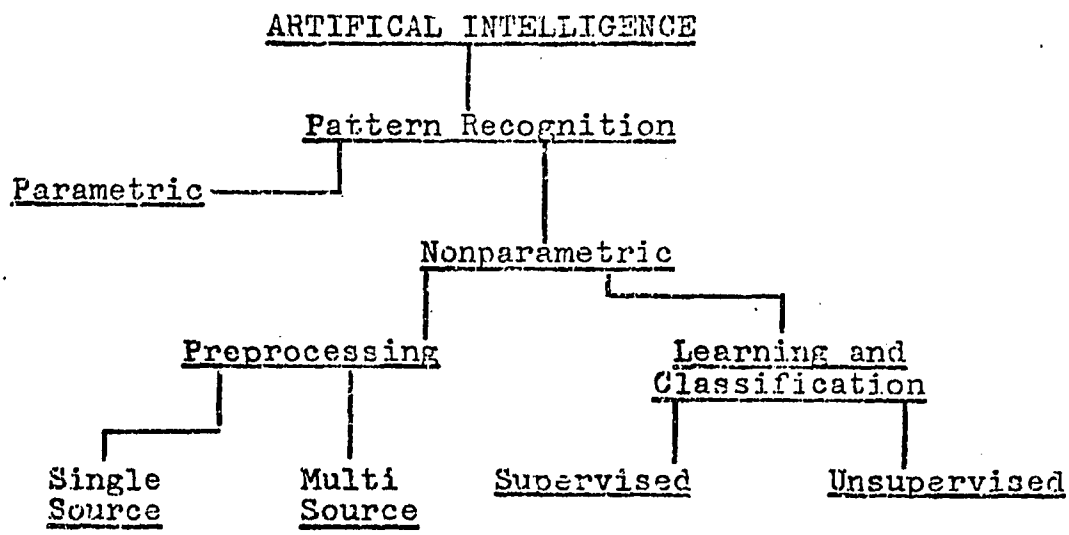
---

(1) G. Nagy, Proc. IEEE, No. 5, 56, 836 (1968).

(2) B. R. Kowalski and C. F. Bender, J. Am. Chem. Soc., 94, 5632 (1972).

FIGURE 1

Functional Breakdown of Pattern Recognition




A general statement of the problem which pattern recognition seeks to resolve is: "Given a set of objects and a list of measurements made on these objects, is it possible to find and/or predict a property of the object that is not directly measured but is related to the measurements via some unknown relationship?" In most of the early work in the field it is assumed that a statistically representative data set is available for designing or training the pattern recognition apparatus. This leads to parametric methods which recognize patterns based on the underlying probability distributions of the data. Thus far these techniques have not proven useful with chemical data and they are not discussed in this work.

Nonparametric techniques usually attempt to develop a linear categorizer which assigns objects to category one or two based upon the sign of the weighted sum of measurements made on the object. Let the set of measurements made on any one object be represented by the vector $\vec{f}$, amd let the linear categorizer be represented by the vector $\vec{W}$. Both vectors have n members. Then:

$$\vec{f} = (f(1), f(2), \ldots f(n))$$

and

$$\vec{W} = (W(1), W(2), \ldots W(n))$$

The linear categorizer assigns an unknown vector to class one if $\sum_{i=1}^{m} W(i) \cdot f(i) > 0$, otherwise to class two.

From the beginning we will discuss only two-class problems. In principle, however, any multiclass problem can be treated as a number of two class problems involving the separation of each class from the remainder of the universe. A discussion of optimal assignments in multiclass problems can be found in the article by Braverman[3] or the thesis by Kiessling[4]. Jurs et al[5] also treated this topic.

---

(3) E. M. Braverman, Automat. i. Telemekh, 23, 349 (1962).

(4) C. Kiessling, M. S. Thesis, Cornell University, 1965.

(5) P. C. Jurs, B. R. Kowalski, T. L. Isenhour and C. N. Reilley, Anal. Chem., No. 6, 41, 695, (1969).

Preprocessing simply refers to the necessary manipulation of the data which must be done in certain cases to render the problem amenable to an orderly solution. It is referred to variously as preprocessing, filtering or prefiltering, feature or measurement extraction or dimensionality reduction. Some form of preprocessing is frequently needed when the data is taken from more than one source, for example when infrared and mass spectra were used to classify chemical compounds[6]. Occasionally data from a single source, such as nuclear magnetic resonance spectral data[7] must be preprocessed.

Learning and classification is supervised if a set of objects whose classes are known is used to develop the weight vector $\vec{W}$. In unsupervised learning the data is clustered or classified into groups without a priori knowledge as to what defines a group. For example the classification of archaeological artifacts by applying pattern recognition to trace element data[8] was done using unsupervised learning. We will concern ourselves with supervised learning only.

---

(6) P. C. Jurs, B. R. Kowalski, T. L. Isenhour and C. N. Reilley, ibid., No. 14, 41, 1949 (1969).

(7) B. R. Kowalski and C. A. Reilly, J. Phys. Chem., No. 10, 75, 1402 (1971).

(8) B. R. Kowalski, T. F. Schatzki and F. H. Stross, Anal. Chem., No. 13, 44, 2176 (1972).

## 2. Application of Pattern Recognition

Pattern recognition has application in computer assisted medical diagnosis and treatment, drug interaction studies, neurobiological signal processing, etc.[9] Clinical data such as electrocardiograms and electroencephalograms have been analyzed using pattern recognition techniques[1, 10]. Manning[11] studied teleseismic event classification in order to differentiate nuclear explosions from earthquakes on the basis of the characteristics of the wave which travels through the earth's mantle. Problems arising from satellite photography for weather, earth-resource and even the sensing for life on remote planets[12, 13, 14, 15] have been attacked using pattern-recognition. And these diverse applications only touch the surface of the hundreds of areas where pattern recognition is being used.

---

(9) E. A. Patrick, Fundamenals of Pattern Recognition, Prentice Hall, Engle Cliffs, New Jersey, 1972.

(10) Special Issue on Technology and Health Services, Proc. IEEE, No. 11, 57, (1969).

(11) J. E. Manning, M. C. Grignetti and P. R. Miles, Rept. 1372, Bolt, Beranek and Newman, Cambridge, Mass. 1966.

(12) Y. C. Ho and A. K. Agrqwala, Proc. IEEE, No. 12, 56, 2101 (1968).

(13) L. M. Uhr, Pattern Recognition: Theory, Simulations, and Dynamic Models of Form Perception and Discovery, Wiley, New York, 1966.

(14) G. C. Cheng, J. Y. Pollock, R. S. Ladley and A. Rosenfeld, eds., Pictorial Pattern Recognition Process, Thompson Book Co., Washington, D. C., 1968.

(15) L. Kanal, ed., Pattern Recognition, Thompson Book Co., Washington, D. C., 1968.

In the late sixties papers began appearing in the chemical literature, and now many types of chemical data have been explored using pattern recognition. Sybrandt and Perone[16] evaluated a pattern recognition technique for qualitative analysis of mixtures by stationary electrode polarography. Molecular formula determination from low resolution mass spectrametry was investigated by Jurs, Kowalski and Isenhour[17, 18] and Kowalski and Reilly[7] considered nuclear magnetic resonance spectral interpretation by pattern recognition. Wangen and Isenhour[19] applied pattern recognition methods to the semiquantitative determination of seventeen light elements by resolution of 14-Mev neutron induced gamma ray spectra. Infrared data were first treated by Kowalski, Jurs, Isenhour and Reilly[20] in 1969 then again in 1973 by Liddell and Jurs[21]. Melting and boiling point data[6, 22] have also been investigated.

---

(16) L. B. Sybrandt and S. P. Perone, Anal. Chem., No. 5, 43, 382 (1971).

(17) P. C. Jurs, B. R. Kowalski and T. L. Isenhour, Anal. Chem., No. 1, 41, 21 (1969).

(18) P. C. Jurs, B. R. Kowalski, T. L. Isenhour and C. N. Reilley, Anal. Chem., No. 6, 41, 690 (1969).

(19) L. E. Wagnen and T. L. Isenhour, Anal. Chem., No. 7, 42, 737 (1970).

(20) B. R. Kowalski, P. C. Jurs, T. L. Isenhour, and C. N. Reilley, Anal. Chem., No. 14, 41, 1945 (1969).

(21) R. W. Liddell and P. C. Jurs, Appl. Spec., No. 5, 27, 371 (1973).

(22) T. L. Isenhour and P. C. Jurs, Anal. Chem., No. 10, 43, 20A (1971).

## 3. The Problem

In late 1972 the Sadtler Reasearch Laboratories announced the publication of a collection of infrared and Raman spectra. It was decided that the first volume of this collection would be used as the data base to investigate the use of pattern recognition in interpreting infrared and Raman spectra. While earlier work was done with infrared spectra[20, 21], no work has yet been reported in which Raman data were investigated using pattern recognition. In the first work done with infrared data[20] the coding of the data was crude (a 0, 1, 2, or 3 was used to represent spectral peaks), and the thrust of the first attempt was simply to explore the feasability of the approach. Using these data these first workers were able to achieve about 75% recognition of various chemical classes (acids, esters, amids, etc.) and thus demonstrated the feasability of using the approach. In general our recognition rate was slightly above 90%.

Liddell and Jurs[21] improved their coding system and by doing so were able to improve their recognition rate to approximately 87%. The bulk of this second look at infrared data, however, concentrated on attempting by various empirical methods to improve the recognition rate and reduce the number of features (i.e. elements in each vector) used in the decision making process.

None of the earlier workers attempted to use the trained vectors as an aid in interpreting infrared spectra. The unique relationship between infrared and Raman makes these data particularly interesting for this purpose. Additionally, the study of Raman data with pattern recognition techniques provides an orderly approach to the development of spectra-structure correlations. The rapid advance of Raman spectroscopy since the development of the laser has outpaced the development of correlation tables; and workers frequently refer to infrared correlation tables, when Raman data are being analyzed since Raman correlation tables are not available. The differences between the selection rules of infrared and Raman point out the hazards of this practice.

With this background in mind the objectives of this investigation were: (1) to use the vectors trained with infrared and Raman data and combinations of those data to distinguish characteristic group frequencies, and (2) to investigate the predictive ability of the trained vectors.

## 4. The Data

Volume One of the Sadtler Standard Raman Spectra includes 400 organic compounds (including water) selected with the intention of providing simple compounds with representative functional groups at the beginning of the publication. When liquid compounds are presented both the

parallel polarized and the perpendicularly polarized Raman spectra are given. When solids are presented only the nonpolarized Raman is given. There are no gases in the collection. The infrared spectrum of each compound is given. Among the 400 compounds in Volume One there are 317 liquids and 83 solid compounds. Thus there are a total of 1117 spectra in the data bank for this work.

The Raman spectra in this volume were obtained using the Cary 83 instrument which utilizes the 4880 Å argon line. The laser power was a nominal 100 mW at the laser head and 30-60 mW at the sample. The source used in this volume of spectra was a Coherent Radiation model 59B Argon Ion laser and the detector was a bi-alkali cathode type photodetector. Sample scatterning was always detected at right angles to the direction of the incident radiation.

Coding Procedure

1. A base line is drawn which connects the two points of least intensity in the spectrum and which runs along the base of the spectrum (the top of the IR prints; the bottom of the Raman prints).

2. The largest peak in the spectum is assigned an intenstiy of 9.

3. The spectrum is divided into 112 bands. From 4000 $cm^{-1}$ to 2000 $cm^{-1}$ the band width is 50 $cm^{-1}$. From 2000 $cm^{-1}$ to 200 $cm^{-1}$ the band width is 25 $cm^{-1}$.

4. Within each band the intensity is measured on a scale from 0 to 9 depending on the relative intensity compared to the strongest peak in the spectrum.

5. When the intensity is measured within a band, the strongest intensity within the band is used as the position at which the measurement is made.

To aid in measuring intensities a system based on the one suggested by Durkin et al.[23] was used. A template similar to the one shown in Figure 2 was drawn. Copies were xeroxed on Albanene tracing paper (Keuffel and Esser Co. #10 5351). By sliding the template across the strongest peak until the perpendicular distance from the base line of the spectrum to the top of the peak just spans the perpendicular distance on the template from the base to

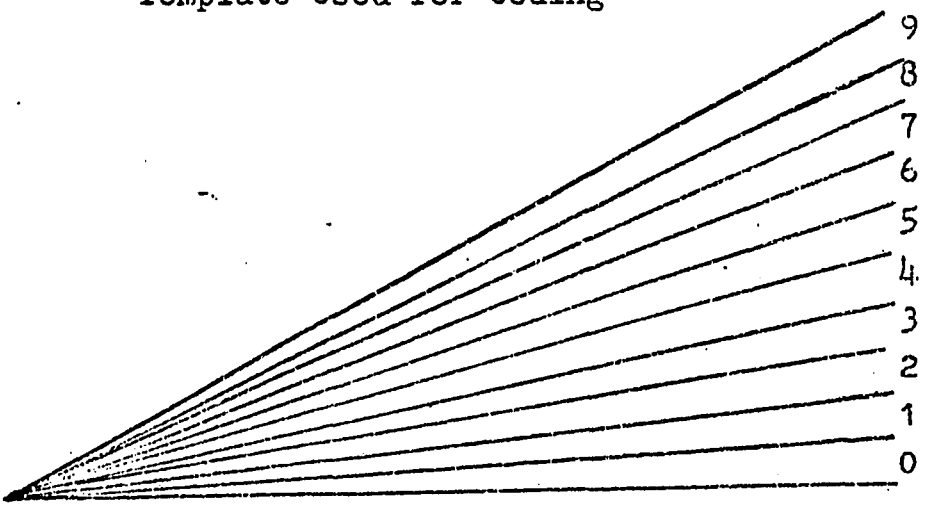
---

(23) T. Durkin, L. DeHayes and J. Glore, J. Chem. Ed., 48, 452 (1971).

the top line, one is able to quickly find a 0 to 9 scale appropriate for the spectrum being coded. It was convient to draw a line on the template and then fold it on the line. The templates were disguarded after a few uses. We found that this system was superior to the plexiglass type of arrangement suggested in the literature.

FIGURE 2

Template Used for Coding

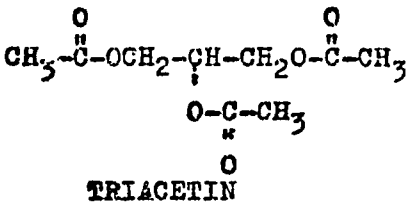



When measuring the intensities it is important to measure the strongest intensity within each band (step 5), rather than to make the measurement precisely at a particular frequency. Several trials with different types of coding suggested that this system gives the most faithful reproduction of the actual spectrum when manually digitalizing the spectrum. Figure 3 shows the spectrum of a typical

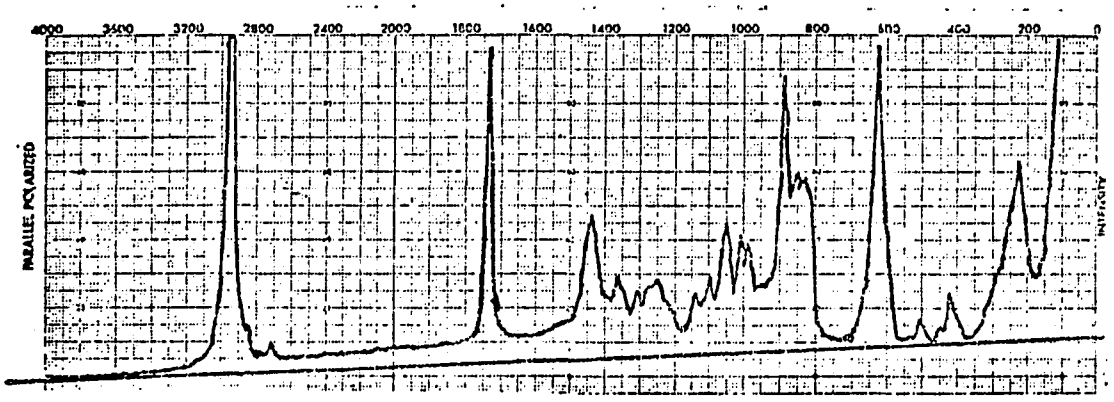
compound, the digital code and a plot of the code vs wave number. It is clear that some detail is lost by the coding process, but most of the peak defination is retained.

FIGURE 3

Coding Illustration



TRIACETIN



| | |
|---|---|
| 0000000000 | 52 |
| 0000000012 | 0110001235 |
| 9921000000 | 0138830000 |
| 0000000000 | 2157554100 |
| 0000000049 | 1111223333 |
| 4100000111 | 3431211222 |

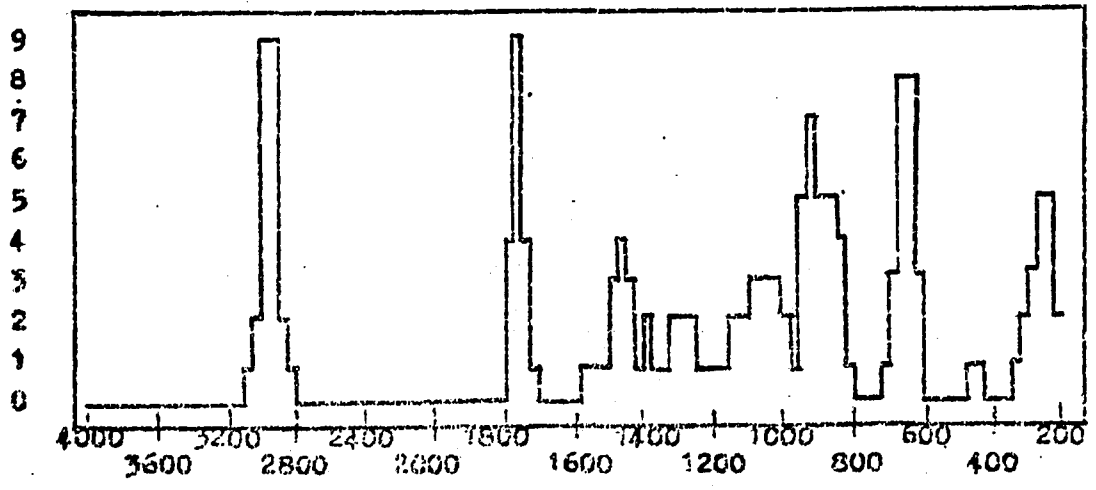


Table I shows the 17 classes of compounds in the data, and gives the number of compounds in each class. The classes which have been treated in this work are noted with an asterisk. There are two ortho-esters in the data, and they have been treated as ethers. Since there are only 4 tertiary alcohols in the data, these were not treated seperately. The classes are not mutually exclusive, and there are 21 compounds which do not belong to any of the classes included in Table I. Appendix A gives the data bank and lists the classes to which each compound belongs.

TABLE I

Classes of Compounds in Data Set

| Description of Class | Number in Class |
|---|---|
| 1. esters | 94* |
| 2. nitrogen containing compounds | 84 |
| 3. primary alcohols | 32* |
| 4. secondary alcohols | 12* |
| 5. tertiary alcohols | 4 |
| 6. ethers and ortho esters | 49* |
| 7. chlorine containing compounds | 47 |
| 8. fluorine containing compounds | 1 |
| 9. bromine containing compounds | 29 |
| 10. iodine containing compounds | 4 |
| 11. aromatic compounds (containing benzene ring) | 157 |
| 12. compounds containing a saturated hexane ring | 8 |
| 13. compounds containing C=C bond (not benzenes) | 38* |
| 14. compounds containing a triple bond | 3 |
| 15. ketones | 15* |
| 16. compounds containing a phenol group | 13 |
| 17. compounds containing sulfur | 28 |

(* indicates those classes treated in this work)

## INFRARED and RAMAN SPECTROSCOPY

### 1. Prologue to the Discussion

The purpose of this section is to briefly outline the differences between infrared and Raman spectra and to acquaint the unitiated reader with the selection rules which govern vibrational spectra. The ideas presented here will not be developed, but rather will be stated in as concise a form as seems practicable. These concepts have been thoroughly discussed elsewhere, and for an extensive and authoritative treatment the reader should consult the classical works: Infrared and Raman Spectra of Polyatomic Molecules by Herzberg[24] and Molecular Vibrations by Wilson, Decius and Cross[25]. Supplementally the more recent works by Woodward[26] and Steele[27] both provide an excellent introduction to the topic.

---

(24) G. Herzberg, Infrared and Raman Spectra, D. Van Nostrand Co., Princeton, New Jersey, 1956.

(25) E. B. Wilson,Jr., J. C. Decius, P. C. Cross, Molecular Vibrations, McGraw Hill Book Co., New York, 1955.

(26) L. A. Woodward, Introduction to the Theory of Molecular Vibrational Spectroscopy, Clarendon Press, Oxford, Englend, 1972.

(27) D. Steele, The Theory of Vibrational Spectroscopy, W. B. Saunders Co., Philadelphia, Pa., 1971.

## 2. Infrared and Raman - Different Techniques

Classical electrodynamics teaches that a system emits radiation by virtue of periodic changes in its electric dipole moment, the frequency of the emitted radiation being the same as that of the dipole oscillations. Absorption is the inverse of emission, and the system is able to absorb radiation of the same frequency as it is able to emit. Infrared spectroscopy is concerned with the absorption of radiation incident upon a sample and is thus a phenomenon which is dependent on the behavior of the vibrating dipole moment, $\mu$, within a molecule. When monochromatic radiation passes through a homogenous material some of the radiation is scattered. Most of the scattered radiation is unaltered in frequency but a portion suffers a frequency shift. This shift results from transference of internal vibrational energy (and also rotational and electronic energy) from the beam to the sample material and vice versa. This effect is known as Raman scattering. It is a molecular light-scattering phenomenon in which a change of frequency occurs in a small portion of the incident radiation. It does not involve absorption at all, and the intrinsic dipole moment $\mu$ of the molecular vibration is of no consequence. Unlike infrared the Raman effect is concerned with the induced dipole moment $\pi$ which is induced in the molecule by the electric field of the incident light. For field strengths of the magnitude ordinarily used in Raman

spectroscopy the relation between $\pi$ and E, the applied electric field, may be written

$$\pi = \alpha E$$

where $\alpha$ is the electric polarizability of the molecule.

Thus while infrared and Raman spectroscopy are techniques for investigating the vibrational characteristics of molecules, the underlying principles for each method are different. In principle, the observed frequencies in each technique (i.e. the observed shifts in the Raman and the observed absorption in the infrared) are the same provided they are allowed by both techniques. Vibrational transitions that are forbidden (not allowed) in infrared absorption may be permitted in the Raman effect, and vice versa. Frequently, however, the same frequencies are observed by both techniques. The two experimental methods are thus essentially complementary in character.

## 3. Selection Rules

The rules or guidelines by which one predicts that a vibration will or will not be allowed using either infrared or Raman spectroscopy are called selection rules. Because the mechanisms of the two techniques are different the selection rules are different. The development of these rules, however is similar in each case.

Three distinct approaches to deriving the selection rules exist. The first is through classical mechanics, the second is through quantum mechanics and the third is through symmetry properties and group theory. The first two methods lead to what are called restricted selection rules because they are restricted by the assumptions used in their derivation. The development based on group theory leads to the general selection rules. While it is outside of the scope of this work to review in detail these approaches, a few comments on each method may help the reader appreciate the meaning and limitations of these rules.

## 4. Classical Approach

The electric dipole $\mu$ in a molecule is a vector, and therefore has three components $\mu_x$, $\mu_y$, and $\mu_z$ in a Cartesian system. The molecule will only be able to absorb radiation of frequency $\nu$ provided that $\mu$ (i,e, at least one of its three components) can occillate with this frequency. Now the dipole moment is a function of the nuclear configuration and so, when the molecule vibrates, the dipole varies correspondingly. In the simple harmonic approximation, all molecular vibrations can be regarded as superpositions of a limited number of fundamental modes, each with its own fundamental frequency $\nu_k$. It follows that the electric dipole moment can only oscillate with these fundamental frequencies, and only radiation of these fre-

quencies can be absorbed.

In general the magnitudes of the components of the molecular dipole moment will be functions of all the vibrational coordinates Q, and thus capable of expansion as a Taylor series:

$$\mu_x = (\mu_x)_0 + \sum_k \left\{ \left( \frac{\delta \mu_x}{\delta Q_k} \right)_0 Q_k \right\} + \text{higher terms}$$

$$\mu_y = (\mu_y)_0 + \sum_k \left\{ \left( \frac{\delta \mu_y}{\delta Q_k} \right)_0 Q_k \right\} + \text{higher terms}$$

$$\mu_z = (\mu_z)_0 + \sum_k \left\{ \left( \frac{\delta \mu_z}{\delta Q_k} \right)_0 Q_k \right\} + \text{higher terms}$$

The zero subscripts in the above equations indicate values at the equilibrium configuration of the molecule. We adopt the general convention whereby the three seperate expressions for $\mu_x$, $\mu_y$, and $\mu_z$ are all implied by the single condensed form

$$\mu = \mu_0 + \sum_k \left( \frac{\delta \mu}{\delta Q_k} \right)_0 Q_k$$

The condition that the molecular dipole moment shall be able to oscillate with the frequency $\nu_k$, i.e. the condition that this normal frequency shall be capable of being absorbed is that $(\delta \mu / \delta Q_k)_0$ shall not be zero. This implies that

$$\left( \frac{\delta \mu_i}{\delta Q_k} \right)_0 \neq 0$$

for at least one of the components (i= x, y, z).

This is a statement of the restricted selection rule for infrared absorption. Its derivation is dependent upon two special approximate assumptions. The first is that the molecular vibrations are simple-harmonic, for otherwise the normal modes would not be separable and the meaning of the individual normal coordinates $Q_k$ would be lost. The second assumption is that in the Taylor expansion of the electric dipole moment all the higher terms are negligible.

Similarly, the corresponding selection rule for Raman scattering may be derived. Like the dipole moment the electric polarizability of a molecule $\alpha$ will in general be a function of all the normal vibrational coordinates. We may therefore expand $\alpha$ as a Taylor series with respect to these coordinates and neglect powers higher than the first. We thus obtain

$$\alpha = \alpha_0 + \sum \left\{ \left( \frac{\delta \alpha}{\delta Q_k} \right)_0 Q_k \right\}$$

where $\alpha_0$ is the polarizability in the equilibrium configuration of the molecule. Since, as we have seen, the <u>induced</u> dipole moment, $\pi$, is given by

$$\pi = \alpha E$$

we may write

$$\pi = \alpha_0 E + \left\{ \sum_k \left\{ \left( \frac{\delta \alpha}{\delta Q_k} \right)_0 Q_k \right\} \right\} E$$

Consider the first term, $\alpha$ E, in the expression for $\pi$. Since every component of $\alpha_o$ is simply a molecular constant and every component of E oscillates with the incident light frequency $\nu_o$, it follows that the corresponding part of every component of $\pi$ must oscillate with this same frequency. Thus light of the incident frequency $\nu_o$ will be emitted and will be observable in directions which differ from that of the incident light. This is the phenomenon known as classical or Rayleigh scattering. Since it is of no interest in this discussion we can ignore this term.

Considering the second term, let us fix attention on the contribution from the particular vibrational mode with the normal coordinate $Q_k$. Every component of $(\delta\alpha/\delta Q_k)_0$ is simply a constant. The time dependent factors are $Q_k$ which oscillates with the normal frequencies $\nu_k$, and all the components of E, which oscillate with k the incident frequency $\nu_o$. These time dependences could be expressed by including the respective factors $\cos(2\pi\nu_k t)$ and $\cos(2\pi\nu_o t)$. In view of the identity $\cos A \cos B = \frac{1}{2}(\cos(A+B) + \cos(A-B))$ we see that all the corresponding contributions to all the components of the induced dipole moment $\pi$ are characterized by the two new frequencies $(\nu_o + \nu_k)$ and $(\nu_o - \nu_k)$. These frequencies are referred to as anti-Stokes and Stokes frequencies respectively, and they constitute the contribution of the kth

normal mode to the Raman spectrum of the scattering molecule.

The condition that a particular normal frequency $\nu_k$ shall be active in Raman scattering is that the factor $(\delta\alpha/\delta Q_k)_0$ shall be different from zero, i.e.

$$\left(\frac{\delta \alpha_{ij}}{\delta Q_k}\right)_0 \neq 0$$

for at least one of the components (i or j = x, y or z) of the polarizability. This statement is the restricted selection rule for Raman scattering and is similar to the selection rule for infrared. As was true with the infrared rule, the Raman rule is subject to the limitation that the vibrations are simple-harmonic and that in the Taylor expansion of the induced dipole moment all higher terms may be neglected.

5. Quantum Mechanical Approach

For the transition between the two states characterized by the wave functions $\Psi^n$ and $\Psi^m$ we denote the wave mechanical quantity known as the transition moment by $\mu_{nm}$. The transition moment is defined by the following equations:

$$(\mu_x)_{nm} = \int \Psi^n \mu_x \Psi^m \quad d\tau$$

$$(\mu_y)_{nm} = \int \Psi^n \mu_y \Psi^m \quad d\tau$$

$$(\mu_z)_{nm} = \int \Psi^n \mu_z \Psi^m \quad d\tau$$

in which $\mu_x$, $\mu_y$, and $\mu_z$ are the magnitudes of the components of $\mu$, and $d\tau$ is a volume element. As is generally done we condense the three expressions into the conventional equation

$$\mu_{nm} = \int \Psi^n \mu \Psi^m \, d\tau$$

The importance of the transition moment is that it determines the intensity of the absorption of radiation by the transition in question. In fact the relation of this intensity to the magnitude $\mu_{nm}$ of the transition moment is similar to that of the intensity of classical absorption to the amplitude of oscillation of an ordinary dipole moment. Thus the total intensity is proportional to the square of $\mu_{nm}$, or more precisely, the total intensity is proportional to the sum of the squares of $(\mu_x)_{nm}$, $(\mu_y)_{nm}$ and $(\mu_z)_{nm}$. Thus a very general statement of the selection rule is that a transition is forbidden in the infrared if $\mu_{nm} = 0$, i.e. if $(\mu_x)_{nm} = (\mu_y)_{nm} = (\mu_z)_{nm} = 0$.

A more particular statement of the selection rule can be derived by substituting the Taylor expansion value of $\mu$ into the equation for the transition moment. This gives

$$\mu_{nm} = \mu_o \int \Psi^n \Psi^m \, d\tau + \sum_k \left\{ \left( \frac{\partial \mu}{\partial Q_k} \right)_o \int \Psi^n Q_k \Psi^m \, d\tau \right\}$$

Because of the mutual orthogonality of the wave functions, the first integral of the right hand of the equation is zero unless n=m. This, however, corresponds to no transition, and may be neglected where absorption is concerned. Thus, as was seen in the classical development, one necessary condition for infrared absorption is that $(\frac{\delta \mu}{\delta Q_k})_o \neq 0$.

The quantum mechanical treatment of Raman scattering is similar to the treatment of absorption. The transition moment arising from the induced dipole moment $\pi$ is given by

$$\int \Psi^n \pi \Psi^m \, d\tau = E \int \Psi^n \alpha \Psi^m \, d\tau$$

$$= E \alpha_o \int \Psi^n \Psi^m \, d\tau + E \sum_k \left\{ (\frac{\delta \alpha}{\delta Q_k})_o \Psi^n Q_k \Psi^m \, d\tau \right\}$$

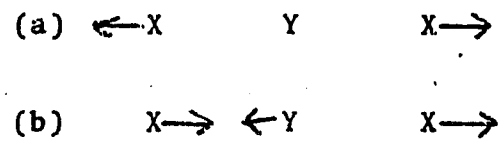
Because of the mutual orthogonality of the eigenfunctions, the integral in the first term on the right hand side vanishes unless $\Psi^n = \Psi^m$, in which case its value is unity. This term accounts for the Rayleigh scattering without change of frequency. When $n \neq m$ one necessary condition for Raman scattering is that $(\frac{\delta \alpha}{\delta Q_k})_o \neq 0$. As occurs in the classical development of selection rules the quantum mechanical treatment relies on the restrictive assumptions that the molecular vibrations are simple-harmonic, and that in the Taylor expansion higher terms may be neglected.

## 6. Group Theoretical Approach

It is the molecular symmetry that determines whether or not $\delta\alpha/\delta Q_k \neq 0$ or $\delta\mu/\delta Q_k \neq 0$ for a particular vibrational mode. This may be illustrated by the very simple example of a linear Y-X-Y molecule in which the two bonds are identical. Its two modes of vibration along the line of the masses are shown in Figure 4.

FIGURE 4

Forms of the Normal Vibrational Modes
Along the Molecular Axis
(Linear X-Y-X Molecule)

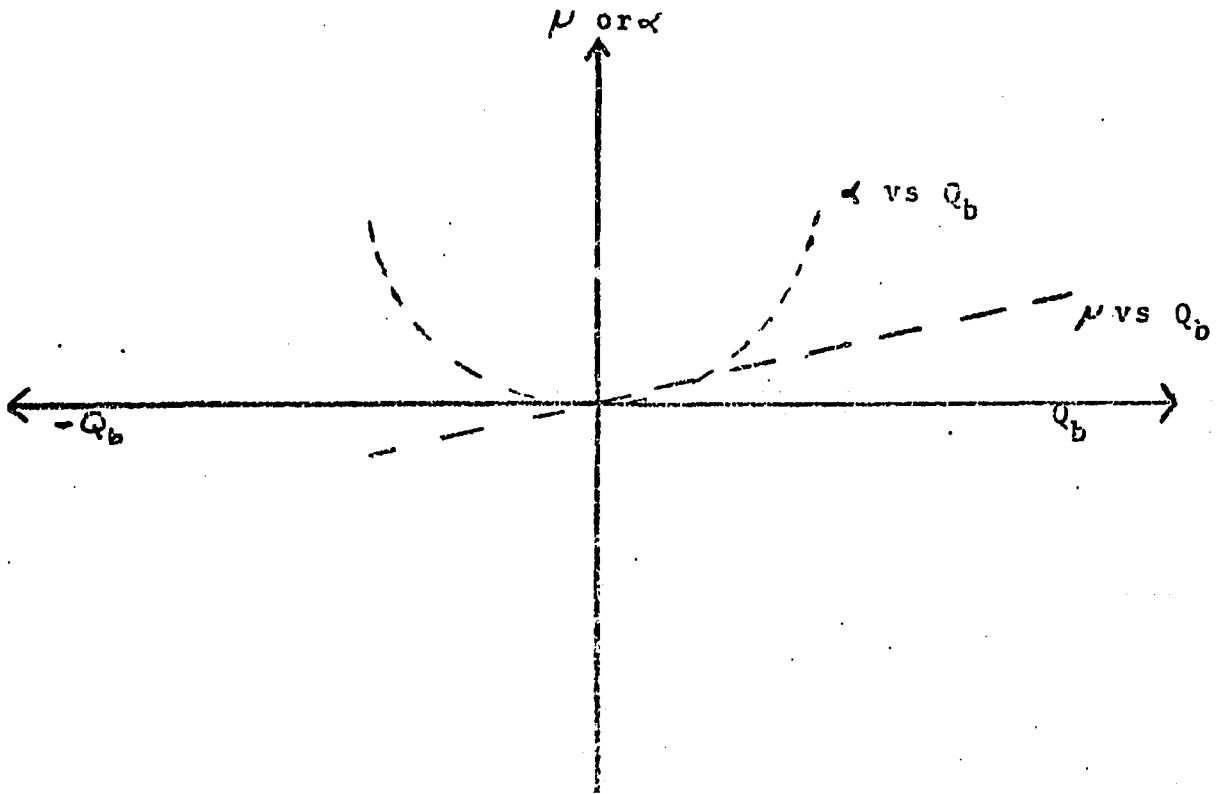



Remembering that the molecule is composed of positively charged nuclei and negatively charged electrons, we see that in its equilibrium configuration its structural symmetry will carry with it a corresponding electrical symmetry and will result in a zero molecular dipole. In the totally symmetric vibration ( a in Figure 4) this symmetry is conserved throughout. The dipole moment therefore remains unchanged. This means that $(\delta\mu/\delta Q_k)_0 = 0$,

i.e. that the mode must be forbidden in infrared absorption. In the antisymmetric mode (b), on the other hand, the molecule does not conserve its equilibrium symmetry. The structural asymmetry through which it passes in the course of a vibration causes a corresponding electrical asymmetry, i.e. produces a non-zero dipole moment which oscillates synchronously with the mechanical vibration, passing through values of equal magnitude but opposite sign at corresponding values of the normal coordinate $Q_b$. This is shown diagramatically in Figure 5, where the abscissa represents the normal coordinate and the ordinate represents the magnitude of the dipole moment.

FIGURE 5

Variation of Dipole Moment $\mu$ or Polarizability $\alpha$ with Vibration along the Normal Coordinate



In this simple case the moment is always directed along the line of the nuclei, which may be called the x direction. Thus we need consider the component $\mu_x$ only, the other two ( $\mu_y$ and $\mu_z$) being always zero. From Figure 5 we see that $(\delta\mu/\delta Q_b)_0 \neq 0$. The mode is therefore active in infrared absorption.

Similarly it is molecular symmetry that determines whether or not $(\delta\alpha/\delta Q_b)_0 \neq 0$, i.e. whether a normal vibrational mode shall be active in Raman scattering. Consider again the modes a and b of Figure 4 in the totally symmetric mode a, both are stretched in one phase of a vibrational cycle and both are compressed in the other. Although this leaves the molecular dipole unaltered, it is clear that the electrical polarizability will be different in the two phases. In general therefore $(\delta\alpha/\delta Q_a)_0$ will differ from zero. In mode b, on the other hand, the structural situation is the same in both phases, in that one bond is stretched and the other is compressed. The end-for-end interchange caused a reversal in the sign of the non-zero electric dipole moment vector, but there is no such reversal for the components of the electrical polarizability. Thus the value of $\alpha$ is the same for $+Q_b$ as for $-Q_b$. This is shown diagrammatically in Figure 5. It is true that the value of any polarizability component at each extreme value of the normal coordinate may be different from its value at

$Q_b = 0$, but from Figure 5 it is clear that the slope at the equilibrium configuration, i.e. $(\delta\alpha/\delta Q_b)_0$, is zero. Thus mode b is forbidden in Raman scattering. In this very simple example it happens that the mode which is forbidden in infrared is permitted in Raman and vice versa. This, however, is a special circumstance for in general many molecular vibrational modes are permitted in both kinds of spectra, and some are forbidden in both. The example illustrates the importance of symmetry in the context of the restricted selection rules. Generalization of the symmetry considerations in the form of group theory makes it possible to predict the spectroscopic consequences of the application of the rules to molecules containing any number of nuclei and belonging to any point group. In addition it should be recognized that symmetry theory is independent of the particular physical nature of the system to which it is applied. The relevant symmetry properties are independent of whether the intramolecular force field is simple harmonic or anharmonic and no special assumptions about Taylor expansions in terms of normal vibrational coorinates are needed.

While a development of the general selection rules based on group theory is well outside the intention of this section, we will simply conclude this section with

Woodward's[26] formal statement of the general selection rules:

"A transition between the states characterized by the wave function $\Psi^n$ and $\Psi^m$ is forbidden in the infrared absorption unless for at least one of the components $\mu_i$ of the molecular electric dipole moment (i=x, y, or z) the product $\Psi^n \mu_i \Psi^m$ belongs to a representation whose structure contains the totally symmetric species.

"A transition between states characterized by the wave functions $\Psi^n$ and $\Psi^m$ is forbidden in Raman scattering unless for at least one of the components $\alpha_{ij}$ of the molecular polarizability tensor (i or j = x, y or z) the product $\Psi^n \alpha_{ij} \Psi^m$ belongs to a representation whose structure contains the totally symmetric species."

# PROOF OF CONVERGENCE

## 1. Terminology

In developing the proof of convergence for the algorithm used in the main computer program, we have availed ourselves of the symbolism common in texts on linear algebra. For a review of this topic and a further development of the principles involved, the reader is directed to the text by Curtis[28]. Table II should suffice, however, for complete understanding of those symbols used in this work.

TABLE II

Notation Used in Proof

| Notation | Meaning |
|---|---|
| $a \in B$ | a is a member of set B |
| $A \cap B$ | denotes the intersection of set A and set B |
| $\emptyset$ | represents an empty set |
| $A \cap B = \emptyset$ | A and B are disjoint sets; no member of A is in set B and no member of B is in A. |
| $\longrightarrow$ | implies |
| $A \subset B$ | set A is a subset of set B |

(28) C. W. Curtis, Linera Algebra, An Intorductory Approach, Allyn and Bacon, Boston Mass., 1968.

It is convenient to think of the data set as a set of vectors, F. Each vector in the set represents one vibrational spectrum and is denoted $\vec{f}$. The set F is composed of two disjoint sets $F^+$ and $F^-$; i.e. $F^+ \cap F^- = \emptyset$, and $F^+$, $F^- \subset F$. Furthermore we denote the members of $F^+$ as $\vec{f}^+$ and the members $F^-$ as $\vec{f}^-$.

To express these ideas in more familiar terms we can consider the problem of developing a means for distinguishing the esters in our data bank from all other spectra. The spectrum of an ester can be represented by $\vec{f}^+$ and the entire set of esters makes up the set $F^+$. All other spectra can be represented by $\vec{f}^-$ and the set of spectra which are not esters is called $F^-$. Our entire data bank is called F.

The task at hand is to develop a rule which will correctly classify any member of F so that all $\vec{f}^+$ are classified as $\vec{f}^+ \in F^+$ and all $\vec{f}^-$ as $\vec{f}^- \in F^-$. That is, we must develop a simple to use decision procedure (frequently called a predicate) by which we can select any spectrum and tell whether it is an ester or not an ester. A number of conventional classification methods which might be used to attack the problem are discussed in the literature. Nagy's review[1] is comprehensive and a large

number of other works are available[29, 30, 31, 32, 33]. These approaches, however, construct the predicate by a specific mathematical analysis of the entire data set F. Each approach is ad hoc to each predicate and suffers from the need to work on the entire data set at once. Such approaches are prohibitive when using large data sets.

In this chapter we will describe such a simple decision procedure based upon the sign of the weighted sum of the components of $\vec{f}$, and an algorithm in which a set of coefficients, $\vec{W}$, can be found by a systematic and easily mechanized procedure. The procedure begins with the selection of an arbitrary set of coefficients. A vector $\vec{f}$ is selected from F and the predicate $\vec{W}$ is tested to see if it works properly. If it does not, $\vec{f}$ is used to modify $\vec{W}$. If it does, $\vec{W}$ is left unchanged. A new $\vec{f}$ is then selected from F. The algorithm proceeds in this manner until $\vec{W}$ will correctly classify all $\vec{f}$ in F. Because $\vec{W}$ is modified every time it makes a mistake the procedure has been termed a "learning machine".

---

(29) M. Minsky and S. Papert, Perceptions, An Introduction to Computational Geometry, MIT Press, Lawrence, Mass., (1969).

(30) F. Rosenblatt, Principles of Neurodynamics, Spartan Books, New York, 1962.

(31) J. Ruben, Rept 39.014, IBM, New York Scientific Center, 1966.

(32) J. MacQueen, Proc. 5th Berkeley Symps. on Statistics and Probability, University of California Press, 281 (1967).

(33) N. J. Nilsson, Learning Machines, Mc Graw Hill, New York, 1965.

Before proceeding it might be helpful to the reader to define a few terms.

The magnitude of a vector $\vec{W}$ is represented by $|\vec{W}|$ where

$$|\vec{W}| = \sqrt{\sum_i (W(i))^2}$$

The dot product of 2 vectors $\vec{W}$ and $\vec{f}$ is given by

$$\vec{W} \cdot \vec{f} = \sum_i W(i) \times f(i)$$

Note that $|\vec{W}|^2 = \vec{W} \cdot \vec{W}$, i.e. the magnitude squared of a vector is the vector dotted on itself.

A unit vector $\vec{W}^*$ is the direction of $\vec{W}$, i.e. it is the vector divided by its magnitude.

$$\vec{W}^* = \frac{\vec{W}}{|\vec{W}|}$$

Provided, of course, $|\vec{W}| \neq 0$.

Note that the magnitude of a unit vector is 1, and

$$|\vec{W}^*|^2 = \vec{W}^* \cdot \vec{W}^* = 1$$

## 2. The Learning Machine

We will show how to pick a $\vec{W}$ which determines whether $\vec{f} \in F^+$ or $F^-$ by $\vec{W} \cdot \vec{f} > 0$ if and only if $\vec{f} \in F^+$. Intuitively one might suggest the simplest type of feedback or correction of $\vec{W}$ such as the following procedure.

START: Let $\vec{W}$ be any vector

TEST: Choose the next $\vec{f}$ from F (If F is exhausted without going to ADD or SUBTRACT, the procedure is done; else begin again with 1st $\vec{f}$.)

If $\vec{f} \in F^+$ and if $\vec{W}\cdot\vec{f} > 0$ Go to test

If $\vec{f} \in F^+$ and if $\vec{W}\cdot\vec{f} \leqslant 0$ Go to add

If $\vec{f} \in F^-$ and if $\vec{W}\cdot\vec{f} < 0$ Go to test

If $\vec{f} \in F^-$ and if $\vec{W}\cdot\vec{f} \geqslant 0$ Go to subtract

ADD Replace $\vec{W}$ by $\vec{W} + R_{\vec{f}}\,\vec{f}$

Go to TEST

SUBTRACT: Replace $\vec{W}$ by $\vec{W} - R_{\vec{f}}\,\vec{f}$

Go to TEST

$R_{\vec{f}}$ is a positive correction factor with an upper bound A, i.e. $0 < R_{\vec{f}} < A$. (A is some number chosen in advance.)

A priori, any procedure of this sort runs the risk of oscillating wildly. An adjustment of $\vec{W}$ for one vector $\vec{f}$ might undo the previous adjustment for another $\vec{f}$. As will be shown, however, this procedure will eventually converge to vector $\vec{W}$ which will correctly classify all $\vec{f}$ in F provided that F is linearly separable into $F^+$ and $F^-$, i.e. provided a hyper-plane exists which will separate $F^+$ from $F^-$. Remember $F^+ \cap F^- = \emptyset$ does not imply that $F^+$ and $F^-$ are linearly seperable, i. e. it is not a sufficient condition to guarantee that a solution $\vec{W}$ exists.

This simple feedback algorithm converges by the "perceptron convergence theorem."[12] The theorem states that

whatever choice is made in START and whatever function is used in TEST the vector $\vec{W}$ will be changed only a finite number of times, provided that F is linearly seperable into $F^+$ and $F^-$.

Before developing the proof of convergence one minor restriction must be placed on the members of any vector, $\vec{f}$. No element of $\vec{f}$ may be unbounded (i.e. $\pm\infty$). Thus for all $\vec{f} \in F$, $\vec{f}(i) \leq L$ , where L is some established upper bound. This restriction is necessary to the proof, but in practice poses no real limitation. In our data bank, for example, every element is in the range from 0 to 9.

Note that if a vector $\vec{W}$ seperates $F^+$ from $F^-$ then any positive multiple of W will also seperate $F^+$ from $F^-$ since

$$(a \cdot \vec{W}) \cdot \vec{f} = a \cdot (\vec{W} \cdot \vec{f})$$

From this it follows that we can find a unit vector as the $\vec{W}$.

The proof of convergence which we present is adapted from Minsky and Papert[29]. Minsky and Papert assumed that $|\vec{f}| = 1$ for all $\vec{f} \in F$, and $R_f = 1$. We make no such assumptions.

3. Proof

The program in the previous section is more complicated than is needed for the proof. Recognize that the following

is equivalent.

START: Let $\vec{W}$ be any vector

TEST: Choose the next $\vec{f}$ from F (Done when F is exhausted without going through ADD)

If $\vec{f} \in F^-$ change the sign of $\vec{f}$ by multiplying $\vec{f}$ by -1

If $\vec{W} \cdot \vec{f} > 0$, go to TEST, otherwise go to ADD

ADD: Replace $\vec{W}$ by $\vec{W} + R_{\vec{f}} \cdot \vec{f}$, where $R_{\vec{f}}$ is a positive weighting factor with an upper bound A

Go to TEST

Proof:

Define $G = \dfrac{\vec{W}^* \cdot \vec{W}}{|\vec{W}|}$

where $\vec{W}^*$ is a unit vector which is a solution, i.e.

$\vec{W}^* \cdot \vec{f} > 0 \longrightarrow \vec{f} \in F^+$

and $\vec{W}^* \cdot \vec{f} \leqslant 0 \longrightarrow \vec{f} \in F^-$

Because $|\vec{W}^*| = 1, |G| \leqslant 1.$

(It may help the reader to note that G is the cosine of the angle between $\vec{W}^*$ and $\vec{W}$.)

Consider the behavior of G on successive passes of the program through add. Considering first the numerator as each trial t progresses,

$$\begin{aligned} \vec{W}^* \cdot \vec{W}_{t+1} &= \vec{W}^* \cdot (\vec{W}_t + R_{\vec{f}}\, \vec{f}_t) \\ &= \vec{W}^* \cdot \vec{W}_t + \vec{W}^* \cdot R_{\vec{f}}\, \vec{f}_t \\ &\geqslant \vec{W}^* \cdot \vec{W}_t + \delta \end{aligned}$$

where $\delta > 0$

This holds because since $\vec{W}^*$ is a solution

$$\vec{W}^* \cdot R_{\vec{f}}\, \vec{f}_t > 0$$

(after the appropriate adjustment of the sign if $\vec{f} \in F^-$)

Thus after the nth application of add we obtain

$$\vec{W}^* \cdot \vec{W}_n \geqslant n\, \delta$$

The numerator of G increases with n, the number of changes of $\vec{W}$, that is the number of errors.

As for the denominator, since $\vec{W} \cdot R_{\vec{f}}\, \vec{f}$ must be negative or zero or the program would not have gone through add

$$\begin{aligned} |\vec{W}_{t+1}|^2 &= \vec{W}_{t+1} \cdot \vec{W}_{t+1} \\ &= (\vec{W}_t + R_{\vec{f}}\, \vec{f}_t) \cdot (\vec{W}_t + R_{\vec{f}}\, \vec{f}_t) \\ |\vec{W}_{t+1}|^2 &= |\vec{W}_t|^2 + 2\, \vec{W}_t \cdot R_{\vec{f}}\, \vec{f}_t + R^2_{\vec{f}} |\vec{f}_t|^2 \\ &< |\vec{W}_t|^2 + L^2 K^2 \end{aligned}$$

after the nth application of add

$$\begin{aligned} |\vec{W}_n|^2 &< n\, L^2 K^2 \\ |\vec{W}_n| &< n^{\frac{1}{2}} L K \end{aligned}$$

combining the results

$$G = \frac{\vec{W}^* \cdot \vec{W}_n}{\vec{W}_n} > \frac{n\, \delta}{n^{\frac{1}{2}} L K}$$

but $G \leqslant 1$ so this can only be so if

$$\frac{n_1 \delta}{n^{\frac{2}{3}} L K} \leqslant 1$$

That is

$$n \leqslant \frac{L^2 K^2}{\delta^2}$$

This completes the proof.

## 4. Variations

The proof presented in this work closely follows the computer program used in the body of this work. A number of minor variants in procedure are possible, however. For example $R_{\vec{f}}$ may be defined as $1/|\vec{f}|$ so that a unit vector is added each time a correction is made to $\vec{W}$. $R_{\vec{f}}$ may, of course, be a constant, i.e. independent of $\vec{f}$. In the program we used, $R_{\vec{f}}$ was defined in the way suggested by Agmon[34]:

$$R_{\vec{f}} = -k \left[ \frac{(\vec{W}_t \cdot \vec{f})}{|\vec{f}^2|} \right]$$

where k is a positive constant. In this manner the magnitude of $R_{\vec{f}}$ is dependent on how "badly" $\vec{W}_t$ failed to

---

(34) S. Agmon, Canadian J. Math., No. 3, 6, 382 (1954).

classify $\vec{f}$. Following the suggestion of Jurs, Kowalski, Isenhour and Reilley[18] we usually set k = 2, but a few experiments were done with other values.

It was pointed out earlier that if $\vec{W}$ is a solution vector, $\vec{W}'$ is also a solution vector where

$$\vec{W}' = k\,\vec{W}$$

k being a positive constant. Similarly it should be noted that if $\vec{W}$ is a solution and $\vec{W}'$ is a solution, then $\vec{W}''$ is also a solution where

$$\vec{W}'' = \vec{W} + \vec{W}'$$

Thus all solution vectors form a convex cone and the program will stop changing $\vec{W}$ as soon as it penetrates the cone. (Convex cone: a set S of vectors for which

(1) $A \in S \Longrightarrow kA \in S$ for all $k > 0$, and

(2) $A \in S$ and $B \in S \Longrightarrow (A+B) \in S$.)

Any solution defines a hyper-plane dividing the space in which we are operating. All $\vec{f}$ in $F^+$ are on one side of the plane while all $\vec{f}$ in $F^-$ are on the other side.

Notice, however, that if we are dealing with n dimensional space (i.e. $\vec{f}$ has n members) and our solution hyper-plane is limited to n members, the hyper-plane must pass through the origin of the space. If, however, we add one member to $\vec{f}$ and make that member +1, the solution hyper-

plane is not constrained to pass through the origin. Thus, while we take 112 readings from each spectrum, the dimensionality of the space we are working in is 113. The resulting vector passes through the origin of the 113 dimensional space but not through the origin of the 112 dimensional space.

## THE PROGRAMS

### 1. Program Names

Three programs were written in Fortran IV to analyze the spectral data. The first program, called TRAIN, trains vectors, the second program, SNOOP, is used for either of two jobs. SNOOP is used to average several trained vectors, or it is used to calculate the average spectrum of a class of compounds. Program TEST tests vectors which have been trained to determine how they perform with a new set of data. A listing of each program with a sample set of input and output for the program TRAIN is included in Appendix B. The comment cards in each program are intended to instruct the interested reader in the details necessary for use of each program.

### 2. TRAIN

Program TRAIN reads a set of options from a card file. It then reads the appropriate spectral data from a file. (The data have been previously stored in the file.) Figure 6 shows a simplied flow diagram for TRAIN and Table III lists the options which are currently available as a part of program TRAIN. In Figure 6 the weight vector being trained is labeled $\vec{W}$ and a vector representing a spectrum is labeled $\vec{f}$.

FIGURE 6

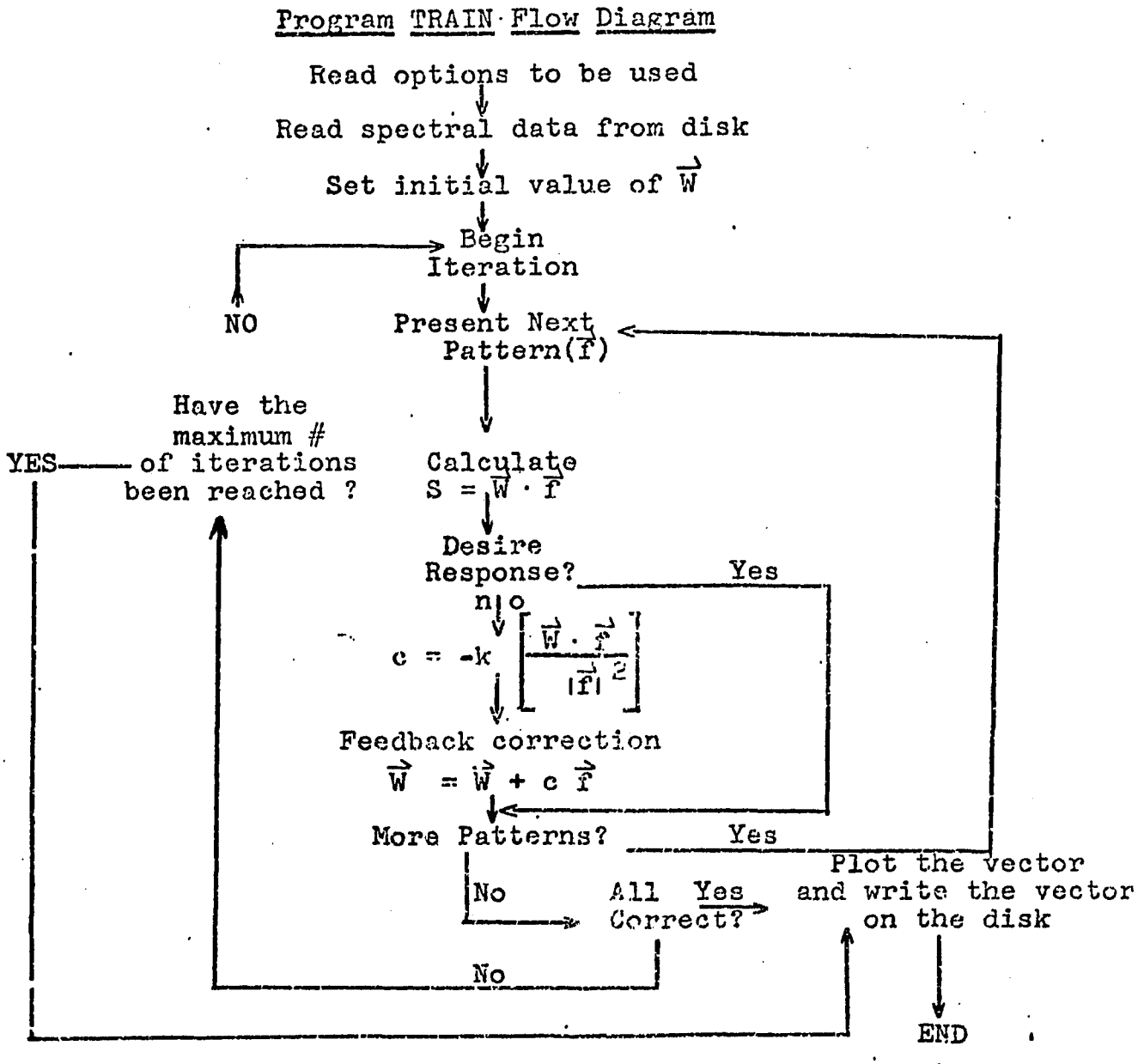

Program TRAIN Flow Diagram
Read options to be used
Read spectral data from disk
Set initial value of $\vec{W}$
Begin Iteration
NO
Present Next Pattern($\vec{f}$)
Have the maximum # of iterations been reached ?
YES
Calculate $S = \vec{W} \cdot \vec{f}$
Desire Response?
Yes
no
$c = -k \left[ \frac{\vec{W} \cdot \vec{f}}{|\vec{f}|^2} \right]$
Feedback correction
$\vec{W} = \vec{W} + c\,\vec{f}$
More Patterns?
Yes
No
All Correct?
Yes
No
Plot the vector and write the vector on the disk
END

TABLE III

OPTIONS IN PROGRAM TRAIN

| Option Number | Fortran Label | Description |
|---|---|---|
| 1 | NOPT1 | Type of spectra to be considered. For complete details see Table III. |
| 2 | NCARD | Number of spectra in the data file stored on disk. For all of the work reported here NCARD = 1117. |
| 3 | NOPT2 | Class of compounds being considered. For complete details see Table IV. |
| 4 | FIN | Initial value of all elements of $\vec{W}$, the vector being trained. If FIN = 0 the program will read a starting vector from disk file. (see Option 12) |
| 5 | NSHUT | Specifies an upper limit to the number of iterations. |
| 6 | LEN | The vector $\vec{W}$ will be printed every "LEN'th" iteration (and again at the end of the calculation). |
| 7 | NOPT3 | The number of spectra not in class (Option #3) which are to be included in the training deck. |
| 8 | NOPT4 | The number of spectra in the class (Option #3) which are to be included in the training set. |
| 9 | NOPT4 | If NOPT6 = 1 every element in the trailing vectors will be used in the training process, else every "NOPT6'th" element is used until the vector $\vec{W}$ is trained or until "NSHUT" iterations have been completed. Then $\vec{W}$ is used as the initial vector for training with every element. |
| 10 | NOPT5 | If NOPT5 $\neq$ 0 another set of data is expected. The necessary disk files will be rewound. |
| 11 | FKOR | A constant used to weight the correction factor. (FKOR is the value of k in Fig.IV) |
| 12 | NFILE | The position in the disk file of the initial vector. If FIN $\neq$ 0 NFILE should not be specified. |

The value of LEN (Option 6) determines how often the vector is printed during the training process. In the early experments it was useful to be able to watch the vector develop. In later work only the trained vector was needed; this was done by setting the value of LEN higher than NSHUT.

Options 8 and 9 are used to limit the number of compounds used to train the vector. In those experiments where all the available compounds were not used in the training process, the resulting vector was later used with program TEST to determine how well the vector classifies those compounds not used in the training set.

Table IV provides a detailed description of the type of spectra treated by TRAIN (option one).

TABLE IV

The Type of Spectra Considered

(Option One in Program TRAIN)

| Value of Option | Description |
|---|---|
| 1 | *Infrared spectra only are used |
| 2 | *Raman nonpolarized spectra only are used |
| 3 | *Raman parallel polarized spectra only are used |
| 4 | *Raman perpendicular polarized spectra only are used |
| 5 | Infrared plus Raman nonpolarized spectra are averaged |
| 6 | Infrared plus Raman parallel plus Raman perpendicular polarized spectra are averaged |
| 7 | *Infrared plus all Raman data are averaged |
| 8 | *All Raman data are averaged |
| 9 | Raman parallel plus Raman perpendicular polarized spectra are averaged |

(* indicates those types of spectra treated in this work)

Table V shows in detail the variations which are possible through option number 3. If the input value of option 3 is from 1 to 17 the program will use the class of compounds corresponding to the input value. (The 17 classes were given earlier in Table I.) As shown in Table V, however, 4 other values of option 3 are allowable. These values simply instruct the program to combine certain classes.

TABLE V

Option Three in Program TRAIN

(The Class of Compounds Considered)

| Value of Option | Description |
|---|---|
| 1 through 17 | Classes 1 through 17 respectively |
| 18 | All halogens (class 7 through 10) |
| 19 | Compounds containing F, Br and/or I (classes 8, 9 and 10) |
| 20 | All alcohols (classes 3, 4 and 5) |
| 21 | Compounds containing O-H (classes 3, 4, 5 and 16) |

A separate program called TRAIN4 was used to combine spectra in a entirely different way. TRAIN4 uses the infrared spectrum of each compound and either the parallel polarized Raman spectrum or, in the case of solids, the nonpolarized Raman spectrum. For each compound the infrared data from 1875 $cm^{-1}$ to 500 $cm^{-1}$ is concatenated, i.e. linked in series with the Raman data from the same region to form a single vector. Thus 56 infrared readings and 56 Raman readings are joined to make a single vector representing the compound. TRAIN4 has all of the options that are in TRAIN except, of course, option one. When trained vectors are printed TRAIN4 seperates the infrared portion of the vector from the Raman portion and plots them separately on the same scale as is used by TRAIN. Clearly TRAIN4 could have been incorporated in TRAIN as yet another option, but a separate program was written to reduce the complexity of the programing effort.

3. TEST and SNOOP

Appendix B includes listings of both TEST and SNOOP. The use of TEST is self-explanatory and needs no discussion. When SNOOP was used to calculate the average representation of a class of compounds, two types of representations were generated. The average representation of each class was calculated and plotted, and the average of all compounds in the class minus the average of all compounds not in the

class was plotted. This second type of average representation was obtained by making a minor program change in SNOOP, and is not shown in Appendix B.

## RESULTS AND DISCUSSION

### 1. Results from SNOOP

As mentioned earlier there are two types of "average" spectra which have been obtained. We have called the first type the Simple Average; and the second, the Muted Average. The Simple Average is obtained by adding all the spectra from a given class and specified type of spectrum and dividing the total by the number of spectra. Let $\vec{f}_i$ represent the ith spectrum of the n spectra in a particular class. Then the Simple Average representation of a class is given by:

$$\text{Simple Average Representation} = \sum_{i=1}^{i=n} \vec{f}_i \; / \; n$$

If $\vec{f}_j$ represents the jth spectrum of the m spectra not in the class, the Muted Average representation of the class is given by:

$$\text{Muted Average Representation} = \left[ \sum_{i=1}^{i=n} \vec{f}_i \; / \; n \right] - \left[ \sum_{j=1}^{j=m} \vec{f}_j' \; / \; m \right]$$

When the Muted Average is calculated, negative intensities are encountered. For example, Figure 7 shows the Muted Average of Raman parallel polarized spectra of esters. Figure 8 shows the same spectrum, but only the positive intensities are printed. Since negative intensities do not give any direct information about the class of compounds being treated, we have adopted the convention of printing only the positive portion of the average spectra or the calculated vector under discussion. In this context it should be noted that a strong positive band does provide evidence that a compound in that class would be expected to have that vibrational frequency. Conversely, however, the presence of a strong negative band does not mean that a compound in the particular class is forbidden to vibrate with the frequency which appeared as a negative band. The only interpretation which can be given to negative bands is that compounds not in the class being treated have vibrational frequencies in that region. In general this information is not useful to the chemist interested in vibrational analysis.

The Simple Average representation has the advantage that it shows all of the bands characteristic of a chemical class, and tends to preserve their relative intensities. No bands are lost or muted simply because they happen to be common in compounds both in and out of the class. The Muted Average, of course, has the advantage that while some common bands

FIGURE 7

Muted Average

Raman Parallel Polarized Spectra

of

Esters

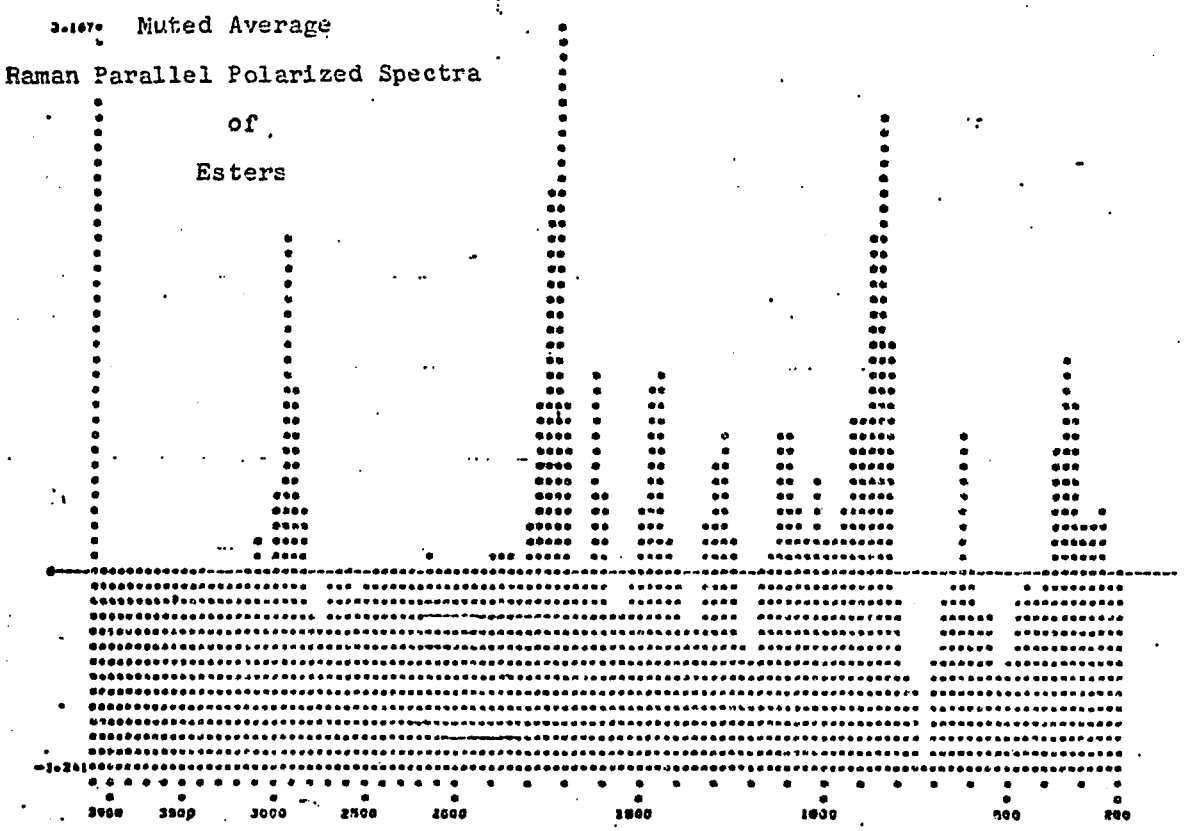


FIGURE 8

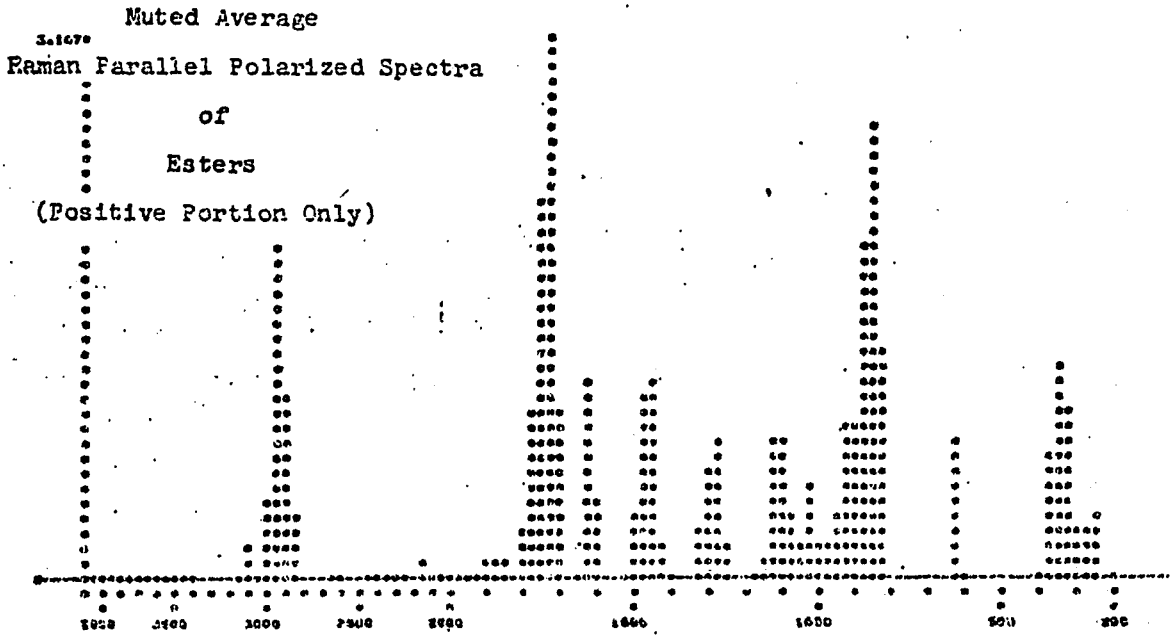
Muted Average

Raman Parallel Polarized Spectra

of

Esters

(Positive Portion Only)



may be reduced in intensity or lost, much of the noise is eliminated. Figures 9 and 10, for example, show the Simple Average and Muted Average spectra of the parallel polarized Raman spectra of ethers. Notice that the intensity scale for the two spectra differ. This must be considered when making comparisons. The Muted Average spectrum shows the peaks more sharply defined than does the Simple Average, but notice the relative intensity change when using the Muted Average. The C-H stretching bands around 3000 $cm^{-1}$ are no longer the strongest peak in the spectrum. More importantly perhaps, the bands between 1600 and 1800 $cm^{-1}$ have disappeared in the Muted Average. This is a clear example of how the Muted Average helps to remove noise. There are no ether vibrational bands above about 1550 $cm^{-1}$. The bands appearing in the Simple Average spectrum in that region occur because 31 out of the 41 ethers used to calculate the average contain either a $C_6H_6$, C=C, or C=O moiety.

Table VI lists the classes and the types of spectra for which Simple Average and Muted Average spectra have been calculated and gives the number of compounds used to calculate each average. Since there are so few nonpolarized Raman spectra in each class it is unlikely that their averages will be useful, and the combination spectra (all Raman or all IR plus Raman) do not have significance outside of the context of this work. Thus only the Muted and

FIGURE 9

Simple Average

Parallel Polarized Spectra

of

Ethers

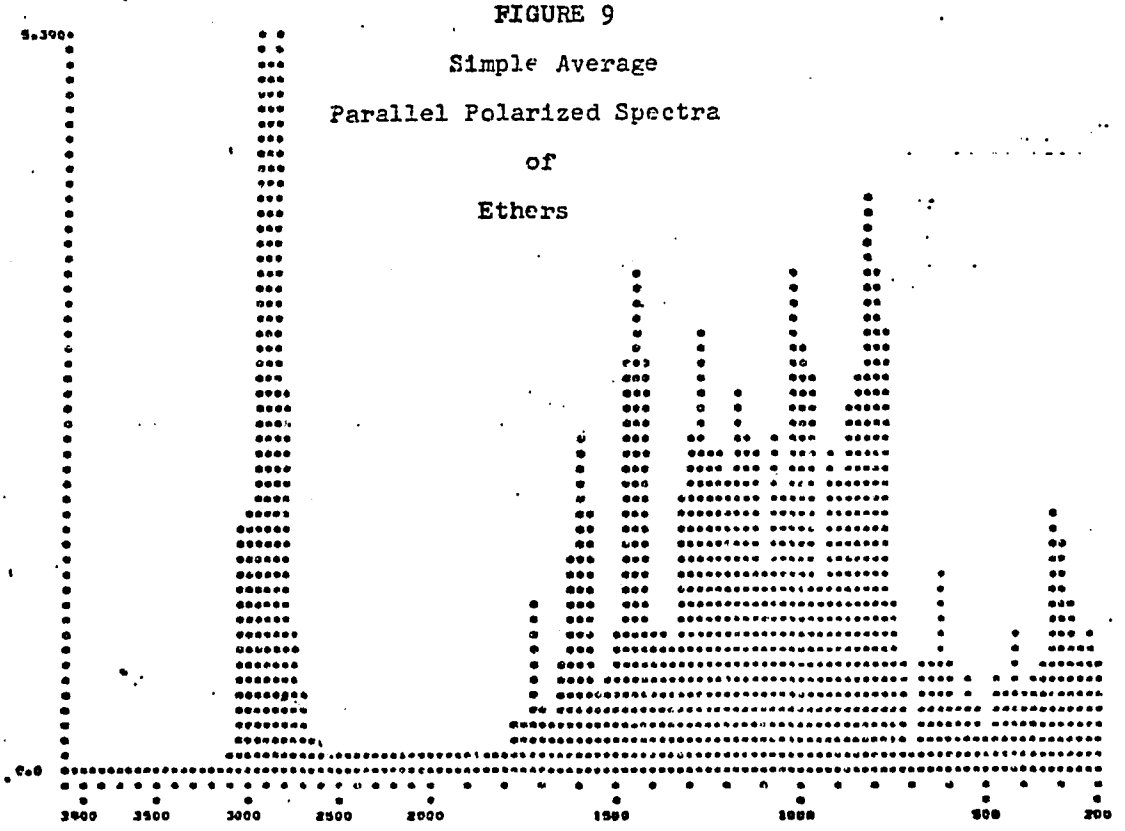


Figure 10

Muted Average

Parallel Polarized Spectra

of

Ethers

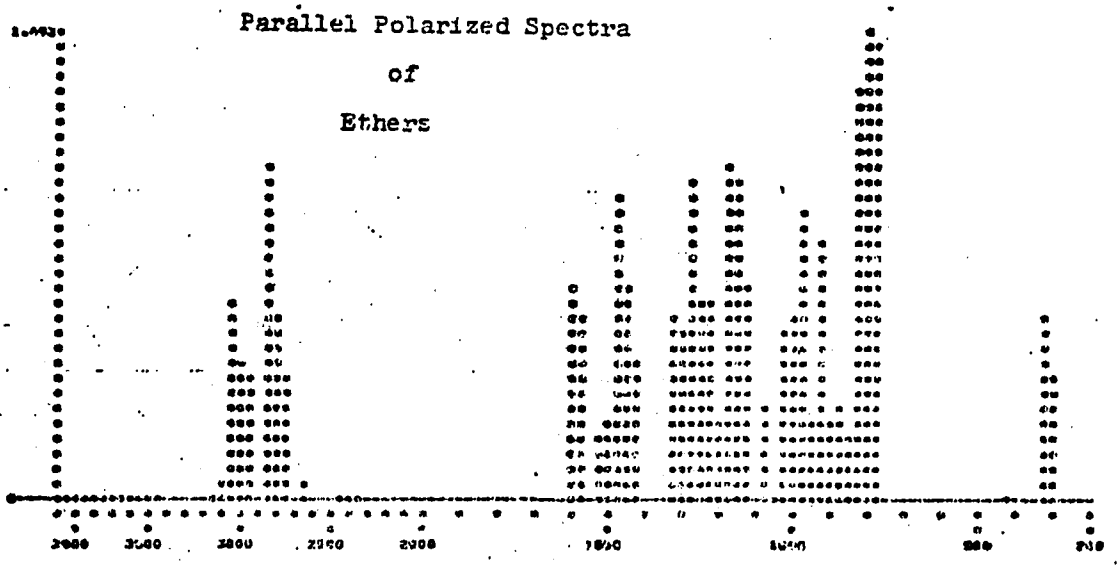


TABLE VI

Number of Compounds Used to Calculate Averages

| Type of Spectrum | IR | All Raman | Parallel Raman | Perpendicular Raman | Nonpolarized Raman | IR plus Raman |
|---|---|---|---|---|---|---|
| CLASS | | | | | | |
| Esters | 92 | 92 | 87 | 87 | 5 | 92 |
| Alcohols | 45 | 45 | 43 | 43 | 2 | 45 |
| Ethers | 48 | 48 | 41 | 41 | 7 | 48 |
| C =C bonds | 39 | 39 | 33 | 33 | 6 | 39 |
| Ketones | 15 | 15 | 14 | 14 | 1 | 15 |
| 1° Alcohols | 32 | 32 | 30 | 30 | 2 | 32 |
| 2° Alcohols | 12 | 12 | 11 | 11 | 1 | 12 |

Simple Averages of IR, the parallel Raman, and perpendicular Raman are included in Appendix C.

2. TRAIN and TEST

Table VII shows the number of iterations in one series of experiments needed for TRAIN to converge on a solution vector. In this set of experiments the initial value of the vector was set at -1. All of the available data were used in these tests. A plus (+) sign in Table VII indicates that TRAIN did not converge and was halted at the number of iterations shown.

The type of spectra labeled "all Raman" in Table VI were obtained by averaging all the Raman spectra for a particular compound. "IR plus Raman" indicates that the IR spectrum and all available Raman spectra were averaged to form a single spectrum for each compound. As was mentioned earlier, "concatenated IR plus Raman" spectra were formed by taking the infrared data from 1900 $cm^{-1}$ to 500 $cm^{-1}$ and combining them with Raman data from the same region. In the case of solids, nonpolarized Raman data were used. For liquid compounds parallel polarized Raman data were used.

The extremely low number of iterations in all cases where nonpolarized Raman spectra were tested is due to the very limited number of data available. (See Table VI.)

TABLE VII

Number of Iterations to Converge

| Type of Spectrum | IR | All Raman | Para. Raman | Perp. Raman | Nonpol. Raman | IR plus Raman | IR plus Raman Concatenated |
|---|---|---|---|---|---|---|---|
| CLASS Esters | 40 | 500* | 100+ | 100+ | 4 | 80 | 32 |
| Alcohols | 15 | 100+ | 94 | 100+ | 4 | 19 | 45 |
| Ethers | 54 | 100+ | 92 | 100+ | 7 | 100+ | 27 |
| C = C | 87 | 22 | 11 | 54 | 4 | 37 | 9 |
| Ketones | 22 | 49 | 24 | 66 | 2 | 40 | 16 |
| 1° Alc. | 32 | 100+ | 100+ | 100+ | 4 | 52 | 53 |
| 2° Alc. | 67 | 18 | 11 | 21 | 3 | 23 | 13 |

IR = Infrared
Para. = Parallel Polarized
Perp. = Perpendicular Polarized
Nonpol. = Nonpolarized
Alc. = Alcohol
1° = primary
2° = secondary

The only reason for running these tests was to determine whether these data were linearly separable from the rest of the data. In all cases they are.

When all tha Raman spectra were combined and the esters were used as the class to be trained, TRAIN was permitted to go to 500 iterations. Prior to doing that test, TRAIN was modified so that all the variables in the iterative part of the program were double precision. The program was not able to train a vector which would correctly classify esters even after 500 iterations. Futhermore, the final vector after 500 iterations did not look significantly different from the vector after 100 iterations. After 500 iterations there were 48 compounds out of 400 which could not be correctly classified by the vector, while after only 100 iterations there were only 46 compounds out of 400 which the vector could not correctly classify. In addition, the compounds which could not be classified correctly after 500 iterations were substantially the same ones which could not be classified after 100 iterations.

After the 100 iteration run with the esters the coding and chemical structure of the most troublesome compounds were checked. Only one anomaly appeared significant. The tri-p-tolyl ester of phosphorous acid was not included in the ester class because it is not a carboxylic acid ester.

It failed to be correctly classified 100 times out of 100 iterations. When the 500 iteration run was made the ester of phosphorous acid was included in the ester category, but this did not make the data linearly separable. In all of the other runs with esters the phosphorous acid ester was not included in the ester class. This experience was typical of what was observed a number of other times during this work. On occasion a computer run was made, then on scrutenizing the data a compound was found in the training set which was not coded correctly. Each time this occurred it was found on rerunning the data that the change in one compound did not significantly alter the vector. Even when one compound in the training set was incorrectly identified as not in the class being treated, the error made little difference in the final vector. In a sense the pattern recognition technique is rather forgiving.

The notation $100^{+}$ in Table VII means that the program was halted at 100 iterations without converging. Failure to converge could be due to a failure to allow the program to go far enough or due to a lack of linear separability in the data.

With every class of compounds except alcohols we found that the concatenated data trains considerably faster than other types of spectra. It is not surprising that alcohols do not train as well with the concatenated

vector since these data do not include the very characteristic O-H stretch region in the infrared. We would expect similar results if amines or other N-H compounds were classified. Interestingly, however, the secondary alcohols are an exception to what might have been anticipated. The reason for this appears to be in the characteristic Raman of secondary alcohols, but futher discussion will be deferred until the section on alcohols.

3. Program TEST

In pattern recognition the final criterian of success is the measure of how well the trained vector categorizes new patterns, i.e. patterns not used in the training. Program TEST reads trained vectors previously written on disk by program TRAIN and tests the appropriate set of spectra.

Tables VIII, IX, and X summarize the data obtained with TRAIN. The vector in each case reported in these tables was initiated at -1. In Table VIII the predictive ability of vectors trained with 50% of the available deck is reported. Tables IX and X show the predictive ability of vectors trained with 75% and 80% of the data, respectively. In every case the ratio of compounds in the class to compounds not in the class was maintained. For example, among the 400 compounds in the data there are 94 esters

TABLE VIII

Percentage of Compounds Correctly Classified

50% of Data Used in Training Set

| Type of Spectra | IR | All Raman | Paralled Polarized Raman | Perpendicular Polarized Raman | IR and Raman Averaged | IR and Raman Concatenated |
|---|---|---|---|---|---|---|
| CLASS | | | | | | |
| Ester | 74 | -- | (86) | -- | 72 | 92 |
| Alcohol | 97 | (82) | 86 | -- | 95 | 84 |
| Ether | 80 | (68) | 79 | (70) | (69) | 87 |
| C = C | 89 | 89 | 86 | 86 | 87 | 89 |
| Ketones | 89 | 91 | 89 | 85 | 89 | 91 |

-- indicates that the training set did not converge in 100 iterations.

( ) that larger training sets of the same type failed to converge.

TABLE IX

Percentage of Compounds Correctly Classified

75% of Data Used in Training Set

| Type of Spectra | IR | All Raman | Parallel Polarized Raman | Perpendicular Polarized Raman | IR and Raman Averaged | IR and Raman Concactenated |
|---|---|---|---|---|---|---|
| CLASS | | | | | | |
| Ester | 82 | -- | -- | -- | 83 | 95 |
| Alcohols | 98 | -- | 72 | -- | 96 | 87 |
| Ethers | 81 | -- | 81 | -- | -- | 90 |
| C = C | 91 | 90 | 91 | 85 | 91 | 92 |
| Ketones | 97 | 93 | 89 | 85 | 94 | 97 |

-- indicates that the training set did not converge in 100 iterations.

TABLE X

Percentage of Coumpounds Correctly Classified

80% of Data Used in Training Set

| Type of Spectra | IR | All Raman | Parallel Polarized Raman | Perpendicular Polarized Raman | IR and Raman Averaged | IR and Raman Concatenated |
|---|---|---|---|---|---|---|
| CLASS | | | | | | |
| Esters | 86 | -- | -- | -- | 84 | 96 |
| Alcohols | 97 | -- | 85 | 97 | 96 | 83 |
| Ethers | 84 | -- | 81 | -- | -- | 93 |
| C = C | 90 | 91 | 90 | 87 | 93 | 97 |
| Ketones | 97 | 93 | 89 | 85 | 93 | 97 |

-- indicates that the training set did not converge in 100 iterations.

(see Table I). When IR spectra of esters were trained using 50% of the data deck, 47 esters and 153 compounds which were not esters were chosen as the training set. The data were not randomized before each run, and the first compounds in the deck were always included in the training set. Thus all the compounds included in the training set where 50% of the deck was used were also included in those sets where 75% and 80% of the deck was used.

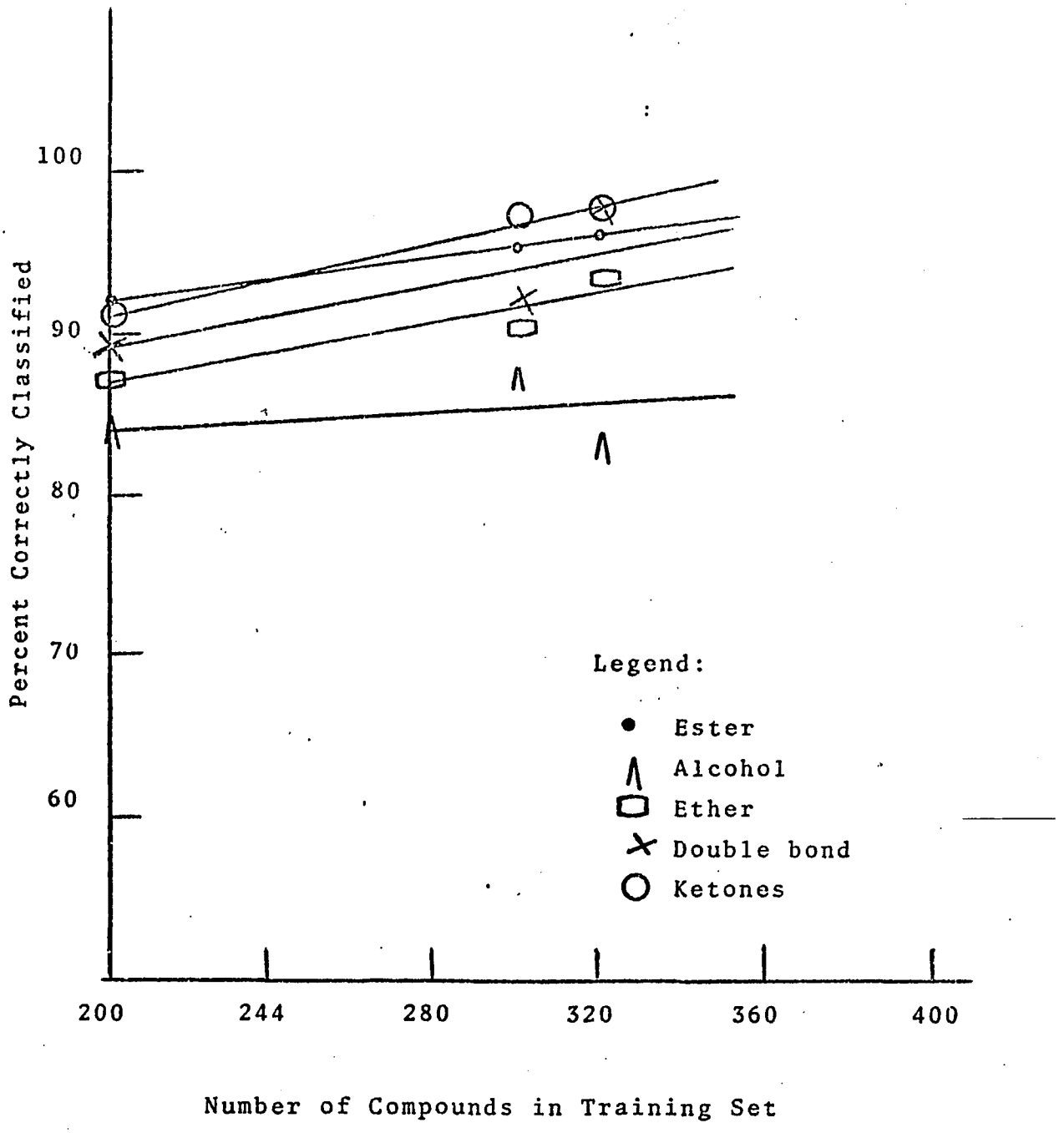
In Table VIII the numbers in brackets signify that while the vector trained with 50% of the deck did converge, the particular experiment did not converge when a larger portion of the data was used. Where data are not presented, signified by a dashed line, the vector did not converge and no run of TEST was made.

With every class except alcohols the type of spectra which are concatenated IR and Raman data are at least as predictive as other types of data. The apparent exception of alcohols is no doubt due to the exclusion of the O-H stretching region in the concactenated data. Interestingly, with esters the Raman data alone did not converge, but the use of Raman data in the concatenated vector seems to significantly improve the predictive ability over what is achieved with IR data alone or Raman data alone. When IR and Raman are combined by simply averaging vectors, the

FIGURE 11

Predictive Ability of Concatenated Vectors

(Vectors Initiated at -1)



data suggests that the predictive ability is not as good as when the IR is used alone.

Figure 11 shows a plot of the percent correctly predicted versus the number of compounds used in the training set. All of the vectors referred to in Figure 11 were obtained from concatenated infrared and Raman data using -1 as the initial value of the starting vector. In general the data suggests that there is improvement in the predictive ability when more compounds are used in the training set and moving from 200 compounds in the training set to 320 compounds improves the predictive ability on the average by about 5%.

It seems that the initial values of the vector plays a relatively minor part in determining what the final vector will look like. In a series of tests using 50% of the deck we trained a number of classes of compounds by initiating the vector at +1, and -1. A third solution vector was then calculated by averaging the positively initiated vector and the negatively initiated vector. In each test in this series the type of spectra used were concatenated IR and Raman data. Table XI shows that the predictive ability of the vector obtained by averaging the two calculated vectors did not differ significantly from the predictive ability of the two calculated solutions, and in no case was the average solution poorer than either of the calculated solu-

tions. Remembering that the average of two solution vectors is itself a solution to the trainig set, and noting that the training set used for each class in Table XI was the same it is understandable that the solution arrived at by averaging is about as predictive as the others. In the two instances where the average solution appears to be more predictive than either calculated solution, we must, at this time, attribute that to scatter.

TABLE XI

Predictive Ability of Concatenated Vectors

50% of Data Used in Training Set

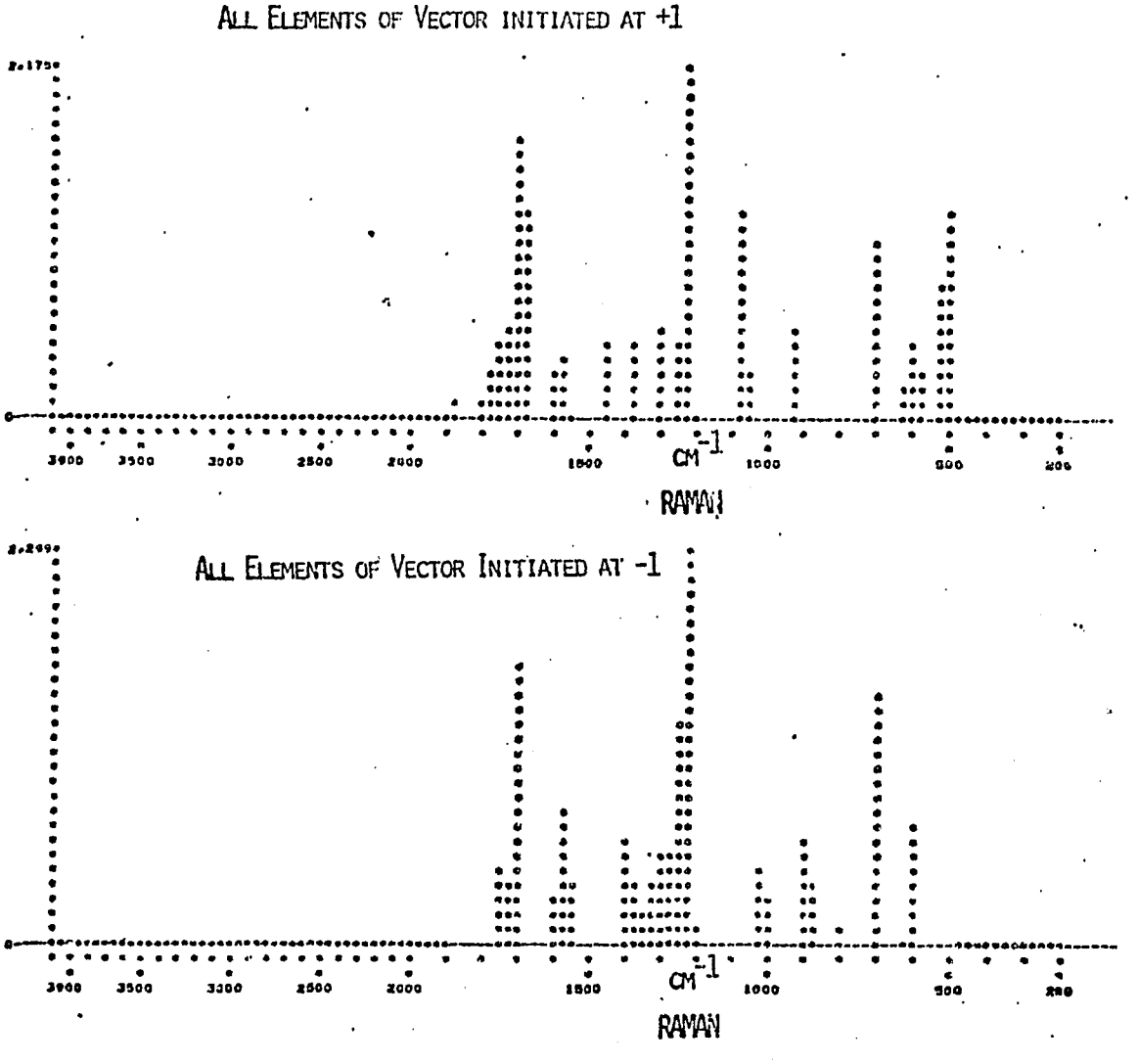
| | % Correctly Predicted | | |
|---|---|---|---|
| | Positively Initiated | Negatively Initiated | Average Solution |
| CLASS | | | |
| Ester | 97 | 92 | 96 |
| Alcohols | 93 | 84 | 92 |
| Ethers | 85 | 87 | 89 |
| C = C | 98 | 89 | 97 |
| Ketones | 100 | 91 | 100 |

Figure 12 shows the Raman portion of vectors which were trained using concatenated data. In both cases the class treated was ketones, and the result of initiating the vector at +1 and at -1 is shown. Notice that the in-

FIGURE 12

TRAINED VECTORS OF KETONES

RAMAN PORTION OF CONCATENATED VECTOR



tensity scales in the two plots are different; this should be taken into account when making comparisons. In general, vectors which are trained with a particular class of compounds but using different starting values are remarkably similar. Usually there are only three or four bands which appear in one spectrum but not the other. In Figure 12 for example, the only major discrepency between the two spectra is a band which appears at 500 $cm^{-1}$ in the vector initiated at +1. With the exception of some minor shifts and line broadening the spectra are about the same. There is no significant difference in the number of iterations to converge on a solution when the vector is initiated with either +1 or -1.

In view of the data presented we have concluded that the best way to combine infrared and Raman data is to concatenate the data from the two sources and treat the result as a single vector. The parallel polarized spectra or the nonpolarized spectra should be used to represent the Raman spectrum. Perpendicularily polarized data in general do not seem to be easily separated, but this may be due mainly to the quality of the starting spectra. They are frequently very low intensity spectra and manually coding them is difficult. The predictive ability of vectors trained with concatenated data seems to be superior to those trained with either infrared or Raman separately.

While averaging vectors which were initiated with different values does not significantly improve the predictive ability it does seem to average out some of the scatter, without adversly affecting the predictive ability.

In the sections that follow emphasis will be placed on interpreting vectors which were obtained by averaging positively initiated and negatively initiated trained vectors. The trained vectors were obtained from concatenated data.

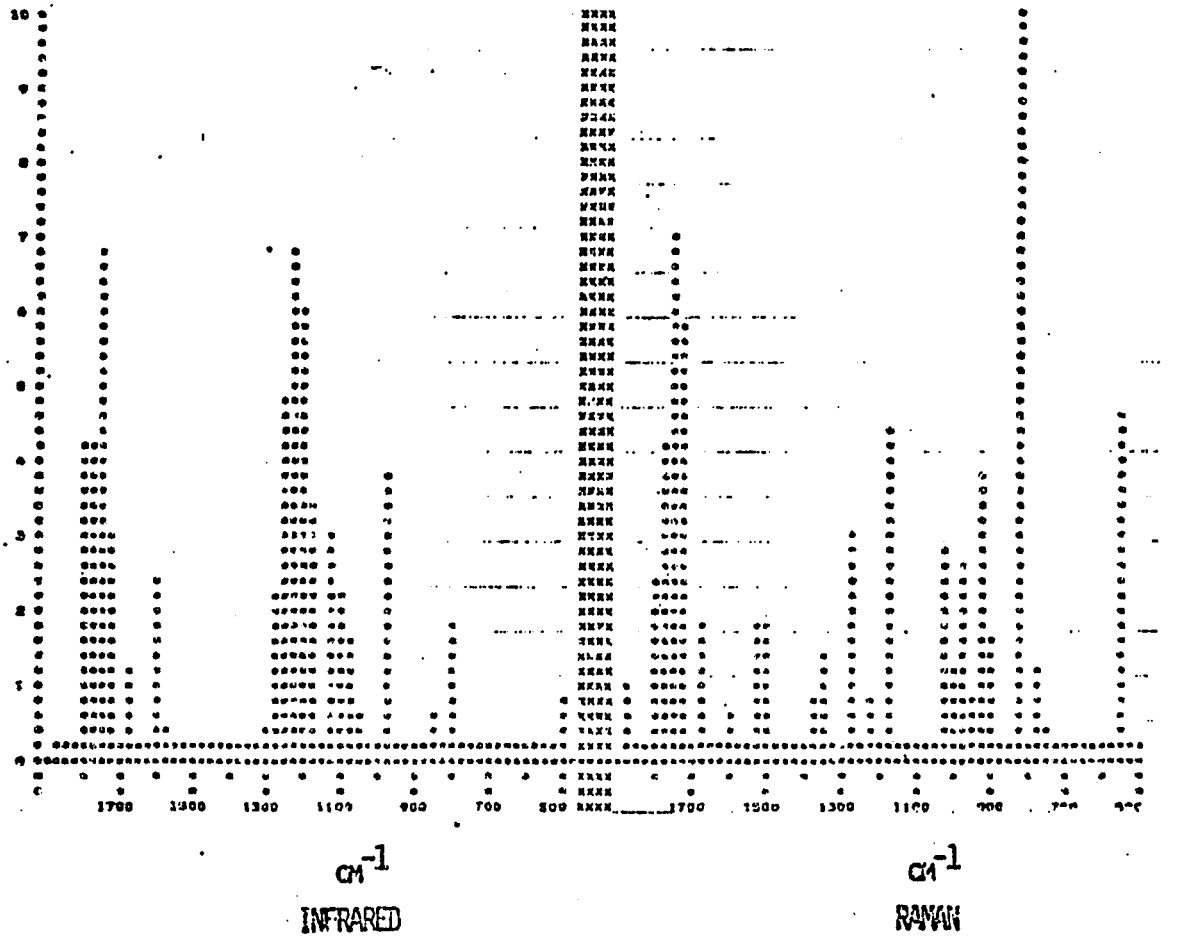
## 4. Esters

Figure 13 shows the average trained vector of esters. The vector was obtained by averaging two vectors, one trained by initiating all elements at +1, and a second trained by initiating all elements at -1. The intensity scale has been adjusted so that the strongest peak in either the infrared or the Raman is set at 10. All other

FIGURE 13

TRAINED VECTOR OF ESTERS

(AVERAGE OF TWO SOLUTIONS)



peaks in both spectra are scaled to the strongest peak.

The C=O stretch from 1800 to 1700 $cm^{-1}$ appears in both the infrared and Raman and is perhaps the most well known peak in the entire IR range [35]. This band is generally expected to occur in the region from 1780 to 1600 $cm^{-1}$, but we might note that carboxylic acid anhydrides, acid halides, acyl peroxides and ketones have fundamental modes between 1850 and 1750 [36], which might account for the relatively low cutoff of this band in the trained vector. Similarly the low end of the carbonyl stretching region is also the region where double bond stretching (1700 to 1500) occurs, and could easily account for the absence of strong bands in the region from 1700 to 1600 $cm^{-1}$. C-O stretching and/or bending is reported [37] from 1300 to 1000 and is generally seen in the IR but not in the Raman. Our vector has this characteristic.

The strongest peak in the vector is a narrow band

---

(35) L. J. Bellamy, The Infrared Spectra of Complex Molecules, 2 ed., Wiley, New York 1958.

(36) L. J. Bellamy, Advances in Infrared Group Frequencies, Barnes and Noble, New York 1968.

(37) R. M. Silverstein, G. C. Bassler, Spectrometric Identification of Organic Compounds, Wiley, New York 1964.

in the Raman at 825 $cm^{-1}$ which does not seem to have a correspondingly strong band in the IR. Looking at the Simple Average and Muted Average representations of Esters (Appendix C) we see that there could well be a characteristic Raman ester band in the region from about 950 to 800 $cm^{-1}$. The band seems greatly reduced in intensity in the Simple Average IR of esters and is totally absence in the Muted Average IR representation of esters. These data prompted us to look at the Raman spectra of all of the 94 esters in our data set. All except 5 of the compounds have a strong generally sharp Raman band in the region from about 950 to 800 $cm^{-1}$. Table XII lists the exceptions and also shows the nearest band which might be assigned to the vibration

TABLE XII

Esters Without a Band in the Region from 950 to 800

| Name of Compound | Nearest Assignable Band ($cm^{-1}$) |
|---|---|
| Acetic acid, o-tolyl ester | 1040 or 780 |
| Resorcinol, diacetate | 1000 |
| Benzoylacetic acid, ethyl ester | 1000 |
| Cinnamic acid, ethyl ester | 1000 |
| Bromophenylacetic acid, methyl ester | 1000 or 780 |

based on proximity, shape and intensity. It would appear that the proximity of the aromatic ring shifts the frequency to higher wave numbers, but further work would have to be done before an assignment could be made. For now we tentatively suggest that in general esters have a characteristic Raman band in the region from 950 to 800 $cm^{-1}$, and that band is frequently weak or absent in the IR.

5. Ethers

The characteristic feature of the vibrational spectra of ethers is the broad strong asymmetric stretching band between 1000 and 1300 $cm^{-1}$ in the infrared[38], which is weak or absent in the Raman. In Figure 14 the trained vector of ethers shows that characteristic. The band at about 525 in the infrared or 575 in the Raman is interesting because it corresponds to a reported [37,38] characteristic infrared vibration for ethers which is found in the region from 625 to 500 $cm^{-1}$. In reviewing the 49 ethers included in the data we find only 4 which do not have a band in the IR and/or Raman in the region from 625 to 500 $cm^{-1}$. Table XIII lists those four exceptions. The symmetry of the compounds

---

(38) A. J. Gordon, R. A. Ford, The Chemists' Companion, A Handbook of Practical Data, Techniques and References, Wiley, New York, 1972.

FIGURE 14

TRAINED VECTOR OF ETHERS

(AVERAGE OF TWO SOLUTIONS)

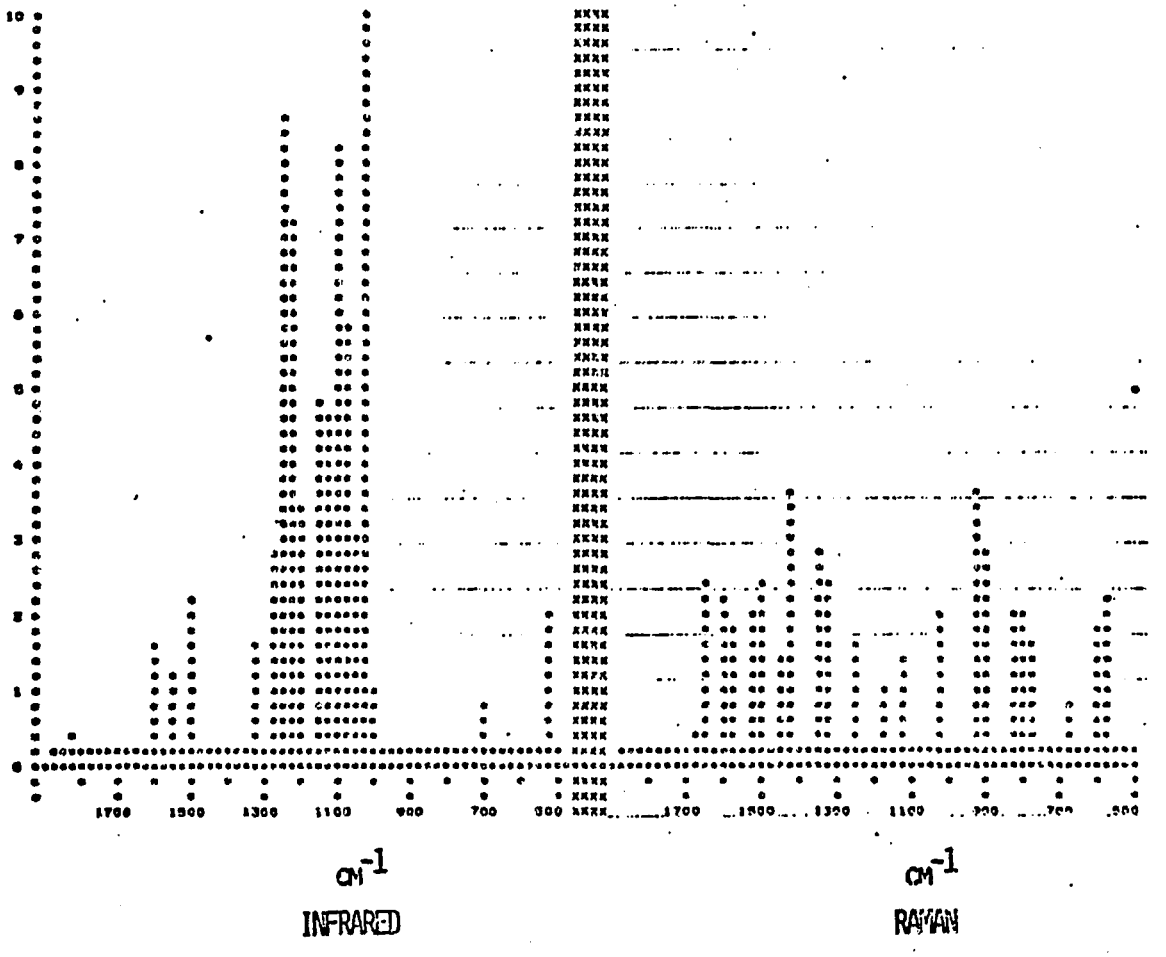



TABLE XIII

Ethers Without a 625 to 500 $cm^{-1}$ Band

| Compound Name | Structure |
|---|---|
| Isopentyl ether | C-C(-C)-C-C-O-C-C-C(-C)-C |
| Butyl ether | C-C-C-C-O-C-C-C-C |
| Isopropyl ether | C-C(-C)-O-C(-C)-C |
| Ethyl ether | C-C-O-C-C |

might be responsible for the suppression of the band; and, indeed, we have found that the band is very weak with other totally symmetric ethers.

6. Ketones

There are only 15 ketones in the entire data set, and therefore any conclusions based on the trained vector must be made tentatively. Of course, this caveat applies to all of the trained vectors, but it applies *a fortiori* to such a small class. Some comments will be made on the appropriate size and composition of training sets in a following section.

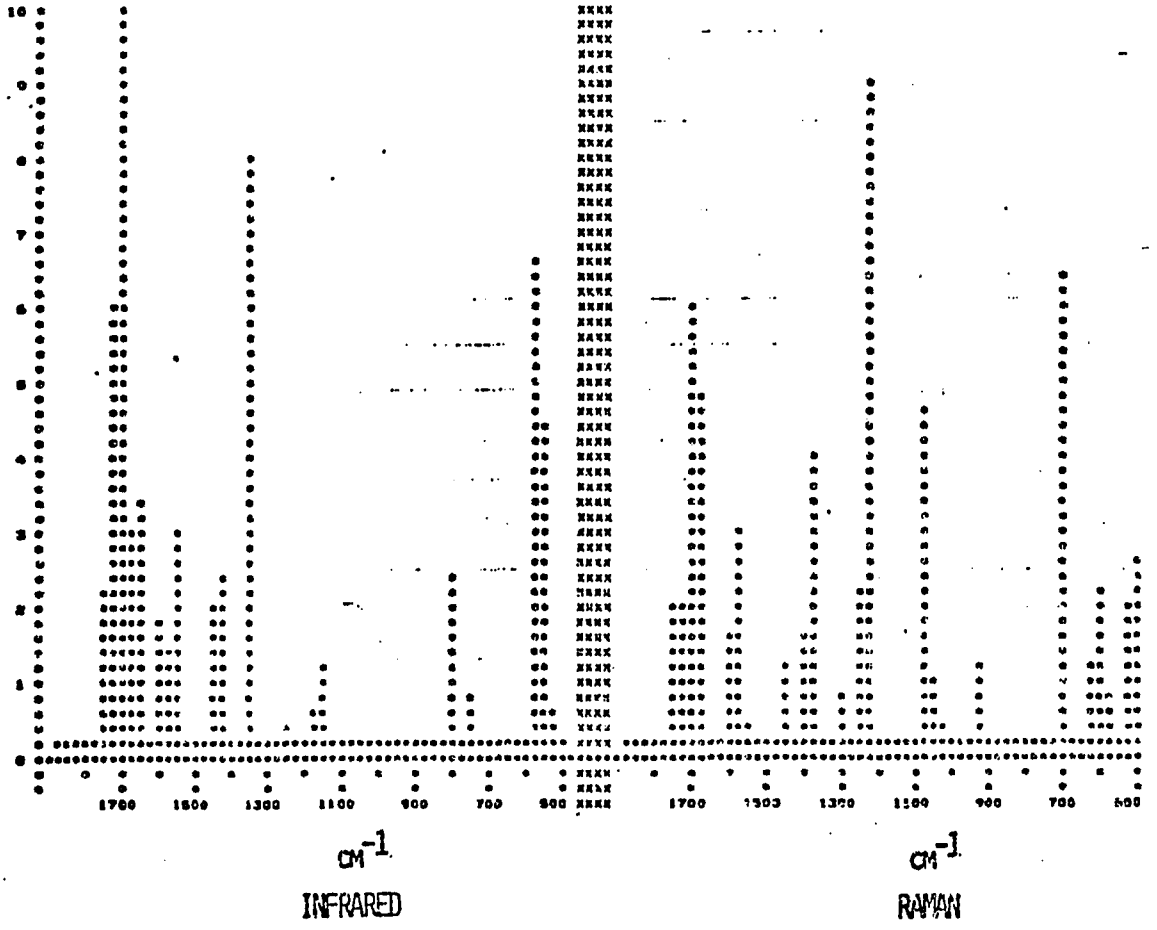
Figure 15 shows the average of two trained vectors using ketones as the class of trained compounds. The C=O stretch is clearly evident in both the IR and Raman at about 1700 $cm^{-1}$. The 15 ketones in our set are all aliphatic, we would therefore expect an IR band [39] of medium intensity in the region from 1225 to 1075 $cm^{-1}$, but the trained vector does not show this band. The Simple Average (Appendix C) spectrum of ketones, however, does show the band. Undoubtably, the reason for this is that the ethers have such a strong band in this area that they dominate.

---

(39) N. B. Colthup, Introduction to IR and Raman Spectroscopy, Academic Press, New York, 1964.

FIGURE 15

TRAINED VECTOR OF KETONES
(AVERAGE OF TWO SOLUTIONS)




In the training process, every time an ether fails to be correctly classified as not-a-ketone (none of the ketones in the set are also ethers), this band gets subtracted out. This points out the care that must be taken in interpretation of trained vectors. While the absence of a band in a particular region would suggest that the class of compounds being trained does not have a characteristic

band there, there are exceptions to this.

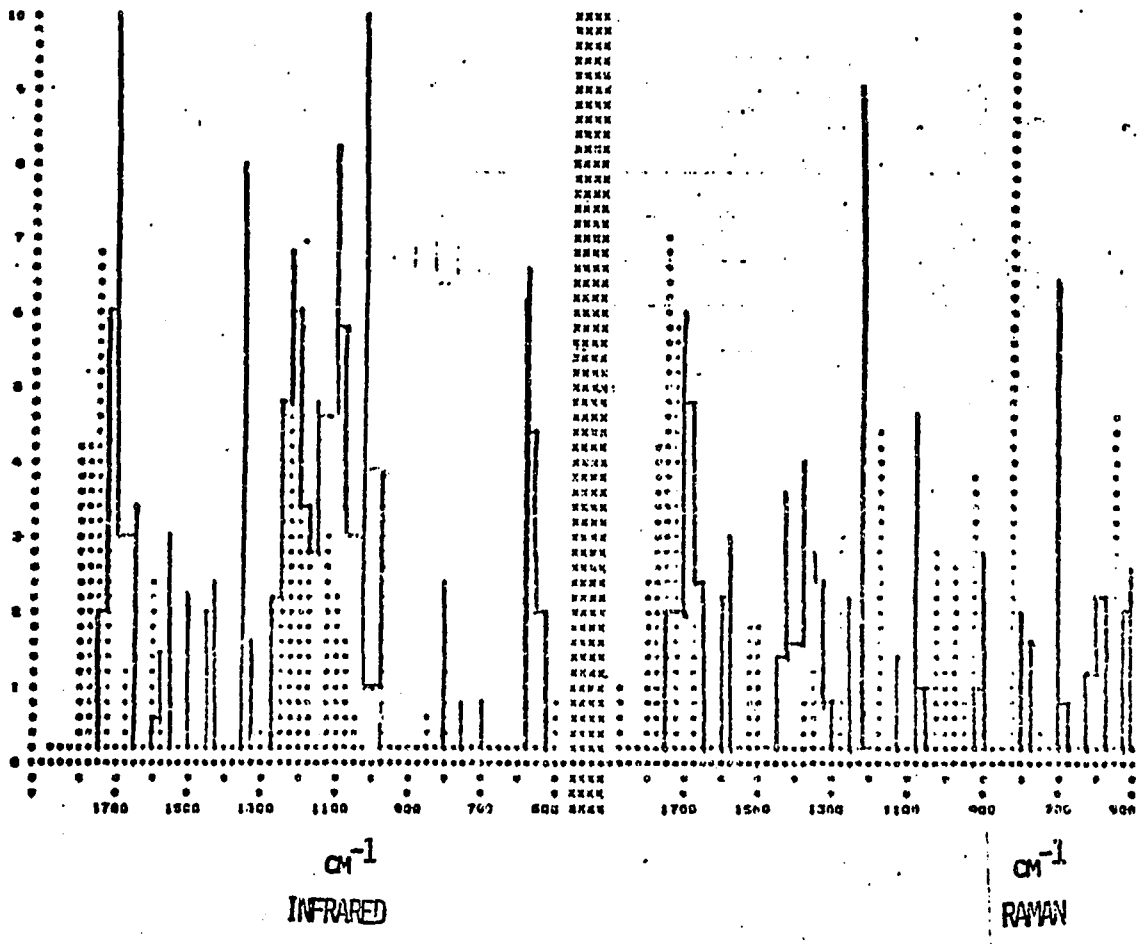
It is interesting to note that the strongest band in the Raman portion of the ketone vector is at 1250 $cm^{-1}$. This is precisely what we expect since the ether band in that region is usually not observable in the Raman spectrum.

We were prompted to train a vector using ketones in spite of the small number of compounds available, because of a desire to estimate the extent to which esters might be considered a combination of ketones and ethers. Figure 16 shows the trained vector of esters as was shown in Figure 13, but traced over the computer print is the outline of all those bands occuring in the trained vectors of ethers and ketones.

The carbonyl band in the ester at 1775 $cm^{-1}$ is shifted to higher wave numbers than is found in ketones, as expected. The band centered around 550 in the IR, which is characteristic of ethers, is absent in the IR of esters, but interestingly (and possibly fortuitously) the Raman of esters shows a band at 550. Many of the esters in the data have a distinct, sharp band around 500, but there are sufficient examples of esters without the band to leave in doubt the question of assignment.

FIGURE 16

TRAINED VECTOR OF ESTERS
WITH KETONE AND ETHER BANDS OUTLINED



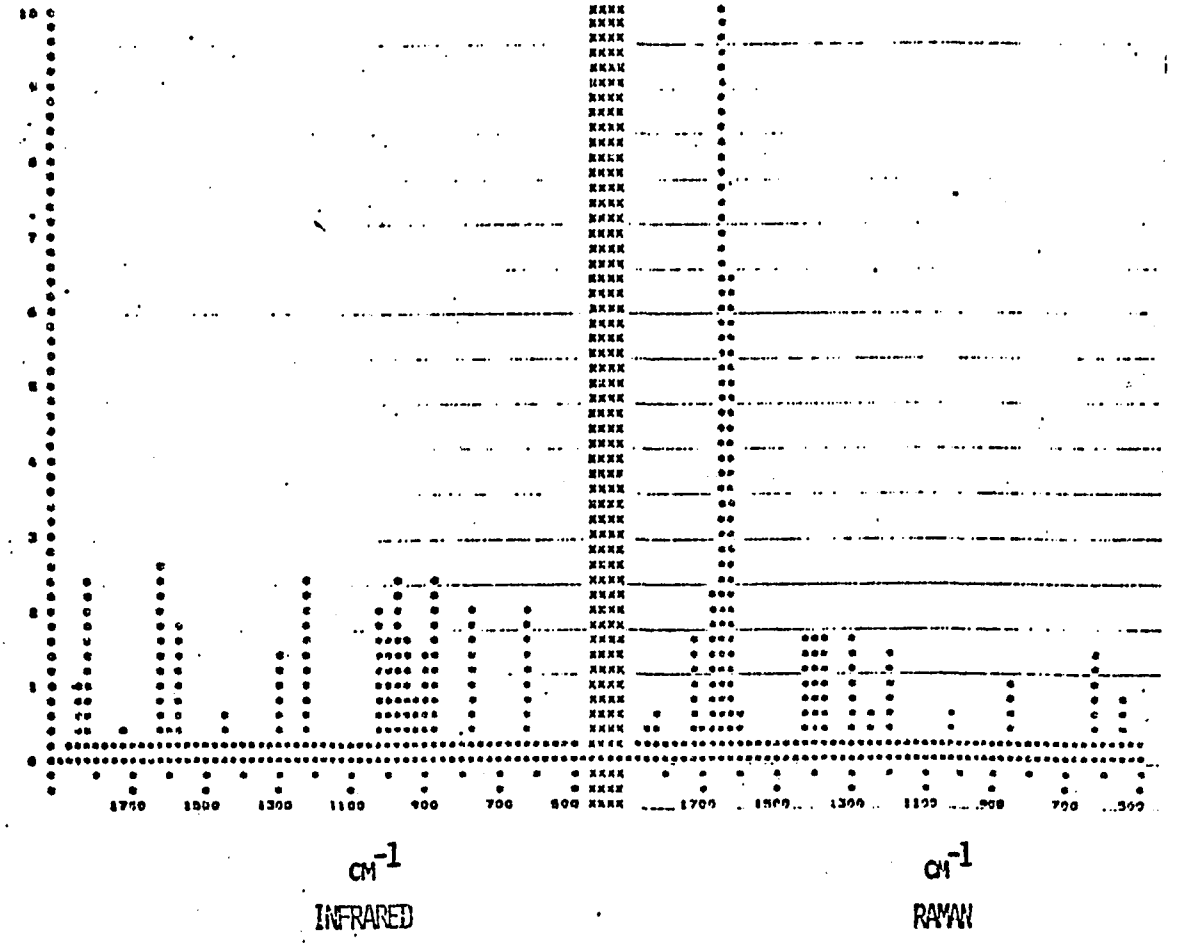

The similarity between the vector of esters and the combination of ethers and ketones is remarkable, particularly in the infrared. Yet, there are apparent differences to permit the differentiation of an ester from a ketone or an ether. This tends to show the power of the pattern recognition technique.

7. Double Bonds

In the data set there are 38 compounds which have a double bond in their structure, excluding double bonds which are a part of a benzene ring. Figure 17 shows the trained vector which resulted from treating these compounds

FIGURE 17

TRAINED VECTOR OF DOUBLE BONDS
(AVERAGE OF TWO SOLUTIONS)



as a class. The vector is dominated by the band between 1700 and 1600 $cm^{-1}$ in the Raman, as might be expected. All 38 of the compounds included in this class have a strong band in the Raman between 1700 and 1600 $cm^{-1}$ and usually the Raman band is much stronger than the corresponding IR band. The infrared protion of the vector has a broad band between 1050 and 875 $cm^{-1}$ which probably corresponds to the bands generally assigned to vinyl substituted, trans substituted (not cis) or geminally substituted double bonds.[36]. All of the compounds in the data set which fit one of these categories, with one exception, have a strong IR band around 1000 $cm^{-1}$, and the corresponding Raman band is weak or absent. The exception to this is cinnamyl alcohol ( Ø-C=C-C-OH ) which has a very strong Raman band at 1000 $cm^{-1}$ with a corresponding IR band. Unfortunately the Sadtler index does not label the compound as cis or trans, which undoubtably means that it is a mixture.

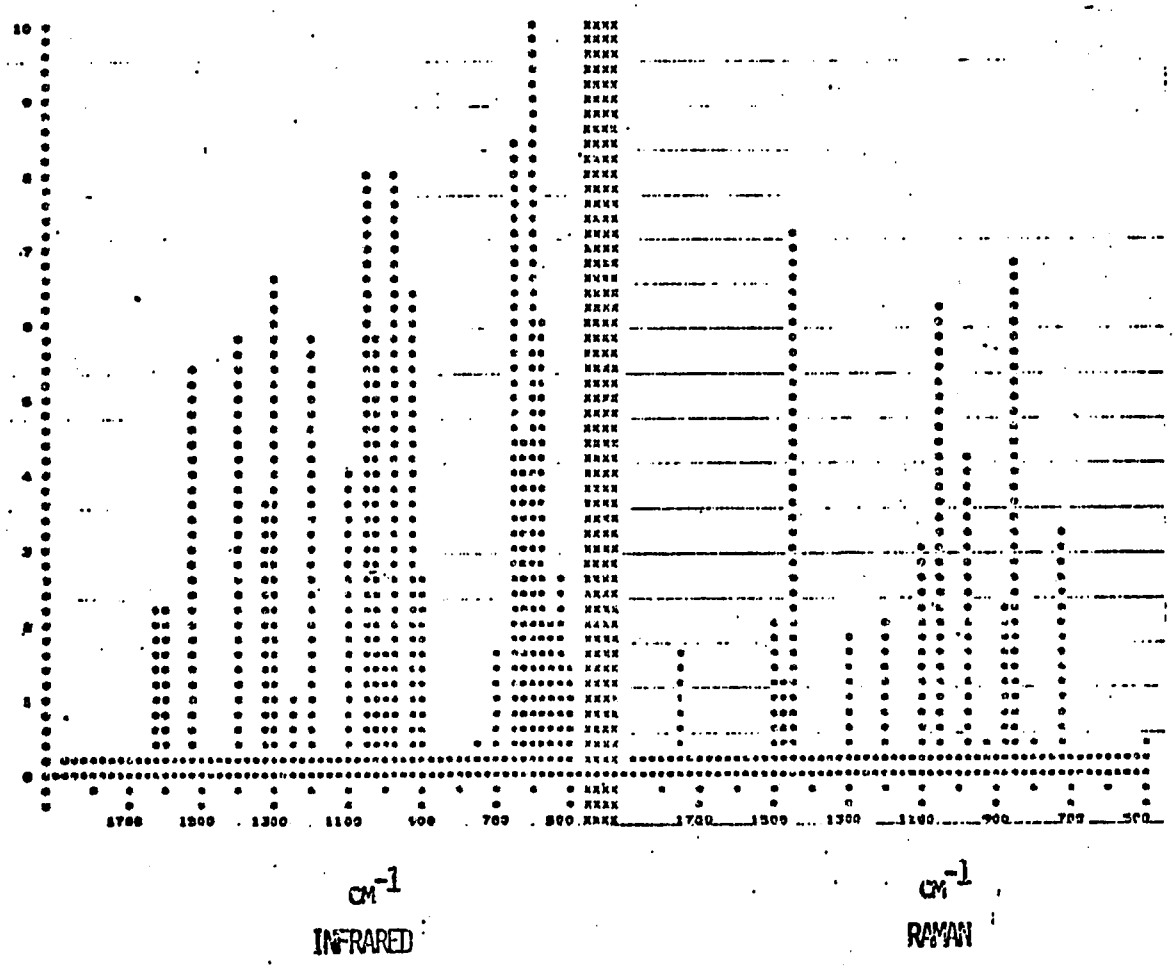
8. Alcohols

Figure 18 shows the trained vector which resulted when all of the 45 alcohols were treated as a single class of compounds. The data were composed of 29 primary alcohols, 9 secondary alcohols, and 4 tertiary alcohols. Three compounds have both primary and secondary moieties.

In order to test the power of the pattern recognition

FIGURE 18

TRAINED VECTOR OF ALCOHOLS
(AVERAGE OF TWO SOLUTIONS)




technique, the primary alcohols and the secondary alcohols were tested as seperate classes. When the primary class was treated those 3 compounds which have both primary and secondary groups were included as being primary alcohols. Similarly, when the secondary alcohols were trained the three compounds were treated as members of the class of

secondary alcohols. In both instances the program converged on a solution. Figure 19 shows the vectors which resulted from this treatment of alcohols.

Alcohols are characterized by the strong O-H band in the region from 3500 to 3100 $cm^{-1}$ in the IR[37,38,39]. This same band is absent in the Raman. Many other bands also exist in the finger print region of the spectrum, and it is these that are treated in this work.

In Figure 19 the band at 1725 $cm^{-1}$ in the IR portion of the secondary alcohol vector is the only band which clearly does not appear in the vector for all alcohols (Figure 18). In all likelihood this occurs because three of the secondary alcohols also contain carbonyl groups. Since the class was so small this should be considered an artifact of the data set.

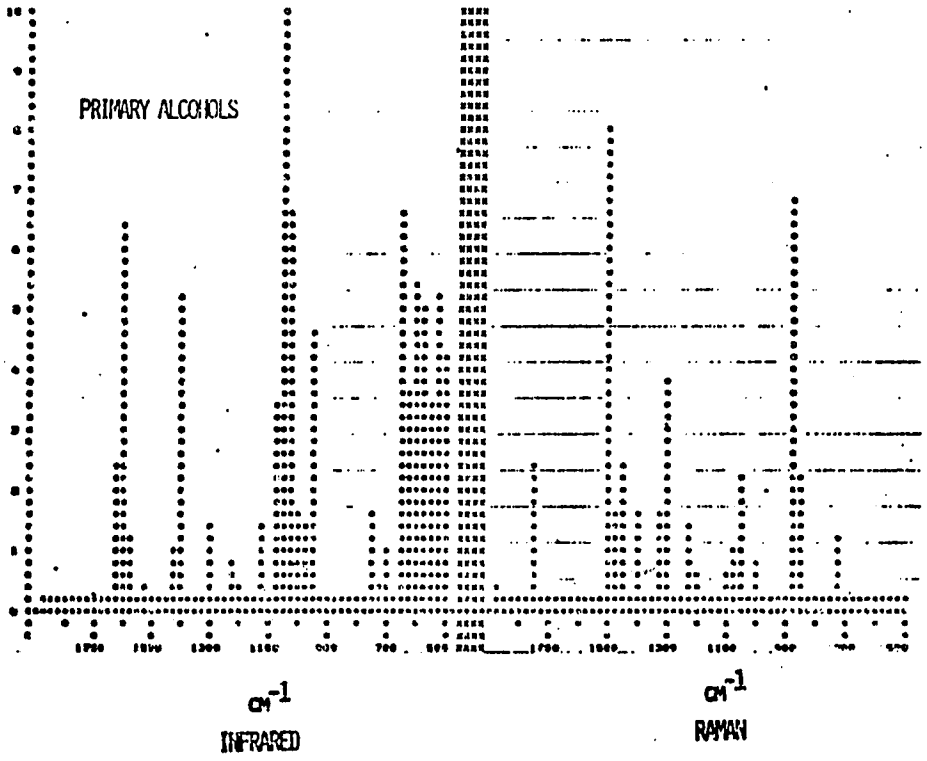
Figure 19 suggests that secondary alcohols differ from primary alcohols in the Raman in that they do not have a band at 1500 $cm^{-1}$ as do the primary compounds. In reviewing the actual spectra it is difficult to justify such a statement, although some band shifting in that region is evidenced. In view of the small number of data we refrain from any conclusion.

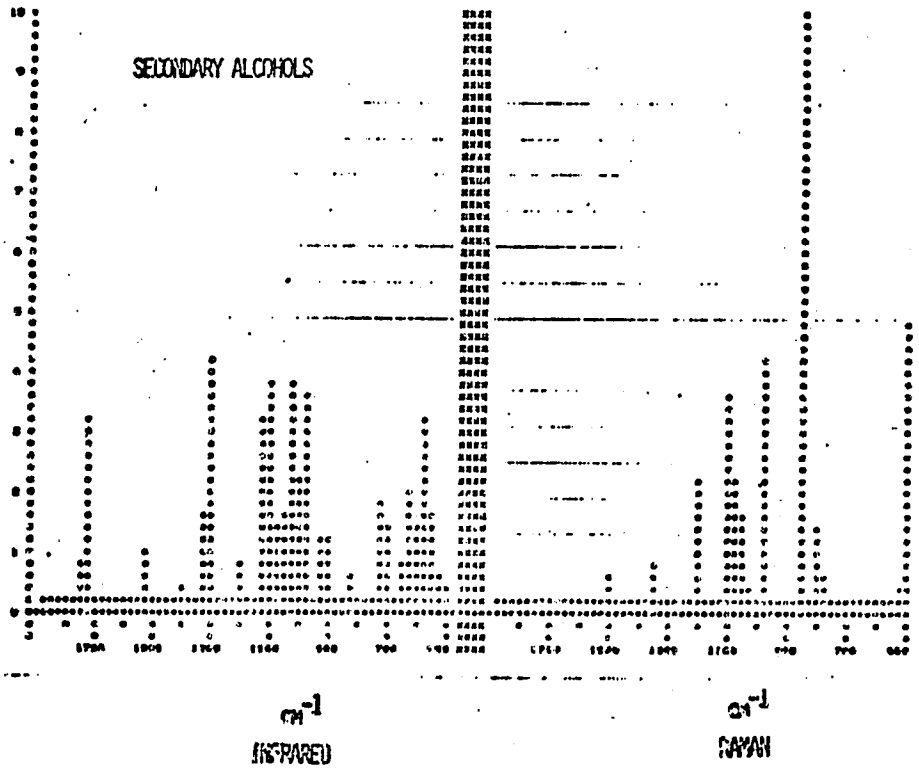
FIGURE 19

TRAINED VECTORS OF PRIMARY AND SECONDARY ALCOHOLS

(EACH VECTOR IS AN AVERAGE OF TWO SOLUTIONS)

(VECTORS WERE TRAINED USING ONLY ALCOHOLS IN THE DATA SET.)



A band in the Raman around 500 $cm^{-1}$ would appear to be characteristic of secondary alcohols and not primary , but again because of the small number of data we withhold further comment.

Pursuing the topic of primary vs secondary alcohols a little farther, we extracted from the data all compounds which were either primary, secondary or tertiary alcohols; but not combinations. In this manner we made a small data deck composed of alcohols only. We found that we could train vectors for either primary or secondary alcohols using IR data alone, using parallel polarized Raman alone, using perpendicularly polarized Raman alone or using linear combinations of these data. In every case the program converged in less than fifty iterations, offering a convincing demonstration of the power of the pattern recognition technique.

# SPECTRAL INTERPRETATION

## 1. Method Used

In this section we have attempted to carry the results of the pattern recognition techniques as far as seems practical, i.e. we attempt to use the trained vectors as supportive evidence for the assignment of vibrational frequencies. The original assignments were made using more traditional types of evidence.

Ab initio frequency assignments using pattern recognition data alone are well outside the power of the technique at this stage, and it seems likely that such assignments will remain outside the scope of pattern recognition for some time to come. In view of the success that the method enjoys when classifying compounds not used in the training set, however, it seems appropriate to use trained vectors to support frequency assignments. This is so particularly in view of the fact that spectroscopists frequently use the "reasonableness" of an assignment as supportive evidence, sighting similar assignments made for other molecules.

In an earlier section when presenting the data on predictive ability, we showed an advantage to using concatenated IR and Raman data. When averaging positively initiated and negatively initiated trained vectors, however, there does not seem to be an increase in the predictive ability

(nor any decrease). The average vector, nevertheless, did give a single representation which could be used to look for characteristic group frequencies. In looking for support for frequency assignments, however, we choose to have the broadest representation we can. To this end we have combined positively and negatively initiated trained vectors which resulted for concatenated data by superimposing the two vectors on the same frequency scale. The two superimposed vectors were traced in such a way that the strongest intensity in either of the two vectors appeared on the trace. Thus when one intensity at a particular frequency is stronger than the other, the strongest appears on the trace; and when one vector shows a positive intensity and the other does not, the positive intensity is traced. We have referred to this type of trace as a "broad representation".

The effect of this type of treatment is to broaden slightly the frequency bands associated with each chemical group and to emphasize strong bands. For this reason the broad representations which follow are presented without frequency scales.

Figures 20 and 21 show the broad representation of the infrared of ethers and the broad representation of the Raman of ethers respectively. Similarly, Figures 22 and 23 show the broad representation of the IR of compounds containing C=C and the broad representation of the Raman

FIGURE 20

Broad Representation of Ethers

INFRARED

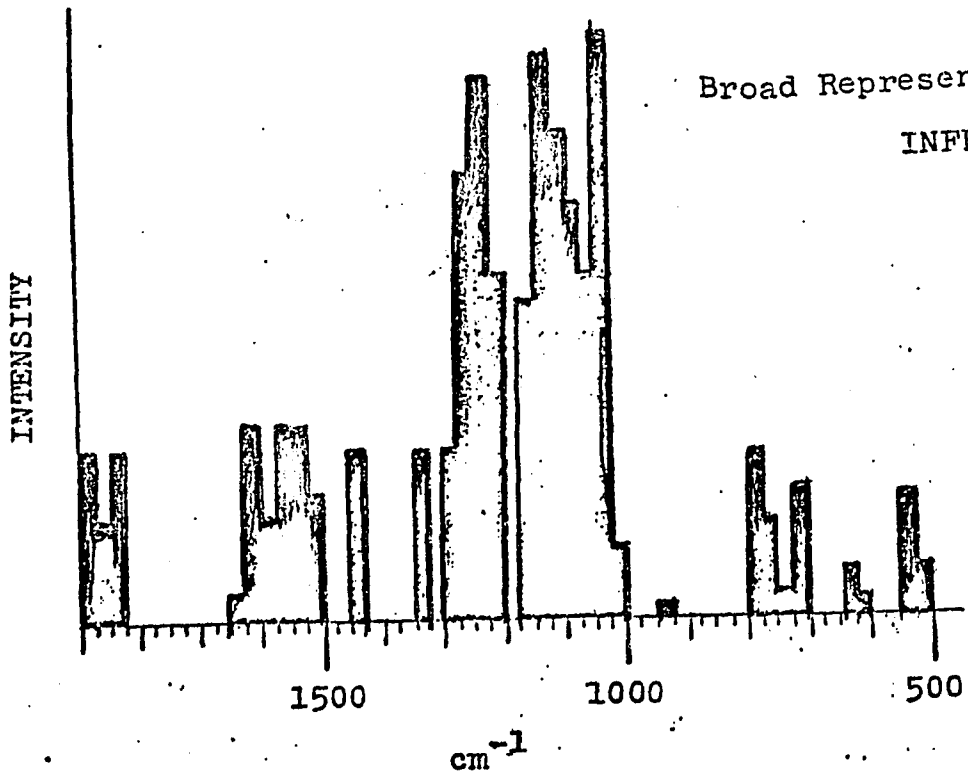



FIGURE 21

Broad Representation of Ethers

RAMAN

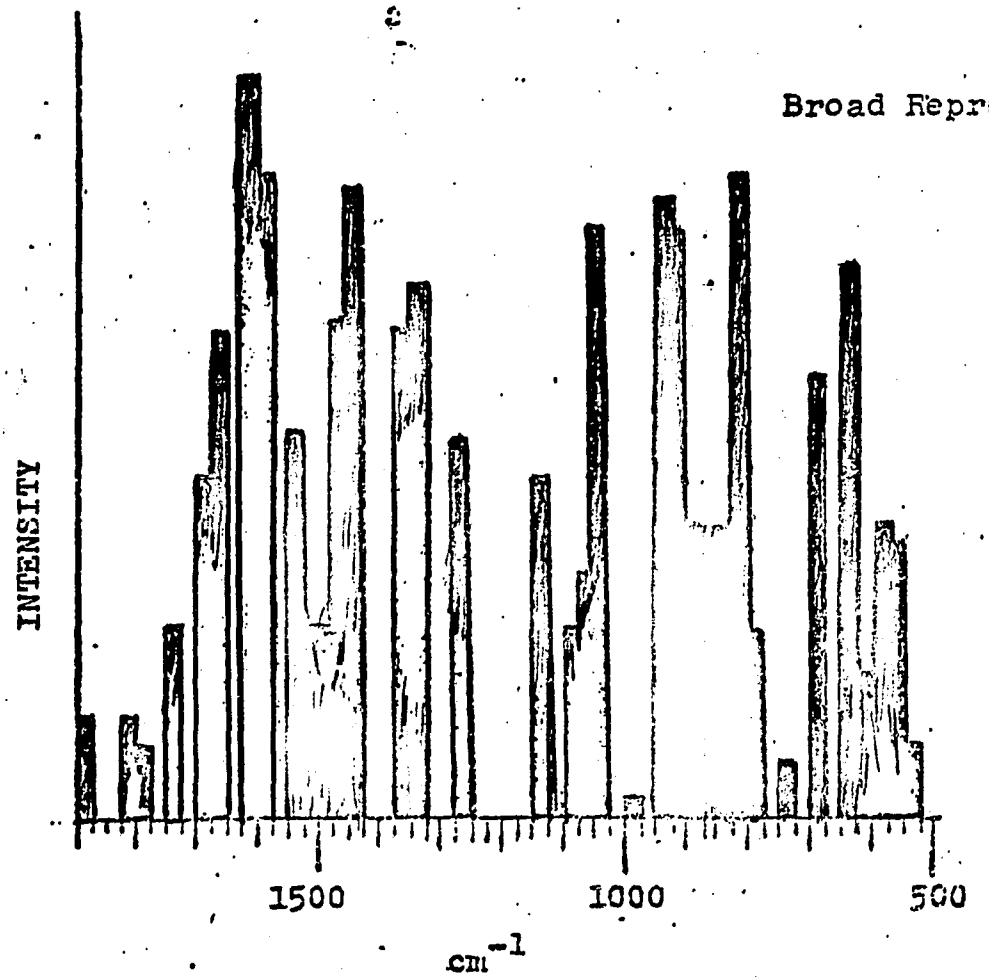


FIGURE 22

Broad Representation of Double Bonds

INFRARED

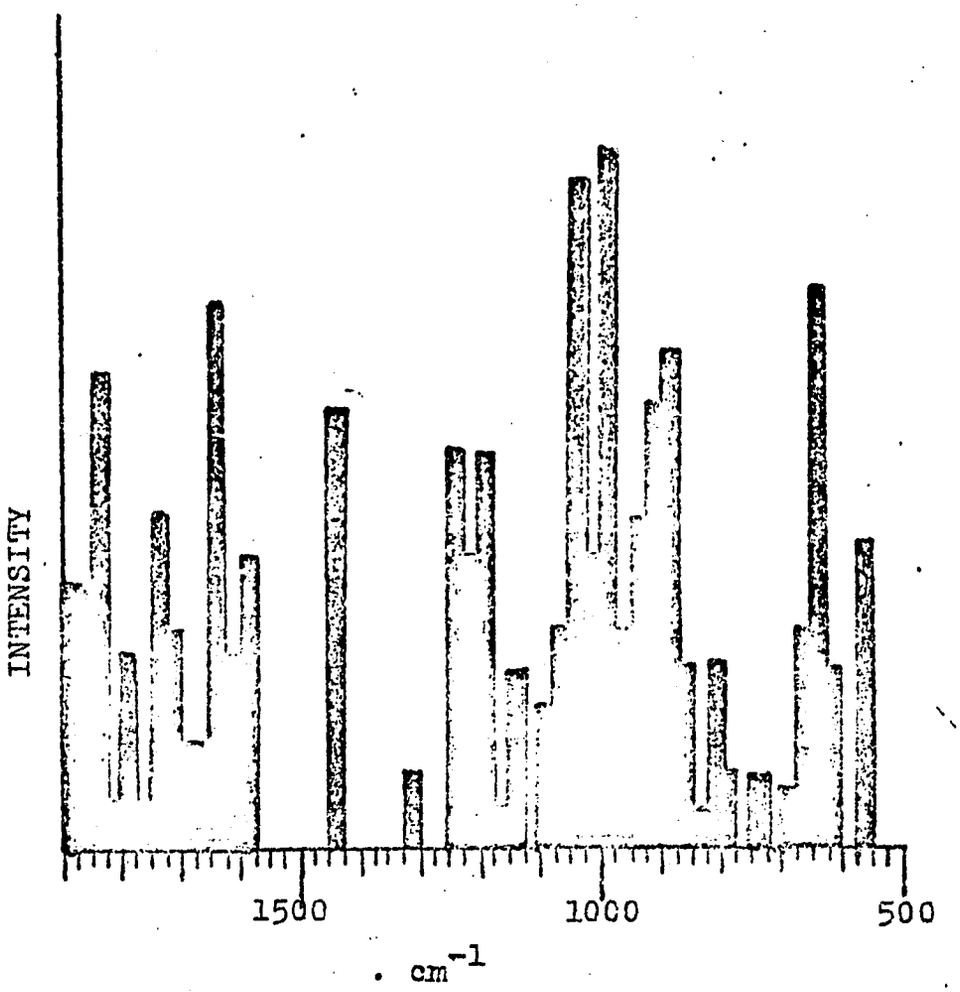
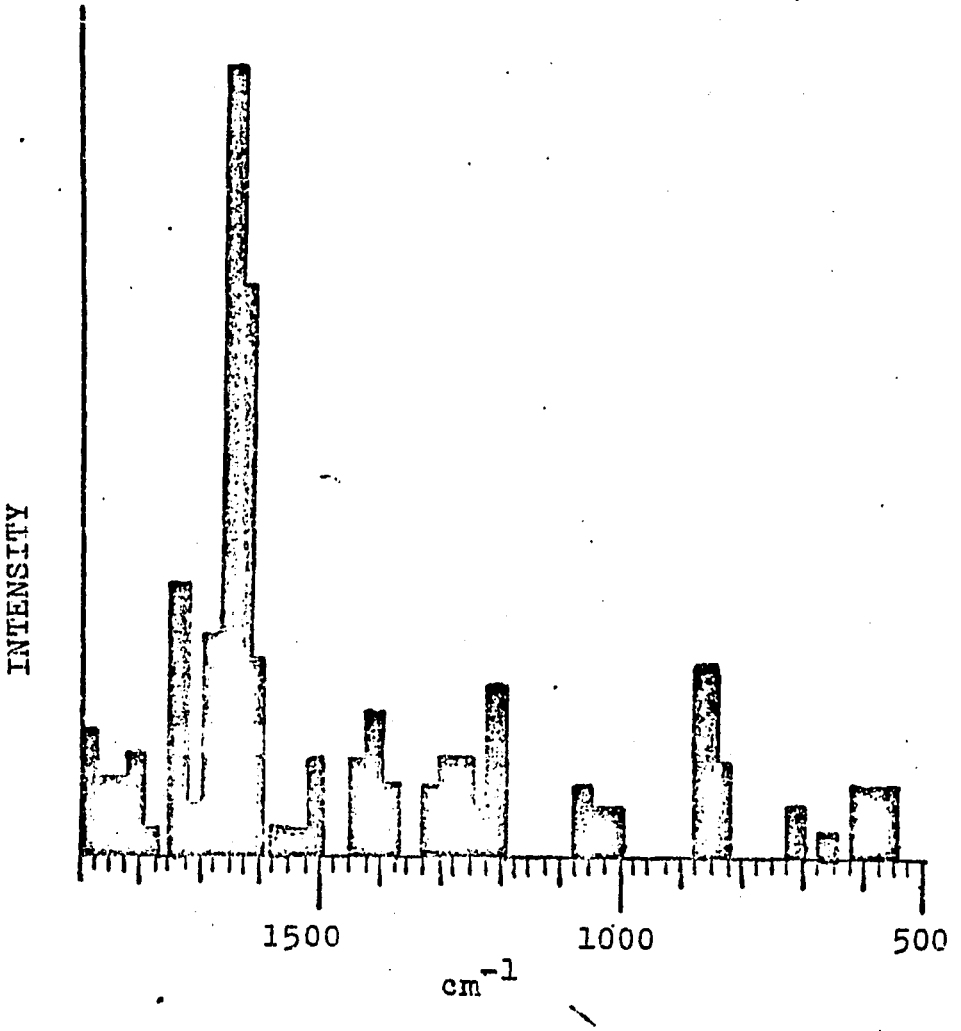


FIGURE 23

Broad Representation of Double Bonds

RAMAN



of compounds containing C=C. These representations were used to provide the supportive evidence for the frequency assignments made in the literature for 3 compounds, and to provide suggested assignments where possible. In each case, when an assignment was made in the literature the appropriate broad representations were checked to determine if the observed frequency appeared in one of the representations. If it did the frequency was labeled "supportable". If it did not appear in the proper broad representation, the observed frequency was labled "no supporting evidence". When an observed frequency was not assigned, all four broad representations were scrutenized and a suggested assignment was made only if there was no conflicting evidence.

2. cis-1, 2-dimethoxyethylene

Kimmel, Waldron and Snyder[40] recently published a vibrational analysis of cis- and trans-1,2-dimethoxyethylenes. The observed frequencies for the cis molecule, the assignments made and the available evidence from pattern recognition are included in Table XIV. The IR data were taken from the reported data on liquids since most of the IR data in the Sadtler Spectra were taken from liquids.

---

(40) H. Kimmel, J. Waldron, W. Snyder, J. Mol. Structure, No. 3, 21,445 (1974).

TABLE XIV

Interpretation of cis-1,1-Dimethoxyethylene

| Observed Frequency ($cm^{-1}$) | | | | Assignment | Evidence from Pattern Recognition |
|---|---|---|---|---|---|
| IR | | Raman | | | |
| 1843 | vw | ---- | | ---- | ---- |
| 1700 | m | 1720 | s | C=C str. | supportable |
| 1675 | m | 1690 | vs | C=C str. | supportable |
| 1562 | w | ---- | | ---- | either band, not C=C |
| 1460 | s | 1470 | w | $CH_3$ asym. def., $CH_3$ sym. def., | not apply |
| 1377 | vs | ---- | | $=C^2H$ i.p. rock | no supporting evidence |
| 1305 | sh | 1300 | m | | C=C band, not ether |
| 1292 | m | ---- | | o-$CH_3$ rock | supportable |
| 1220 | vs | 1220 | vw | ---- | ---- |
| 1188 | sh | 1195 | vs | ---- | ---- |
| ---- | | 1165 | vw | ---- | ---- |
| 1155 | vs | 1120 | vw | C-O-C asym. str. | supportable |
| 1028 | m | 1030 | vw | ---- | ---- |
| 978 | w | 985 | vw | C-O-C sym. str. | supportable |
| 911 | m | ---- | | H-C=C-H twist | supportable |
| 878 | w | 885 | w | C-O-C sym. str. | supportable |
| ---- | | 860 | w | ---- | C=C band, not ether |
| ---- | | 790 | vw | ---- | ---- |
| ---- | | 745 | w | =C-H O.O.P. wag | no supporting evidence |
| 725 | m | ---- | | =C-H O.O.P. wag | supportable |
| 615 | s | ---- | | C-O-C i.p. bend | supportable |
| ---- | | 545 | w | O.O.P skel. def. | not apply |

Several frequency assignments have been made to carbon-hydrogen type vibrations or skeletal type deformations. Since we have no corresponding pattern recognition data to either support or contradict this type of assignment we have simply used the term " not apply" in the column: Evidence from Pattern Recognition.

There are two bands in cis dimethoxyethylene (Table XIV) which have been assigned by the authors for which there is no supporting evidence.

The frequencies around 1700 and 1675 $cm^{-1}$ (IR data) are stronger in the Raman than in the IR, and this is in agreement with what is observed in the broad representations of the IR and Raman of compounds with double bonds. The 1120 Raman band is not seen in the Raman representation of ethers, but we expect this since it was observed as a very weak band. The 978 IR band is observed as a weak band and is weak in the IR representation of ethers, while the corresponding 985 Raman band is very weak in the observed spectrum and missing in the broad representation of ethers. The 885 Raman band is seen in the appropriate Raman representation, but the corresponding weak 878 band is not in the IR representation. Both the 725 and 615 IR bands are in the correct representations, but they are weaker in the representations than would be expected based on their observed intensities.

around 1125 $cm^{-1}$ the IR in that region has a strong band. Thus we might have expected to have observed a corresponding IR band in the spectrum.

Conversly, the assigning of the 635 Raman band to an ether type vibration seems appropriate since the Raman of the broad representation of ethers has a weak band in that region and the representation of the C=C in the Raman has no band there. Additionally, the IR representation of ethers shows only a very weak band at 625 (compared to the corresponding Raman); so it is not unexpected that no 625 IR band was observed in the spectrum.

3. Divinyl Ether

Table XVII shows how the evidence from the broad representations of ethers and double bonds applies to divinyl ether. The spectral data were taken from Clague and Danti[41] and solution IR data were used rather than the vapor data. Unfortunately no data from neat samples were reported in the IR.

---

(41) A. D. H. Clague, and A. Danti, J. Mol. Spectrosc., No. 4, 22, 371 (1967).

No suggested assignment was made for the 1220 band, because both C=C and ethers vibrate in this region. It should be noted, however, that it is characteristic of ethers and not C=C to be strong in the IR and weak in the Raman near 1220 $cm^{-1}$. Similarly, the 1030-1028 band is in a region where double bonds are characterized by strong IR and weak Raman bands as observed; yet the interpretation is not clear due to overlapping ether bands. The very weak 790 Raman band corresponds to a band in the Raman of ethers and double bonds do not have a Raman band here, but the broad representation suggests that a strong band should appear. Since the observed band was very weak, no suggested assignment was made.

Table XV defines the abbreviations used in those Tables which describe frequency assignments.

4. trans-1,2-dimethoxyethylene

The data for trans-dimethoxyethylene were taken from the same source as the cis data[40] and are presented in Table XVI. Again the IR data used were the liquid data.

The evidence from pattern recognition pertaining to the trans molecule is similar to that pertaining to the cis-dimethoxyethylene. Again there are two assignments made by the authors for which there is no supporting

TABLE XV

Abbreviations Used in Describing Frequency Assignments

| Abbreviation | Meaning |
|---|---|
| vvw | very, very weak |
| vw | very weak |
| m | medium |
| s | strong |
| vs | very strong |
| sh | shoulder |
| str. | stretch |
| asym. | asymmetric |
| sym. | symmetric |
| def. | deformation |
| i.p. | inplane |
| o.o.p. | out of plane |

TABLE XVI

Interpretation of trans-1,2-Dimethoxyethylene

| Observed Frequency ($cm^{-1}$) | | | | Assignment | Evidence from Pattern Recognition |
|---|---|---|---|---|---|
| IR | | Raman | | | |
| 1737 | w | ---- | | ---- | C=C band, not ether |
| ---- | | 1688 | m | C=C str. | supportable |
| 1670 | m | 1675 | s | C=C str. | supportable |
| 1467 | m | 1460 | w | $CH_3$ asym. def. | not apply |
| 1455 | m | ---- | | $CH_3$ sym. def. | not apply |
| 1333 | w | 1330 | w | ---- | ether band, not C=C |
| 1310 | m | 1310 | m | =C-H i.p. rock | supportable |
| 1292 | m | ---- | | =C-H i.p. rock | no supporting evidence |
| 1218 | vs | 1215 | vw | C-O-C asym. str. | supportable |
| 1182 | m | ---- | | O-$CH_3$ rock | no supporting evidence |
| 1172 | vs | 1170 | vw | C-O-C asym. str. | supportable |
| ---- | | 1155 | vw | ---- | ---- |
| 1135 | vs | ---- | | ---- | ---- |
| ---- | | 1125 | vw | ---- | ether band, not C=C |
| ---- | | 1025 | vw | ---- | ---- |
| 995 | m | 995 | w | C-O-C sym. str. | supportable |
| 957 | s | ---- | | =C-H o.o.p. wag | supportable |
| 945 | m | 945 | w | =C-H o.o.p. wag | supportable |
| ---- | | 910 | w | C-O-C sym. str. | supportable |
| 900 | s | ---- | | H-C=C=H twist | supportable |
| 785 | w | 785 | m | ---- | ---- |
| 735 | vw | ---- | | ---- | ---- |
| ---- | | 635 | w | ---- | ether band, not C=C |
| 605 | w | 605 | w | ---- | ---- |

evidence from pattern recognition. This, of course, does not indicate that the assignments are wrong; and indeed, we do not think they are wrong. It simply indicates that we can not look to pattern recognition (as we have developed the data thus far) for support in making these assignments.

The 995 band is not seen in the Raman of the broad representation of double bonds, but this is in fairly good agreement with the observation that it is a medium strength band in the IR but weak in the Raman. The 945 Raman band is also not seen in the Raman representation, but it too is weak while its IR band has a medium intensity.

The weak 1737 band in the IR could be assigned to a C=C type vibration but not to an ether type vibration. We note, however, that in the broad representation of C=C in the IR, the band is not weak in the 1737 region. Similarly the 1333 (IR) band is seen in both the IR and Raman of the ether representations, but not in the IR or Raman of the double bond representations. The band strengths, however, do not correspond to the observed weak bands.

A band at 1125 is seen in the Raman representation of double bonds, and is not seen in the Raman representation of ethers. We refrained from suggesting an assignment, however, because the observed frequency is very weak, and because, although the representation in the Raman is weak we would have expected a strong corresponding IR band.

TABLE XVII

Interpretation of Divinyl Ether

| Observed Frequency ($cm^{-1}$) | | Assignment | Evidence from Pattern Recognition |
|---|---|---|---|
| IR | Raman | | |
| ---- | 1712 m | ---- | ---- |
| 1672 vw | 1670 s | ---- | ---- |
| 1641 s | 1643 vs | C=C str. | supportable |
| 1623 sh | ---- | ---- | ---- |
| 1619 vs | 1627 s | C=C str. | supportable |
| 1546 vw | 1545vvw | ---- | ---- |
| 1445 w | ---- | ---- | ---- |
| 1392vvw | 1400 vw | ---- | ---- |
| 1378 m | 1383 m | $CH_2$ sym. def. | supportable |
| 1330 m | 1323 vs | C-H i.p. rock | supportable |
| 1300 m | 1299 vs | C-H i.p. rock | supportable |
| 1200 vs | 1205 m | C-O-C asym. str. | supportable |
| 1168 vs | 1186 m | C-O-C asym. str. | supportable |
| 1111 m | 1116 w | $CH_2$ rock | no supporting evidence |
| 1012 m | ---- | ---- | ---- |
| 990 m | 993 w | C-O-C sym. str. | supportable |
| 961vvw | 965 vw | ---- | ---- |
| 944 m | 953 m | $H_2C$= torsion | supportable |
| 938 w | 940 w | $H_2C$= torsion | supportable |
| 848 s | 850 s | C-O-C sym. str. | supportable |
| 838 sh | ---- | $H_2C$= wag | supportable |
| 713wsh | ---- | vinyl wag o.o.p. | no supporting evidence |
| 688 m | 709 m | vinyl wag o.o.p. | supportable |
| 582 vw | 585 w | C=C-O def. | not apply |
| ---- | 510 s | C=C-O def. | not apply |

Two of the bands in the divinyl ether spectrum have been assigned to C=C-O deformations (585 and 510 in the Raman). Since they can not be classed as ether or double bond type assignments, they simply have been labeled "not apply" in Table XVII. The assignment for the weak shoulder in the IR at 713 is the only assigned band for which we do not have some supporting evidence.

In the original paper the 1445 and 1546 IR bands were not labeled with a description of intensity. We assumed the intensities reported for the vapor data.

The 1330 band in the IR is not seen in the representation of double bonds, but judging from the Raman the band might be centered a little lower. Between 1300 and 1325 our representations do show an IR band and a corresponding much stronger Raman band. For both the 1200 and 1168 (IR) bands we see an IR band in the broad representation but we do not see a corresponding Raman. While this is not entirely in agreement with the spectral observation for this assignment, at least the relative intensities (IR stronger that Raman) are in the proper direction.

The only truly weak point in the support that pattern recognition data has to offer with regard to these assignments concerns the assignment of the 990 (IR) to the C-O-C symmetrical mode. We see a small Raman band in this region, but no IR band at all in the pattern recognition

data. Assigning this mode to the 1012 vibration would be in better agreement with our patterns. The author rejects this based on band shape considerations, but the point must still be considered unresolved.

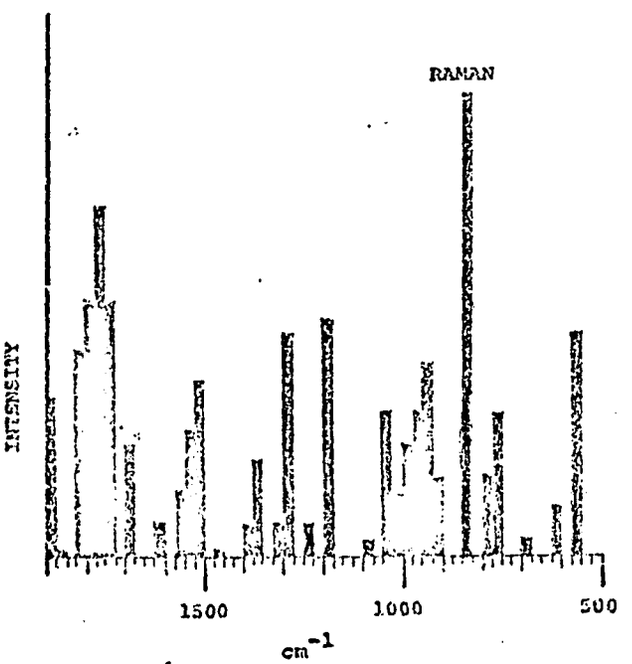
5. Broad Representation of Esters

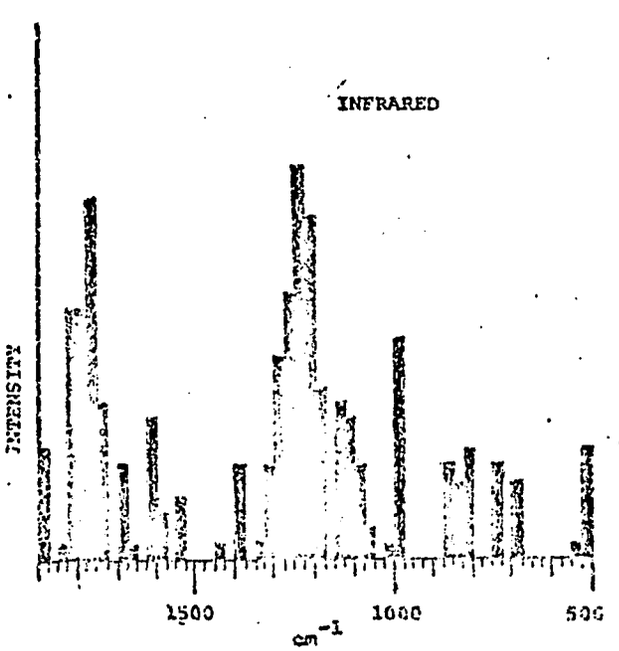
There are 45 alcohols in the data but alcohols as a class represent too broad a group to be useful in spectral assignments. If only primary alcohols are considered we are reduced to a class with only 29 members. This may be too few to be useful. We do have 92 esters, however, and the broad representations of esters may be of considerable use. Figure 24 shows these representations.

Considering only the infrared portion of Figure 24 we have recast the data presented to construct a table of characteristic frequencies for the ester moiety (Table XVIII). One might test the validity of such a table by examining the spectra of many molecules to determine how often the expected frequencies appear in molecules containing esters. More rigorously, however, it is desirable to compare the newly constructed table with frequency assignments which have been confirmed by normal coordinate analysis. Unfor-

FIGURE 24

Broad Representation of Esters



tunately, molecules containing C=O groups[42, 43, 44, 45] appear to present difficulties in all except the simplest cases. Matzke, Chacon and Andrade[46], however, completed the normal coordinate analysis of methyl formate, methyl acetate and dimethyl oxalate; and all of their assignments can be supported using either the IR or the Raman data from Figure 24. In Table XVIII the IR approximate descriptions which have been shown are taken from Matzke et al. For the 3 molecules considered, there are only 5 frequencies which do not correspond to a frequency range in Table XVIII. All 5, however, correspond to Raman frequencies (Figure 24) which are apparently characteristic of esters.

---

(42) J. R. Scherer, Spectrochim. Acta, 19,601 (1963); ibid. 20, 345(1964); ibid. 21, 321(1965); ibid. 23A, 1489(1967); ibid. 24A,747(1968).

(43) J. Overend and J. R. Scherer, Spectrochim. Acta 16, 773 (1960).

(44) J. Overend and J.R. Scherer, J. Chem. Phys. 32, 1296(1960).

(45) J. Overend, R.A. Nyquixt, J.C. Evans and W.J. Potts, Spectrochim. Acta 17, 1205 (1961).

(46) P. Matzke, C. Chacon, C. Andrade, J. Mol. Structure, 9, 255 (1971).

TABLE XVIII

Infrared Group Frequencies for Esters

| Range | Approximate Description |
|---|---|
| 1850 - 1725<br>1700 - 1650<br>1625 - 1575<br>1550 - 1525 | C=O stretch a,b,c |
| 1450 - 1425<br>1400 - 1375<br>1350 - 1175<br>1150 - 1050<br>1025 - 975 | C-O stretch a,b,c<br>O-R stretch a,b,c |
| 875 - 800 | O-R stretch, O-C=O bend, b,c |
| 750 - 725<br>700 - 675 | |
| 550 - 500 | |

a = methyl formate
b = methyl acetate
c = dimethyl oxylate

Critically reviewing Table XVIII, we see that if we were to join bands separated by only 25 $cm^{-1}$ (the limit of the resolution used in coding the data) we could distinguish 5 band areas as shown by the dashed lines in the tables. These may be considered areas where IR bands for esters, i.e. characteristic group frequencies, are likely to be found.

# HEURISTICS

## 1. Purpose

The purpose of this section is to include in the work certain observations and impressions which are peripheral to the main theme; but which, nevertheless, might be useful to future workers in this area. In the course of this investigation over a hundred computer experiments were run which did not contribute directly to the body of data already presented, but rather directed us to abandon certain paths and follow others. What follows draws on that experience without attempting to document and quantify each observation. Indeed, many of the observations are not able to be proven; hence the title of the section, Heuristics.

## 2. Starting Vector

We made several computer runs in which we attempted to use the Muted Average representation of a class (both the positive and negative portion) as the starting vector. In general the program does converge slightly faster when the vector is intitiated in this way, but the savings was only about 5 iterations. In a few instances it actually took more iterations to converge than when initiated at -1. In none of the 7 cases tried were we able to get a set of Raman data to converge within 100 iterations starting with the appropriate Muted Average, if that same set would not

converge starting with a vector initiated at -1.

One of the options in the program permits us to train a vector using every "n th" element in the data bank. This vector is then used as the starting vector and the entire data set (using every element) is trained. With the infrared spectra of esters as the class to be trained, we tried to train a vector using every 10th element. The vector would not train within 100 iterations. Nevertheless, the program takes the vector "as is" at 100 iterations and uses it as the initial vector to train the entire set. When it did this it trained the entire set of esters in 62 iterations. When this set of data is trained using a vector initialized at -1, it takes only 40 iterations to converge. Similar results were obtained when we tried every 5th element. We abandoned the idea without trying other cases.

Using the trained infrared vector as the starting vector for training the same class of compounds with Raman data did not improve the rate of convergence. Nor did it permit us to train Raman sets which would not converge when initiated at -1.

3. Weighting Factor

In the training process the vector being trained is corrected by adding or subtracting a weighted pattern which failed to be correctly classified. The program has an

option which permits the weight factor to be adjusted by a multiplier which is read into the program at the beginning of each run. (See Figure VI.) Jurs, Kowalski, Isenhour and Reilley[18], using data from mass spectrometry, reported that multipliers ranging from 1 to 3 seemed to minimize the number of iterations necessary to reach convergence. We did not repeat their work with our data, but simply used the value 2 throughout the reported runs. We did verify, however, that changing the multiplier does not alter the appearance of the final vector any more than does changing the initial value of the starting vector.

4. Transforms

Two additional data sets were formed by taking the log (base 10) and the anti-log (base 10) of the original data. All of the runs which were done on the original data set were repeated on the two transformed sets in an effort to get certain cases to converge (within 100 iterations). We used the reciprocal transform on the resulting vector and plotted the result in the usual way. With the class of ethers and using perpenducular polarized Raman data we could train a vector using the log set. These data would not train using the untransformed data. In every other case, however, data which would not train without transformation remained untrainable.

In those cases where data sets were trainable without transforms, they remained trainable using either transformed data set. The resulting vectors were substantially different from what resulted from untransformed data, however. We did not attempt to interpret these vectors and did not pursue this, because we found that in every class concatenated data converge very quickly without the need of transforms. There remains here an opportunity for further work.

5. Size of Data Set

A number of workers[2, 18] have suggested that data sets should be over determined by a factor of 10, i.e. there should be 10 times as many patterns in a training set as elements in a pattern. This is rarely done with chemical data, however, because of the scarcity of coded data. Liddell and Jurs[12], for example use as few as 22 patterns in a set with 131 elements, and report that the resulting vector is 90% predictive with patterns not in their training set.

Our experience suggests that the data should be over determined by at least a factor of one. In developing the reported vectors using concatenated data we were always over determined by a factor of 3.5 (400 patterns, 112 elements per pattern). In other trials we found that predictive ability deteriorates as the data set gets smaller (see Figure 11) and falls below 90% predictive when the data is over deter-

mined by a factor of only 1.75.

Ideally, the number of compounds in the class and the number not in the class being treated should both be over determined. In no case did we have enough compounds to do this and still use full spectra. Nagy[1] suggests that the percentage of patterns in the training set which are members of the class being trained should be the same as the ratio in the real world, i.e. should reflect the universal population. Using spectral data such an approach seems impractical, however, since we have no way of estimating such a number nor attaching significance to its meaning.

We found that when the number of compounds in the class was extremely low compared to the number in the training set, the set would always train. Even when we had a set composed of two class members and 398 compounds not in the class, a vector could be trained. Such vectors are useless, however. Our experience suggests that at least 5% of the total set should be members of the class being trained.

6. Errors

In any situation where over 125,000 data points are observed and recorded manually there are bound to be errors. Every effort was made, however, to minize the errors. All of the data keypunching was verified, for example. In order

to estimate the magnitude of the errors we recoded 200 spectra. The second coding was not done by the person who first coded the spectrum. In comparing the two codes of the same spectrum we considered a pair in error if they did not match in every element within 3. Using this criteria we found 6 errors in the original 200 spectra (and another 4 errors in the second coding of the 200). Thus we would estimate that in the 1117 coded spectra our error level is about 3%.

After making the original estimate of error lever, we continued for some time to comb the data. The training program prints a list of the compounds used and the number of times each failed to be correctly classified during the training process. In the early runs we made it a practice to review the coding of the 10 compounds which failed most often. When errors were found they were corrected. We did not return to make a second estimate of the error level, however. When the final runs were made for inclusion in this work, they were all made using the same data set. No changes were permitted once it was decided to develop the results presented.

## SUMMARY AND CONCLUSIONS

We have taken a segment of the Infrared and Raman data which are currently available and manually reduced each spectrum to a digital code. We then explored these data using pattern recognition techniques. We found that the most effective way of treating the data was to consider the infrared code and the Raman code of a compound as a single concatenated vector. In order to save computer space we used the spectral range from 500 to 1900 $cm^{-1}$ only, thus covering the most complicated portion of the vibrational spectrum.

The vectors which result from the training process can be used to categorize compounds not used in the training and they will correctly categorize approximately 93 percent of the time. This is superior to what can be achieved using either infrared or Raman alone, even though we did not use the full vibrational spectrum as we did when using IR or Raman alone.

The positive portion of the trained vector provides new insight into the characteristic frequencies of chemical groups. In every instance tried, all of the normally reported IR group frequencies appear as positive bands in the trained vector of the group. Other bands suggest the presence of characteristic group frequencies not previously

assigned. A band from 950 to 800 $cm^{-1}$ in the Raman, for instance, seems to be characteristic of esters. Usually the corresponding IR band is weak or absent.

It is well known that ethers have a characteristic broad band around 1250 $cm^{-1}$ in the IR which is greatly reduced or absent in the Raman. Our data suggests (based on the 15 ketones available) the ketones have a Raman band in that region, although it is not a strong band.

The IR band around 1050 to 875 $cm^{-1}$ is usually assigned to double bonds (vinyl, trans or geminal), but the band is not easily picked out in a crowded spectrum. The knowledge that the band is generally weak or absent in the Raman is most helpful.

Three spectra not present in the coded data were selected from the literature, because they have been partially interpreted, and contain two groups for which we have trained vectors. It was shown that with nearly every assignment supportive evidence for the assignment could be found in the trained vectors. What is perhaps a greater contribution, however, is that using the trained vectors we were able, in a few instances, to suggest which group might be responsible for certain observed frequencies. Thus advancing the problem of making band assignments.

In short, what we have done here is to explore an orderly, computerized approach to the interpretation of IR and Raman. In a few years Raman spectroscopy will be as common as infrared, and much can be learned from this approach to those data. What is needed is a method for digitizing the spectra so that the data being explored is error free. What we hope we have contributed, is that we have made a begining.

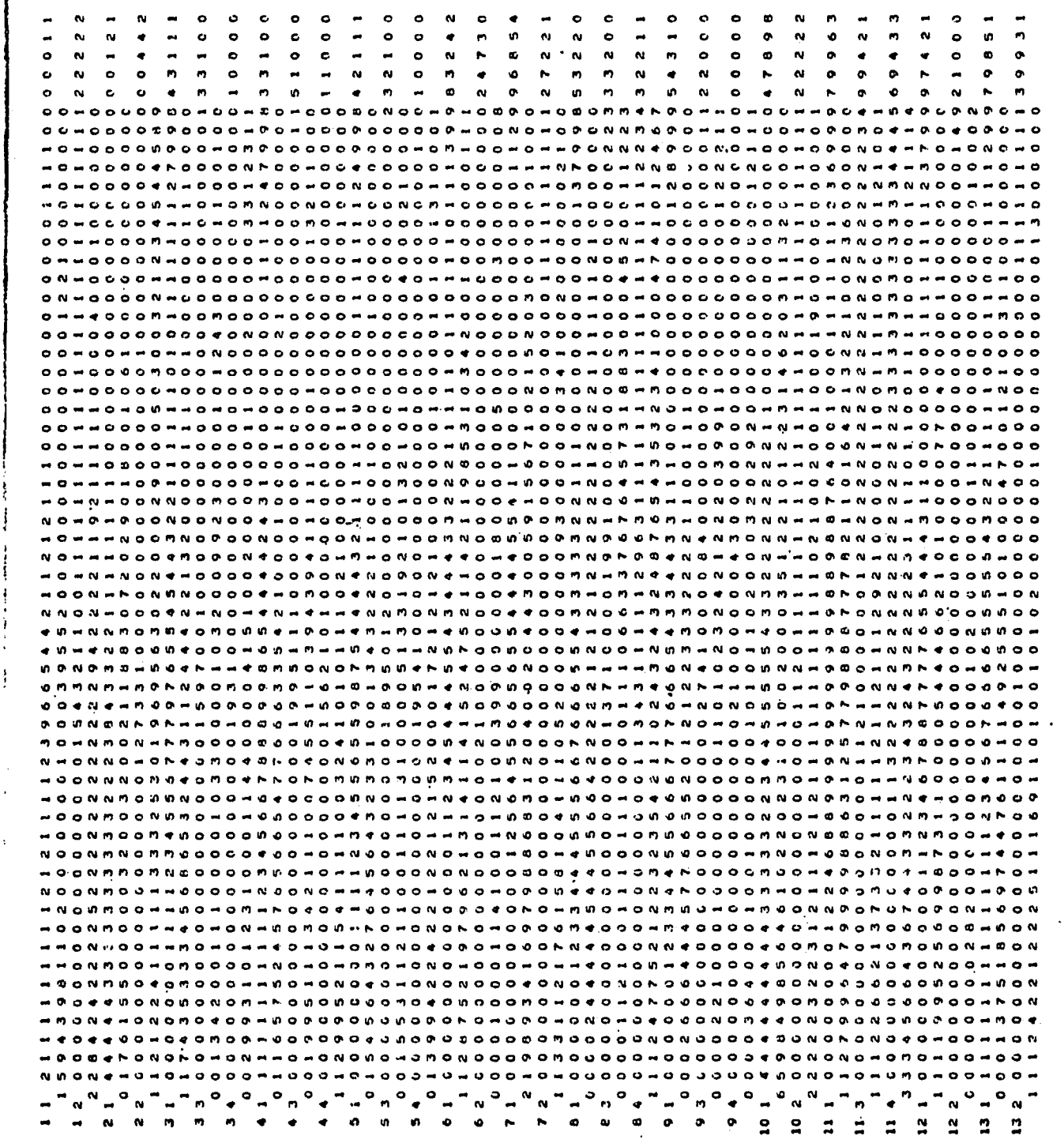
# APPENDIX A

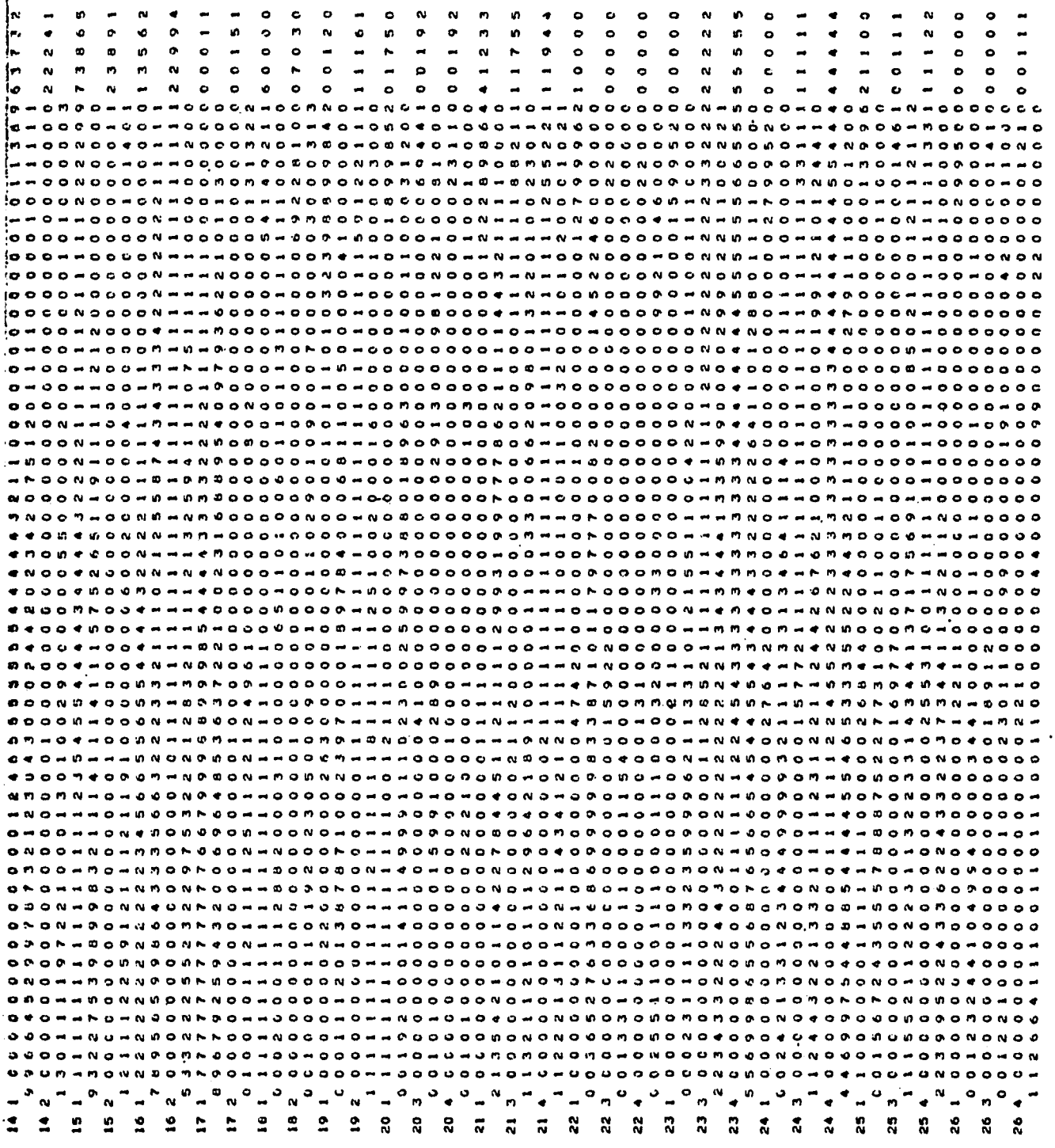
# THE DATA

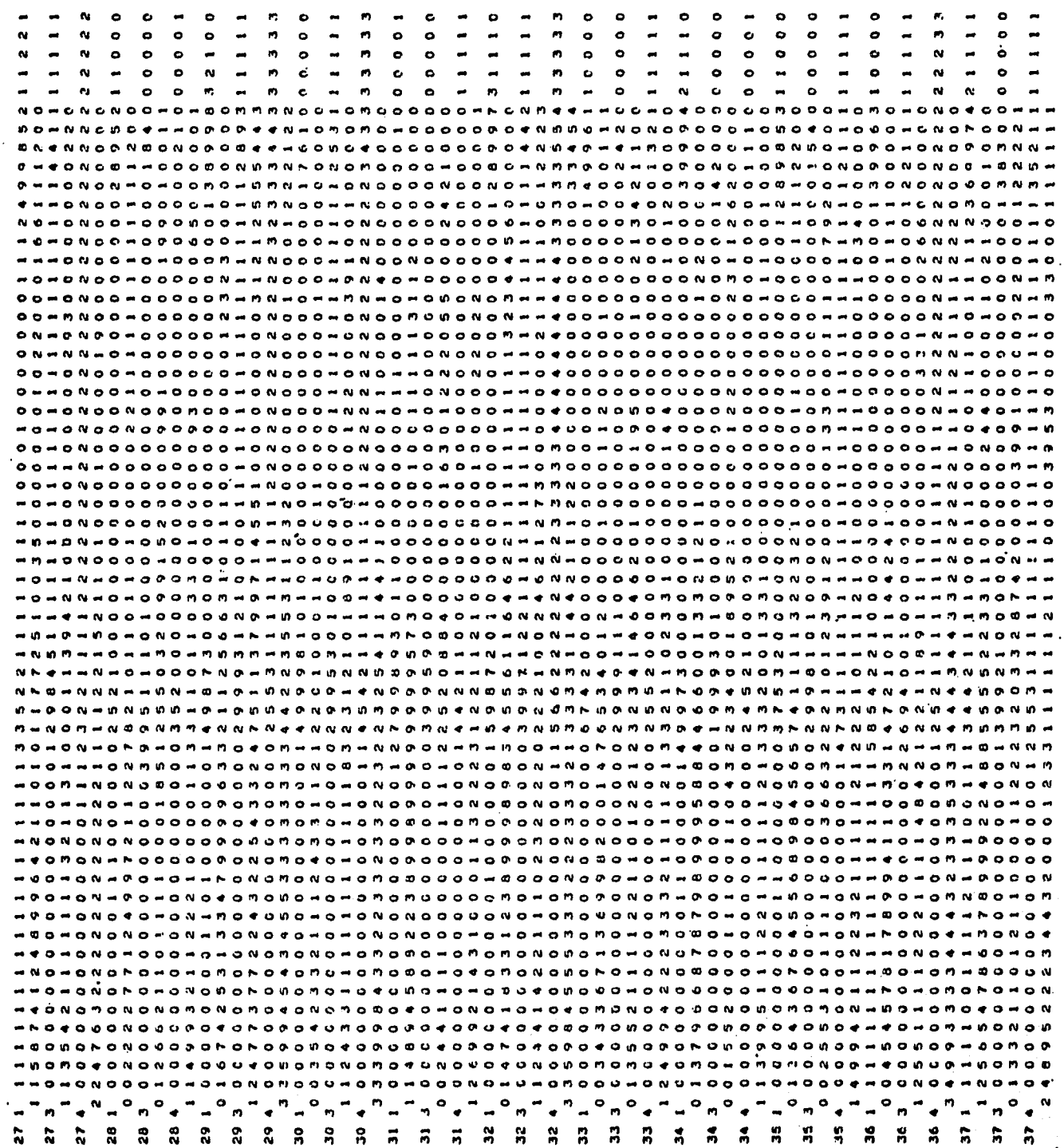
The digital codes of spectra are shown in the following pages. Each code uses the following format:

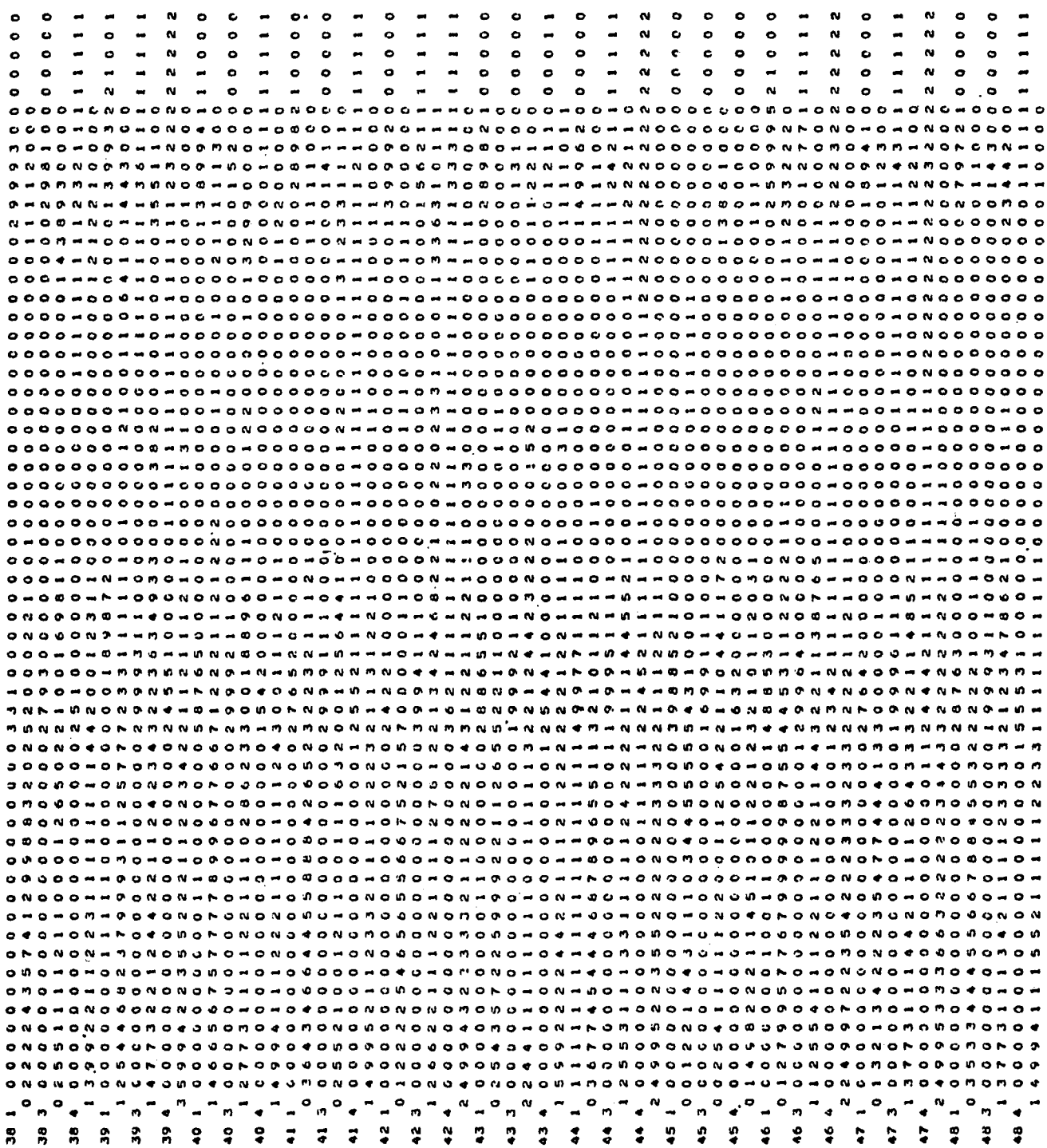
(I3,I2,1x,58I2/4x,54I2)

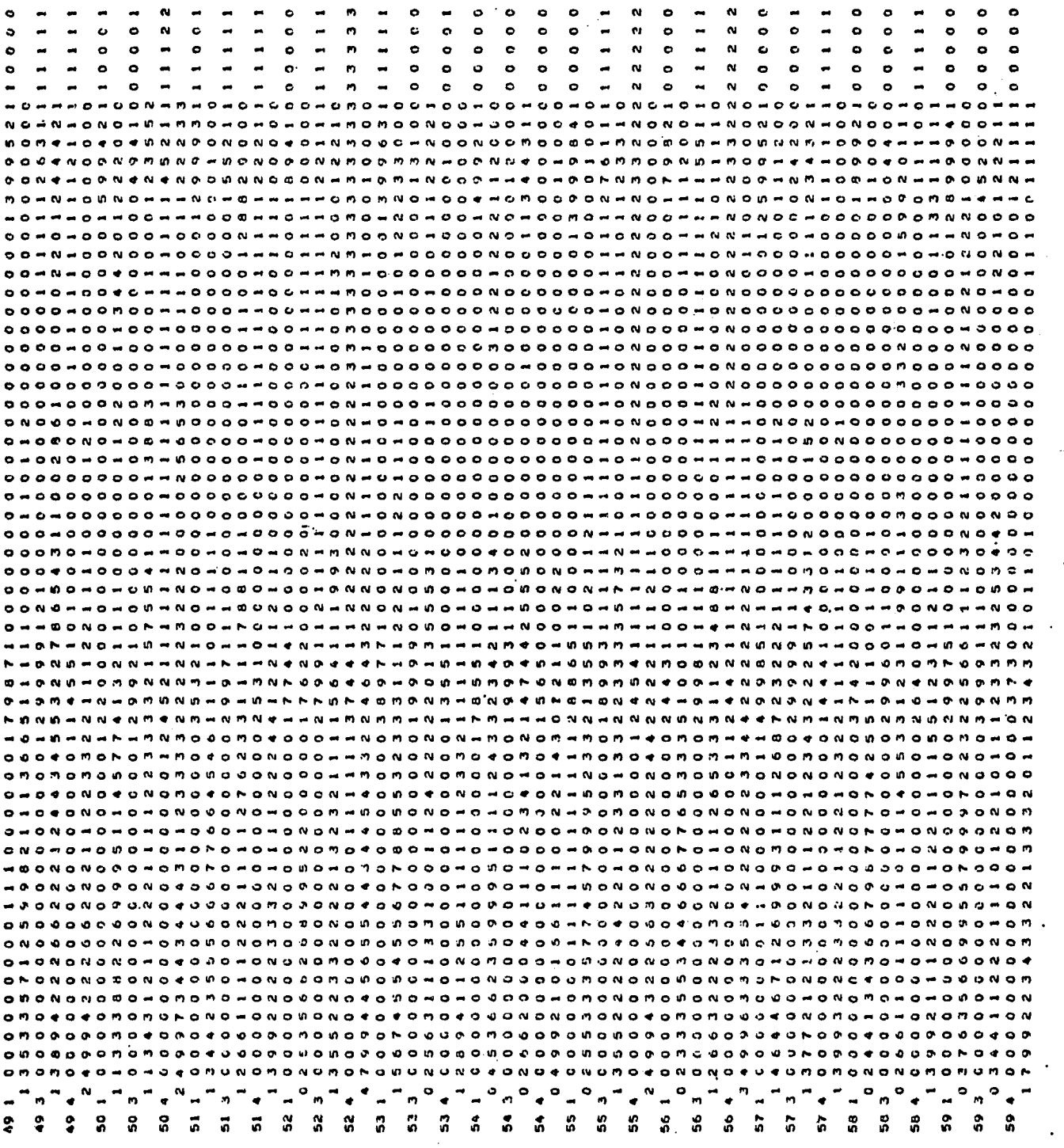
The first intiger (I3) is the compound number, and corresponds to the compound number in the Sadtler Index. The second intiger (I2) is the spectrum type: 1 = IR, 2 = Raman parallel polarized, 3 = Raman perpendicularly polarized, 4 = Raman nonpolarized. The 112 readings from each spectrum are in I2 format.

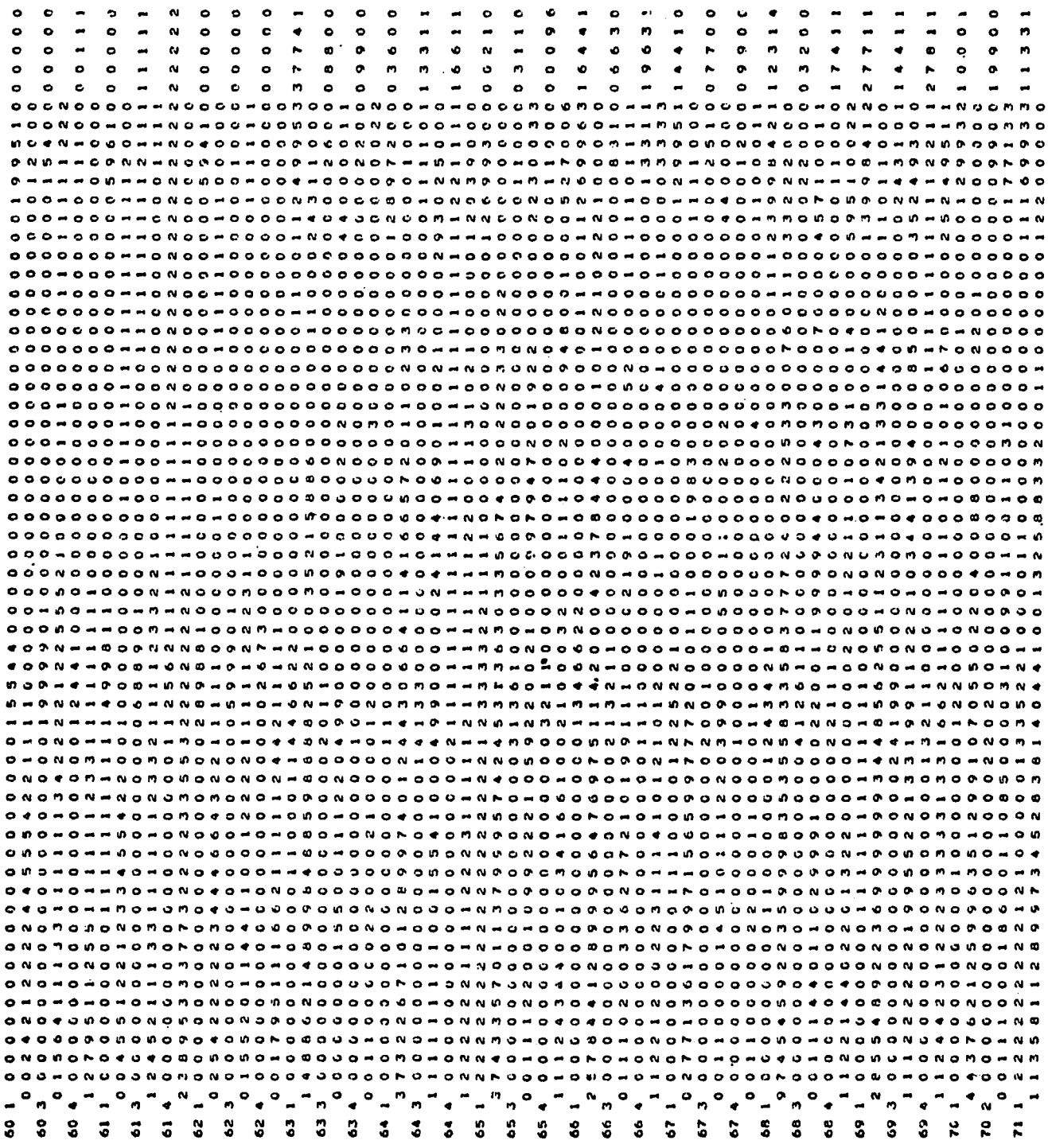


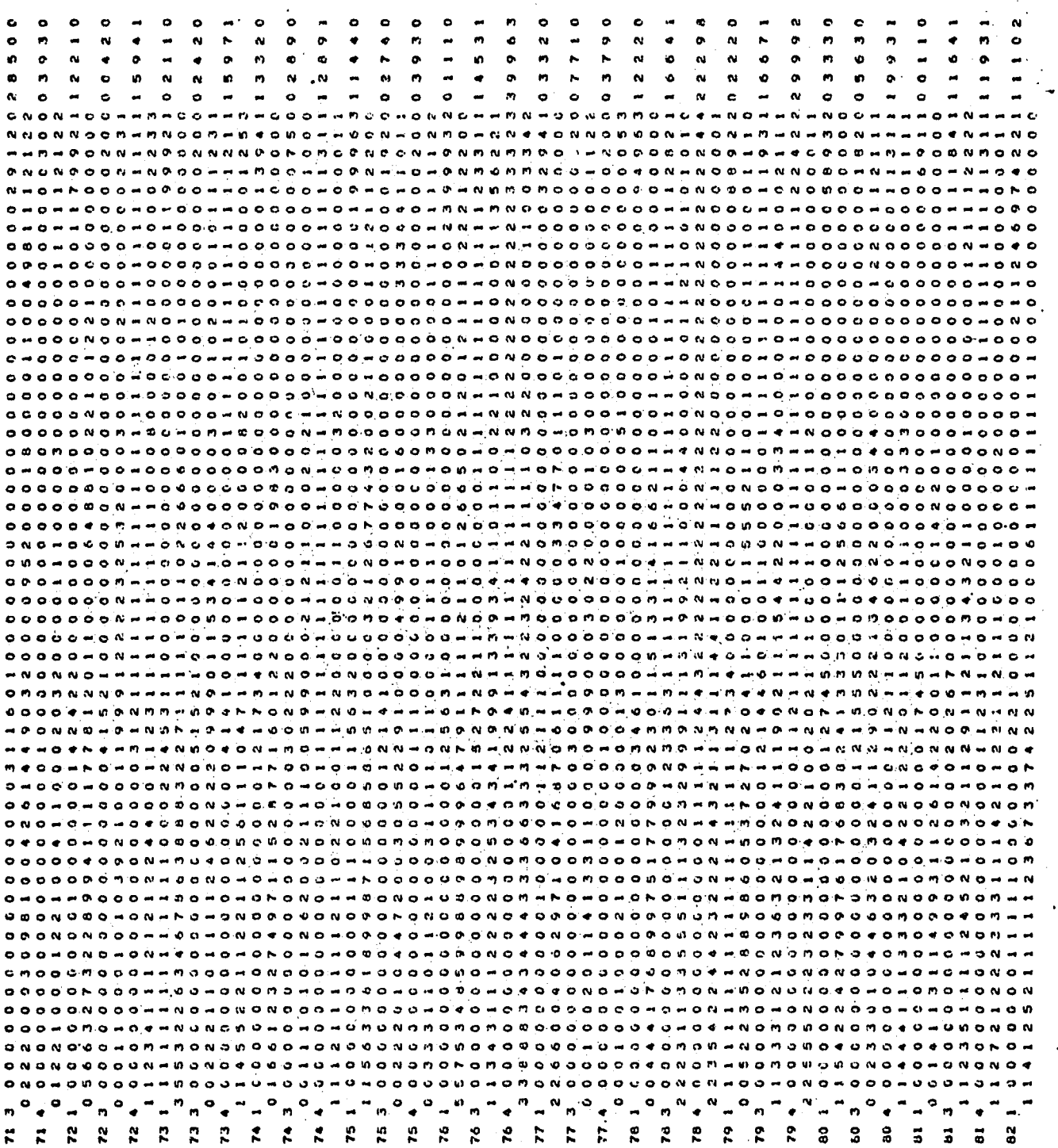


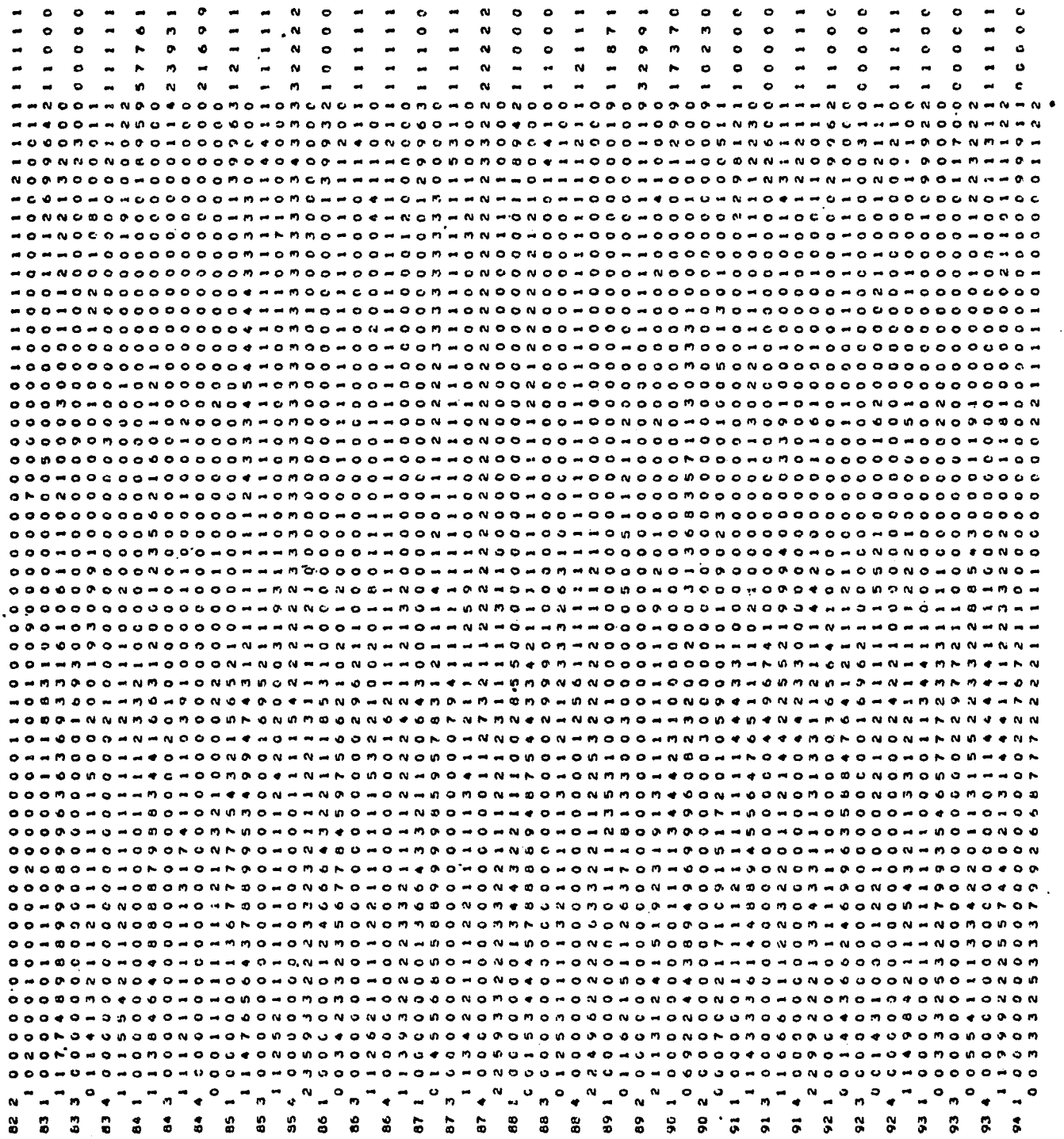


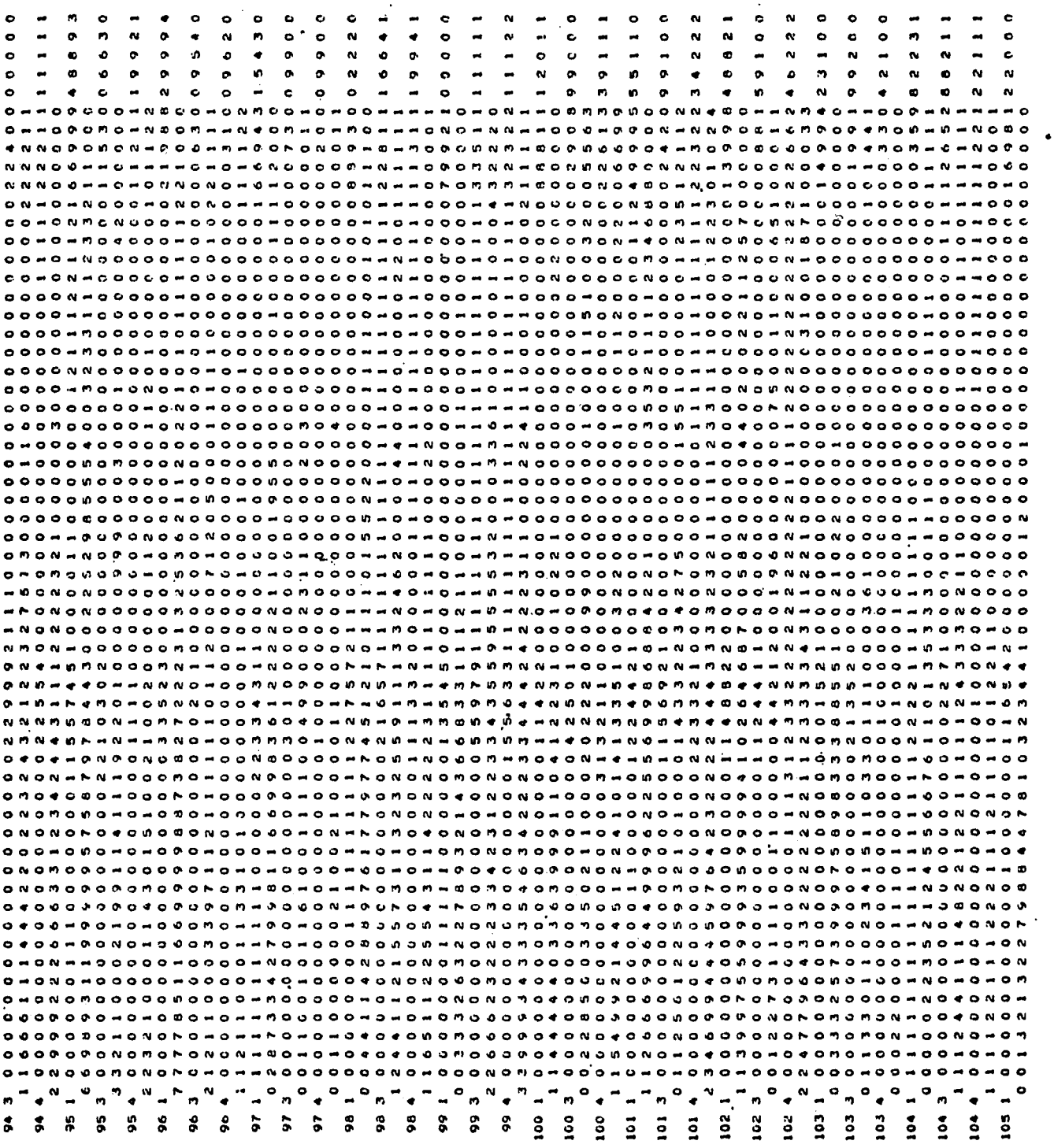


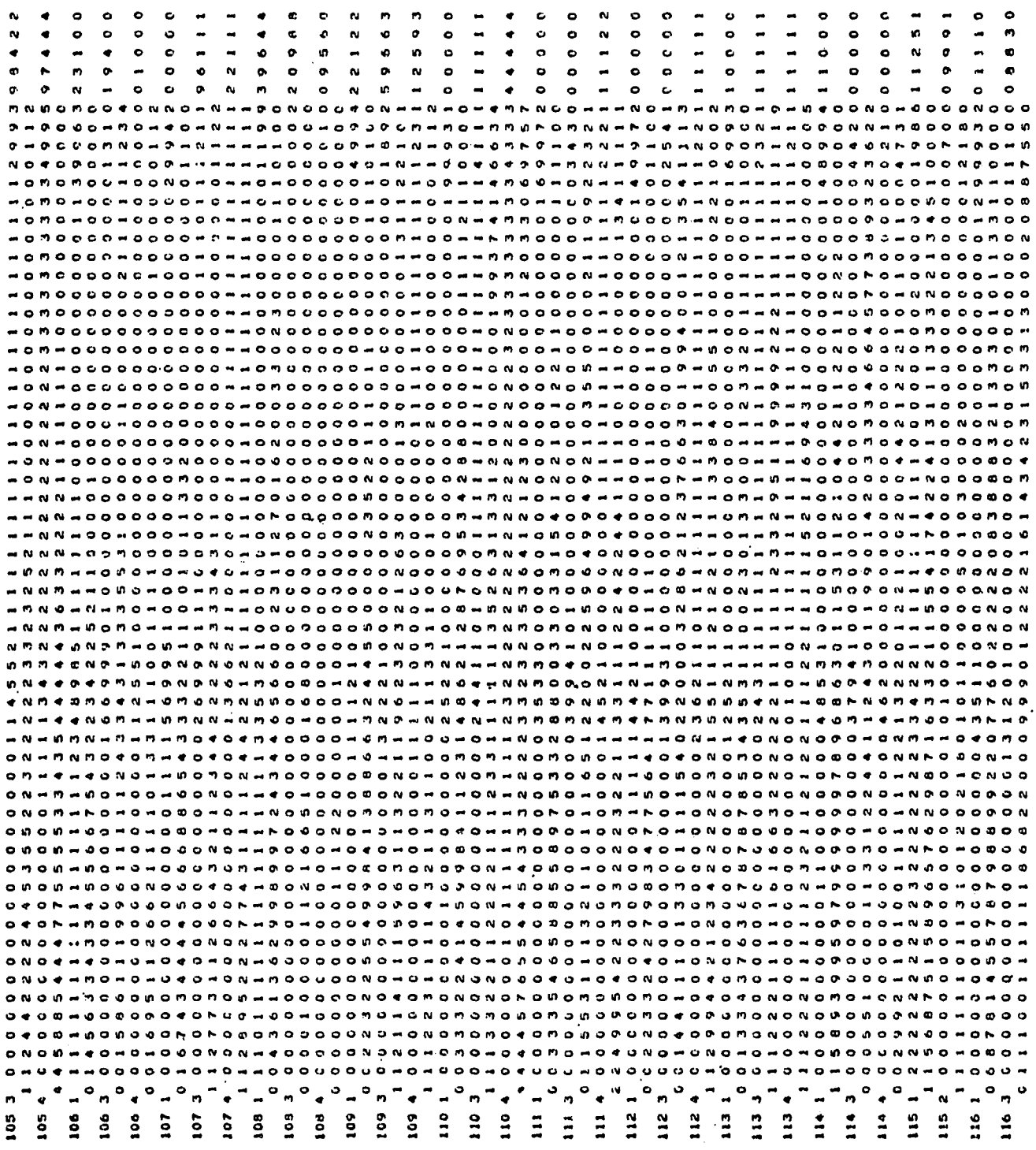


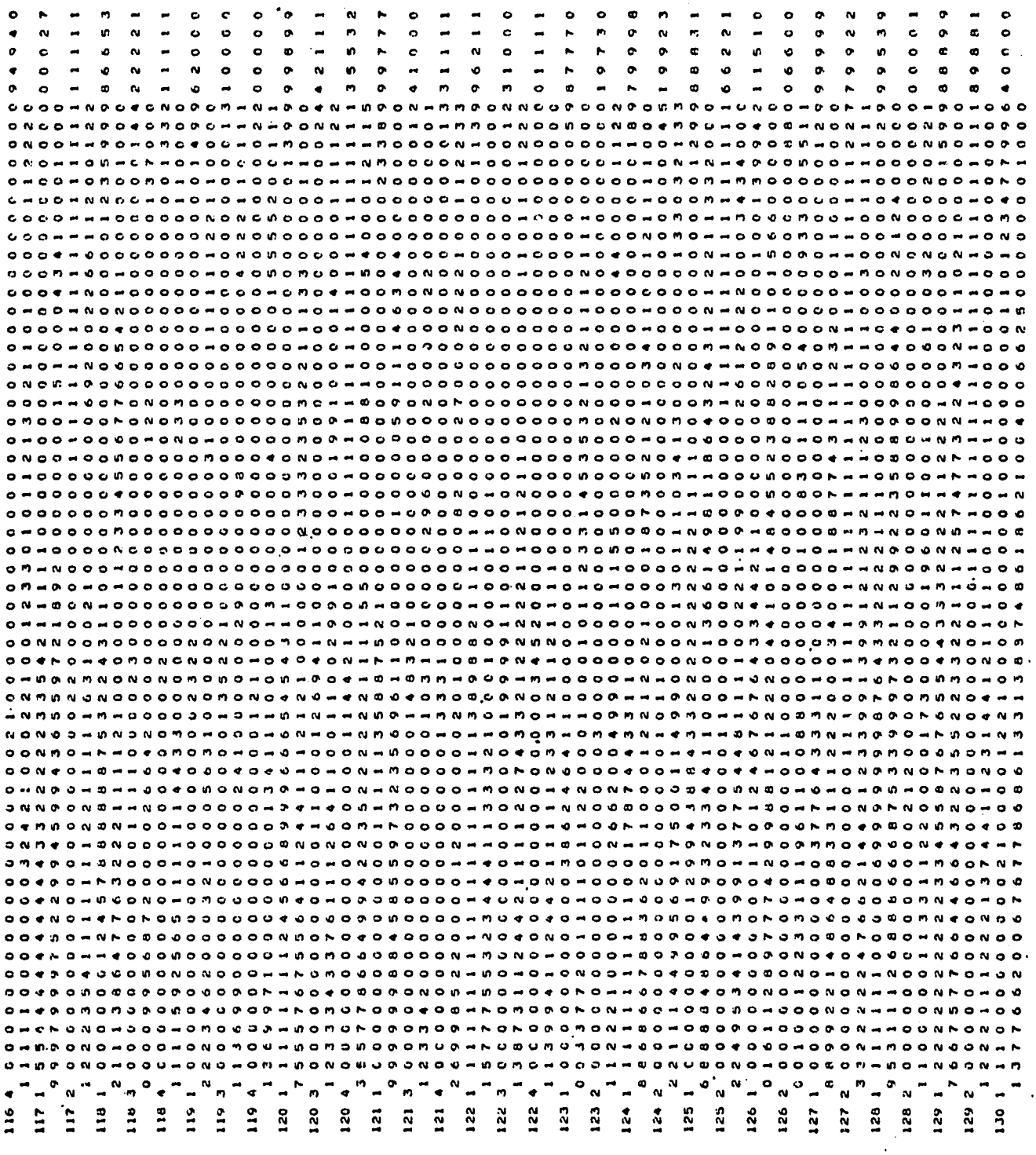


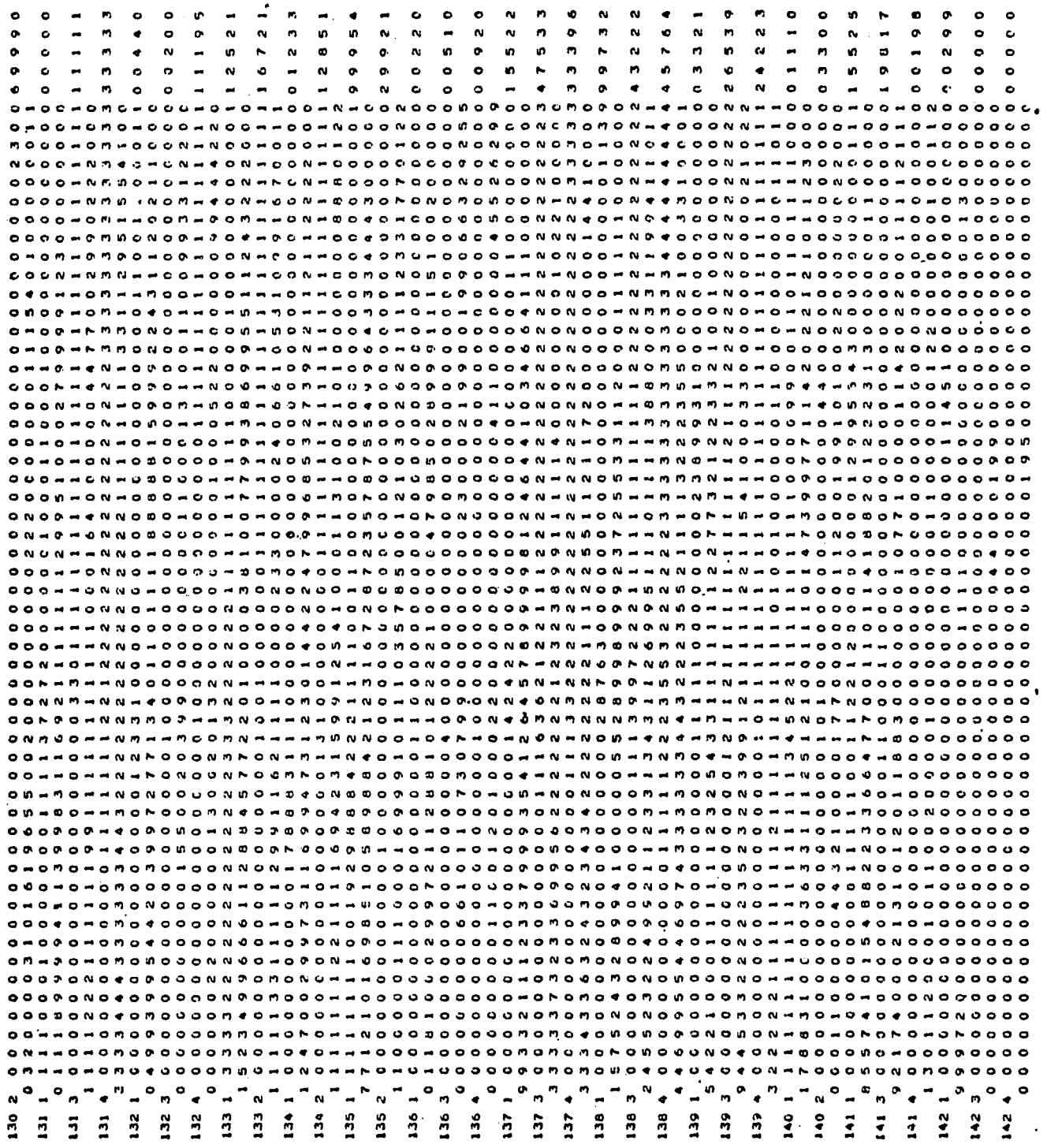


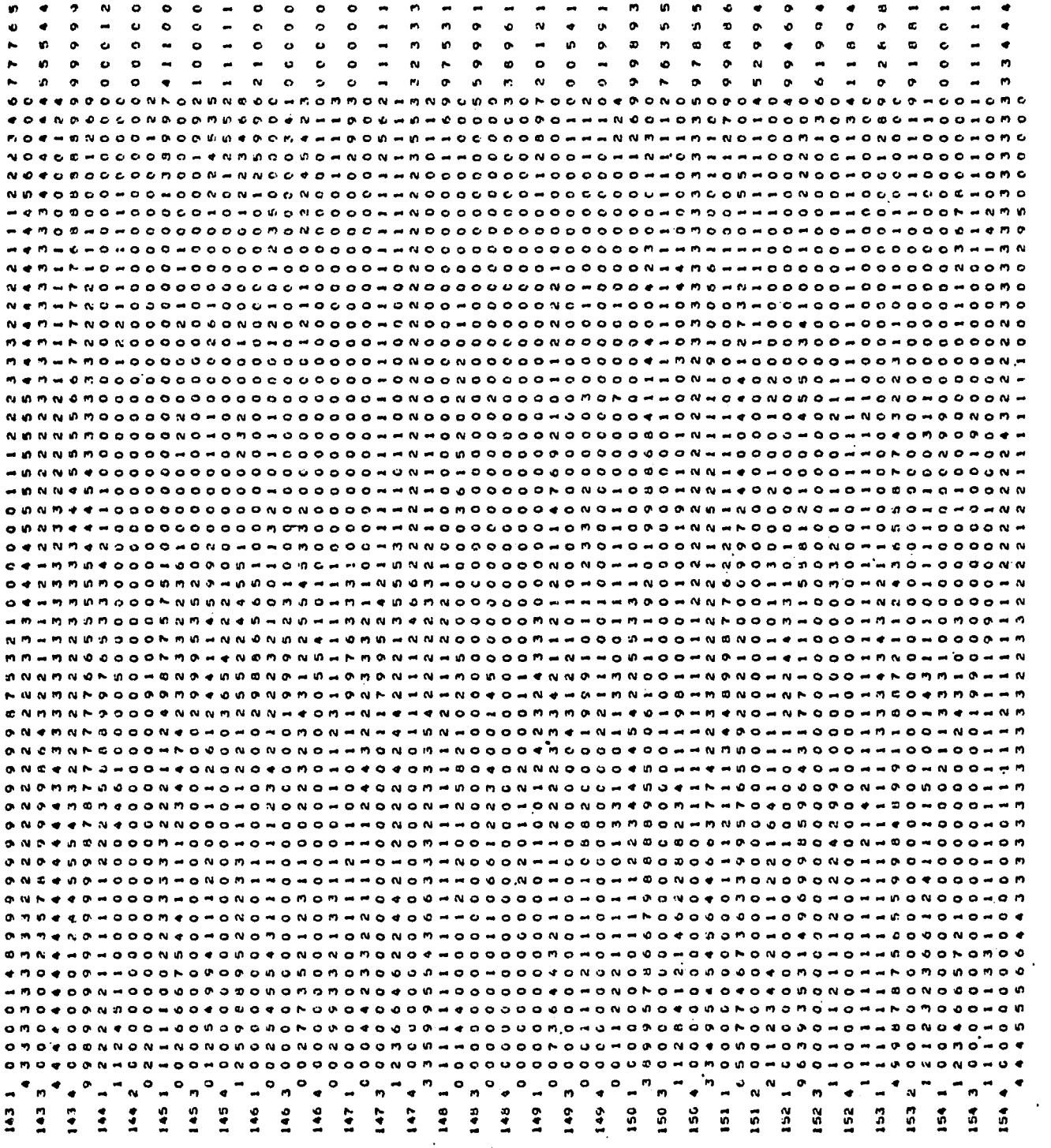


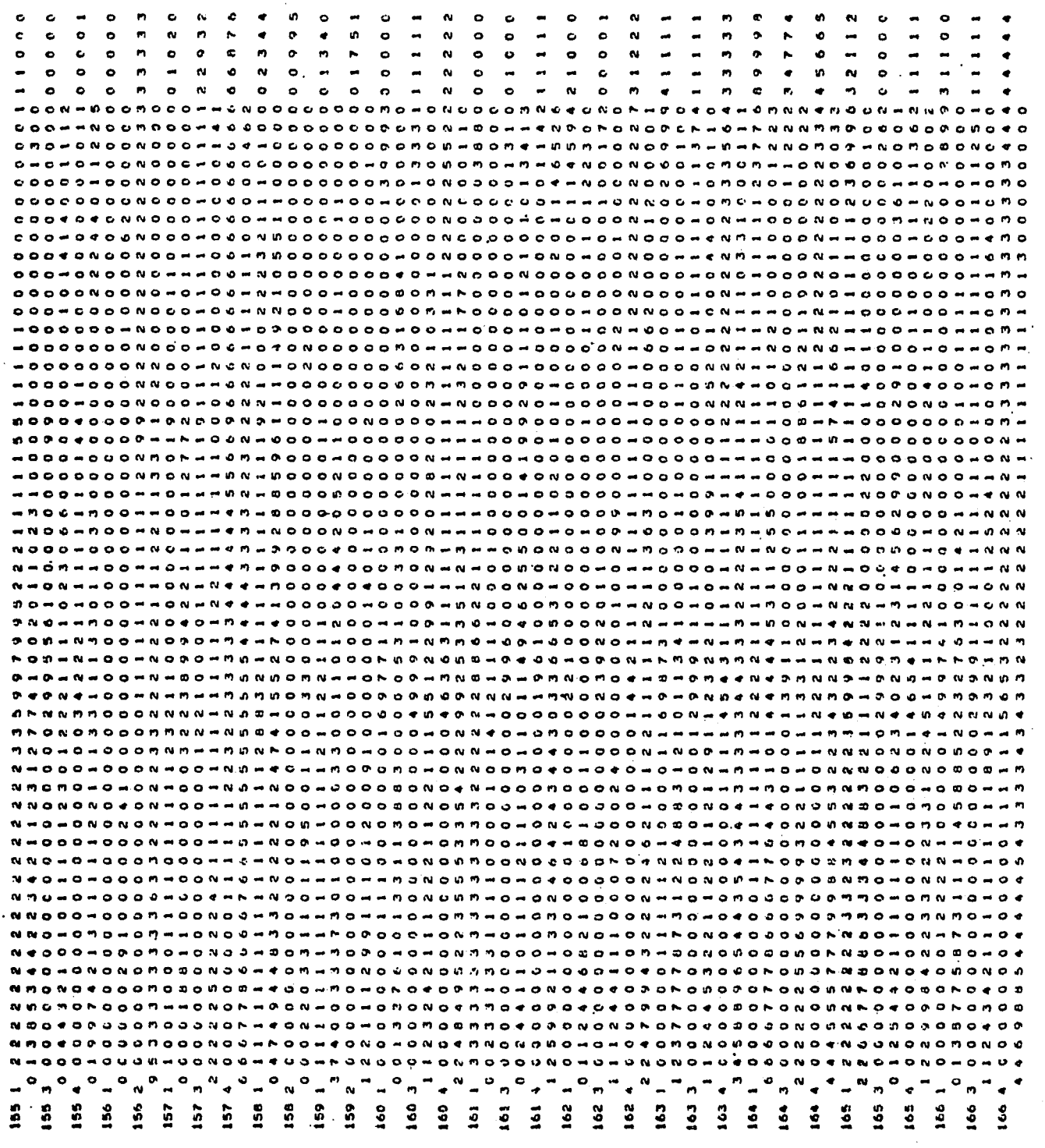


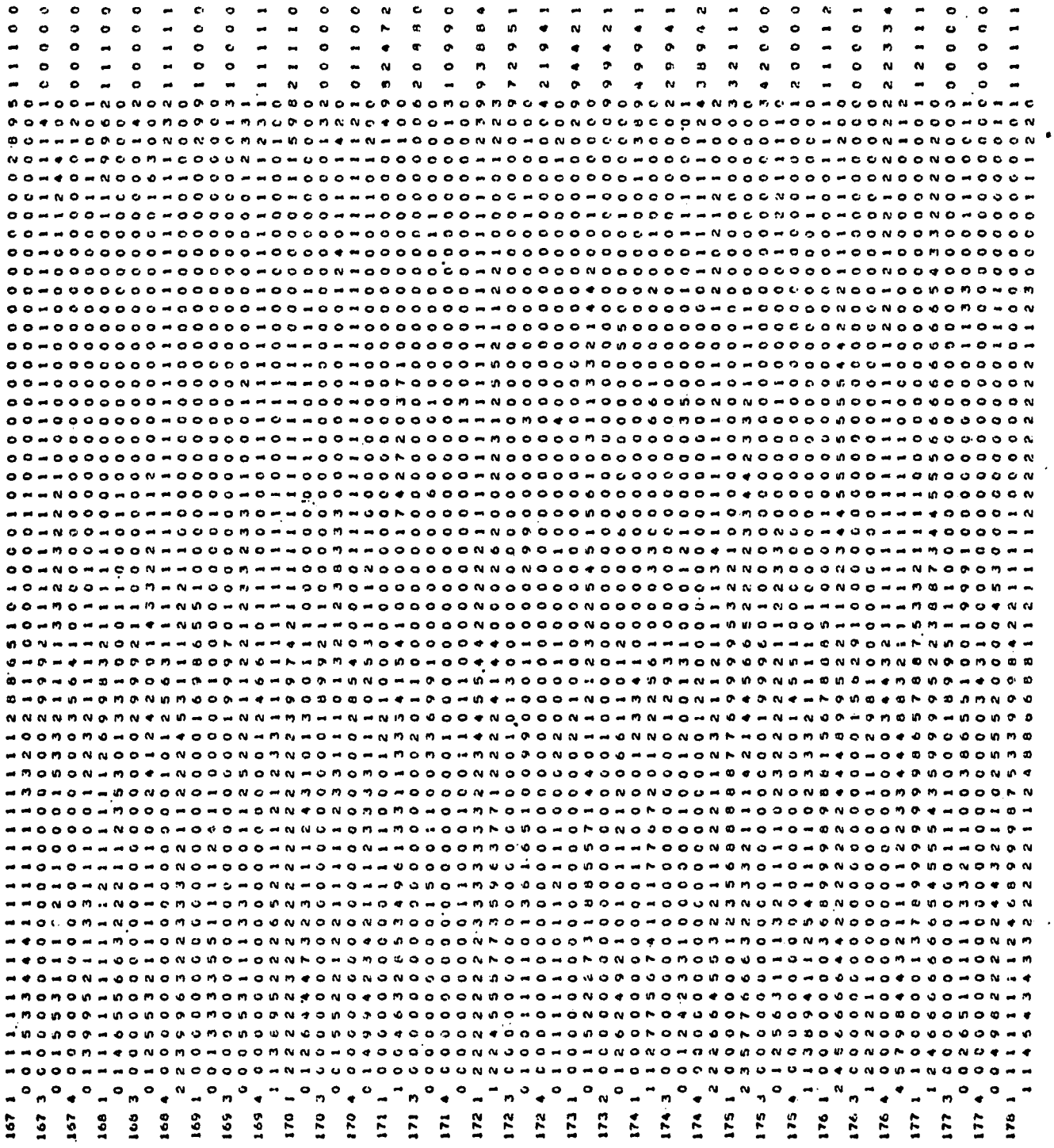


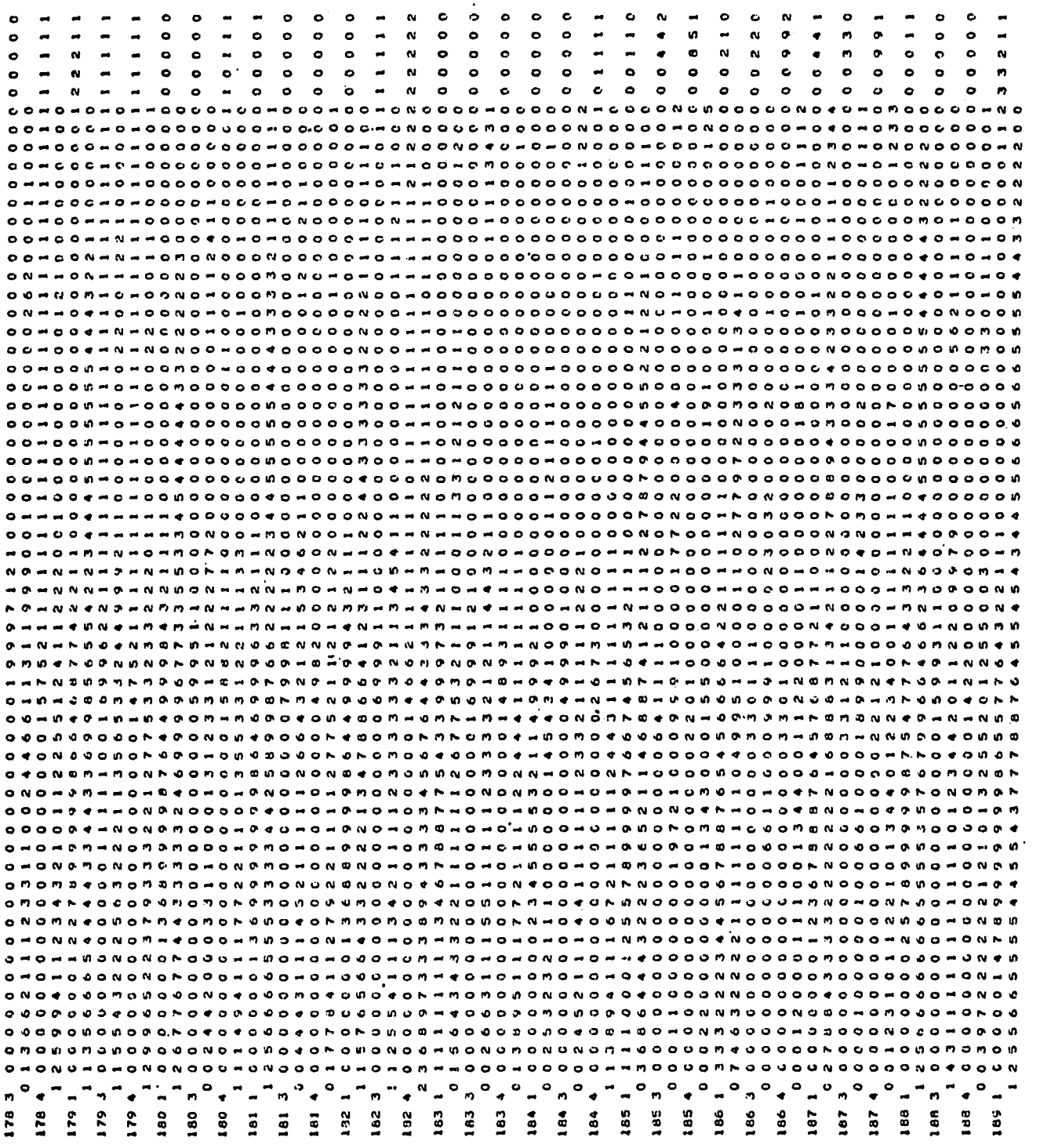


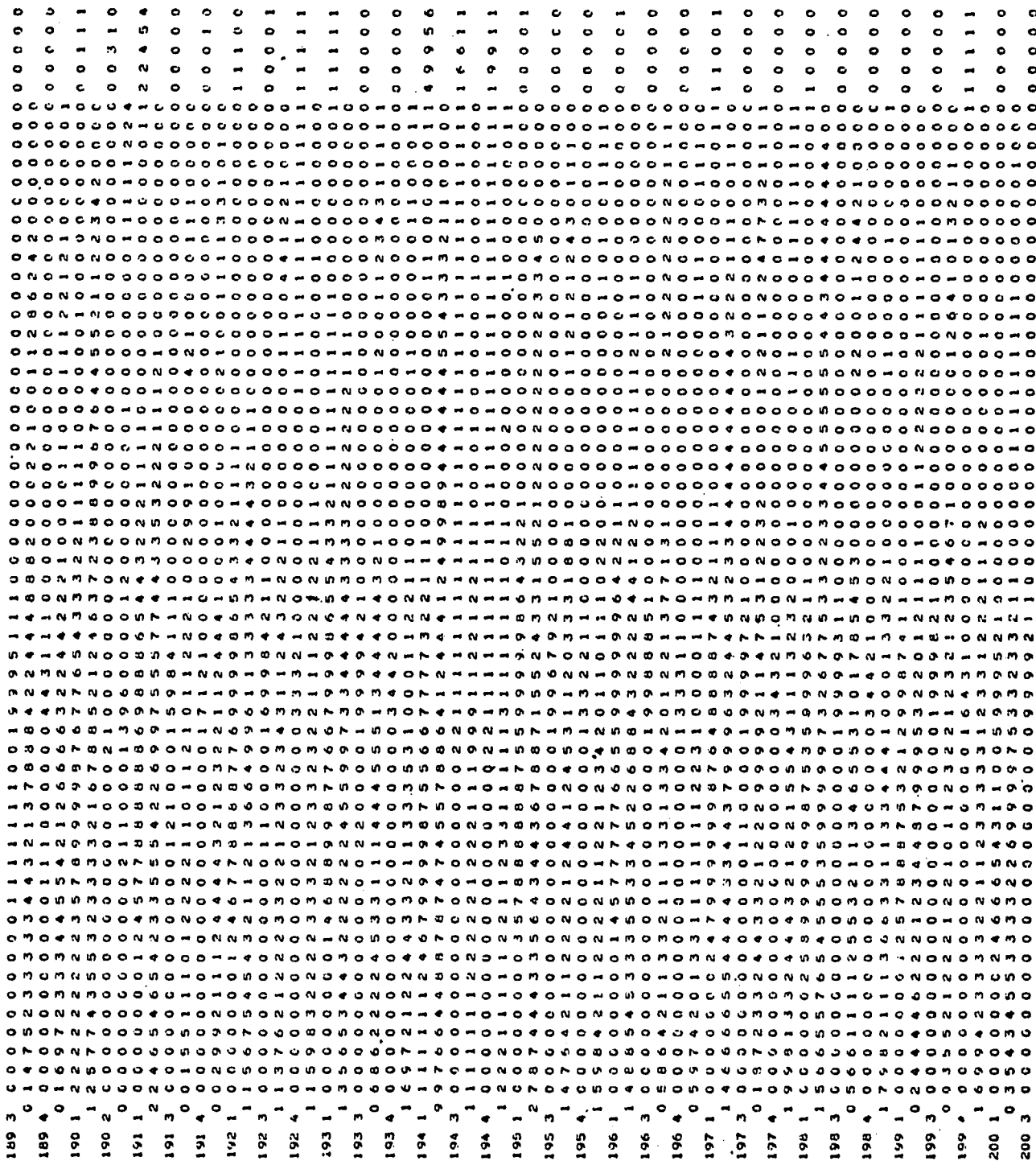


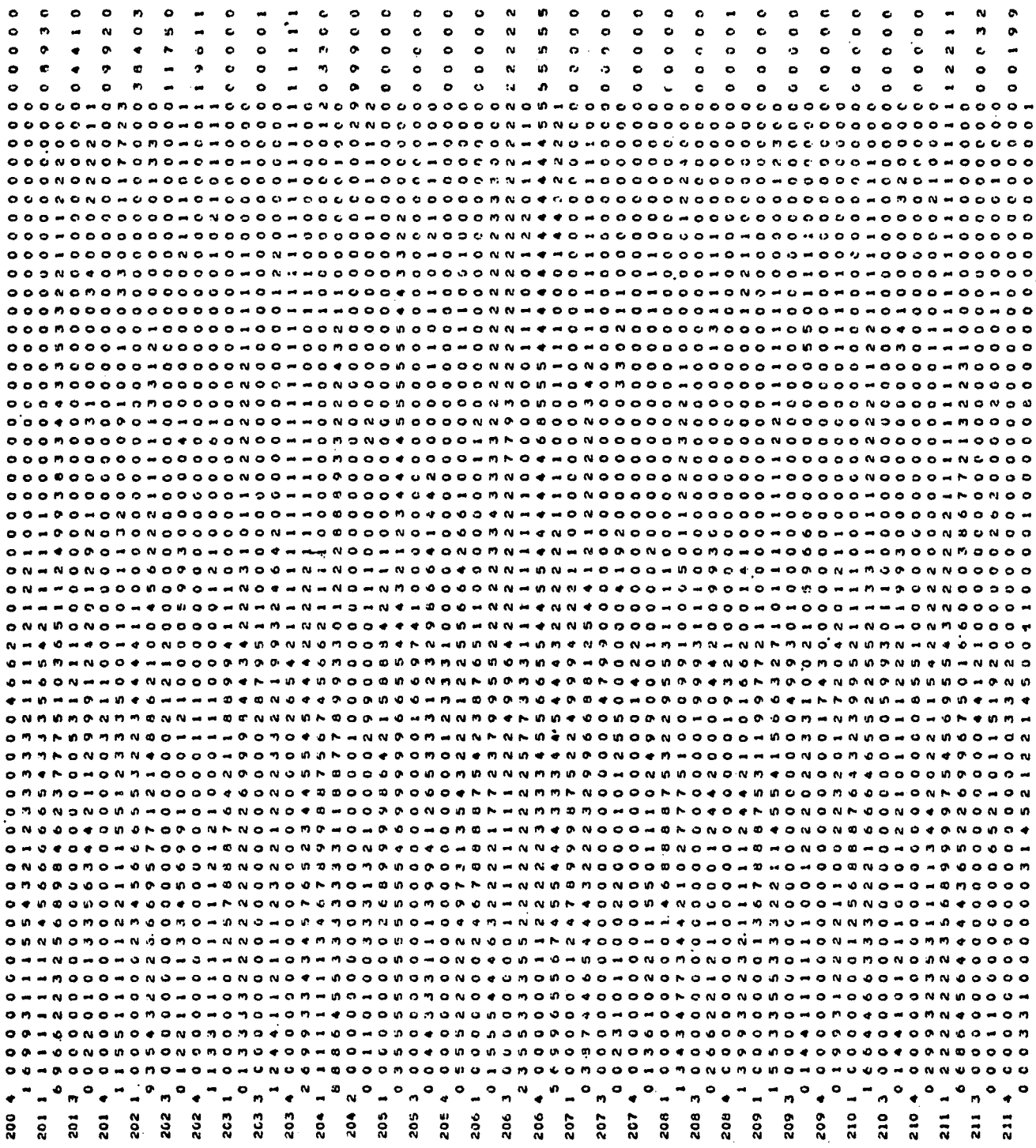


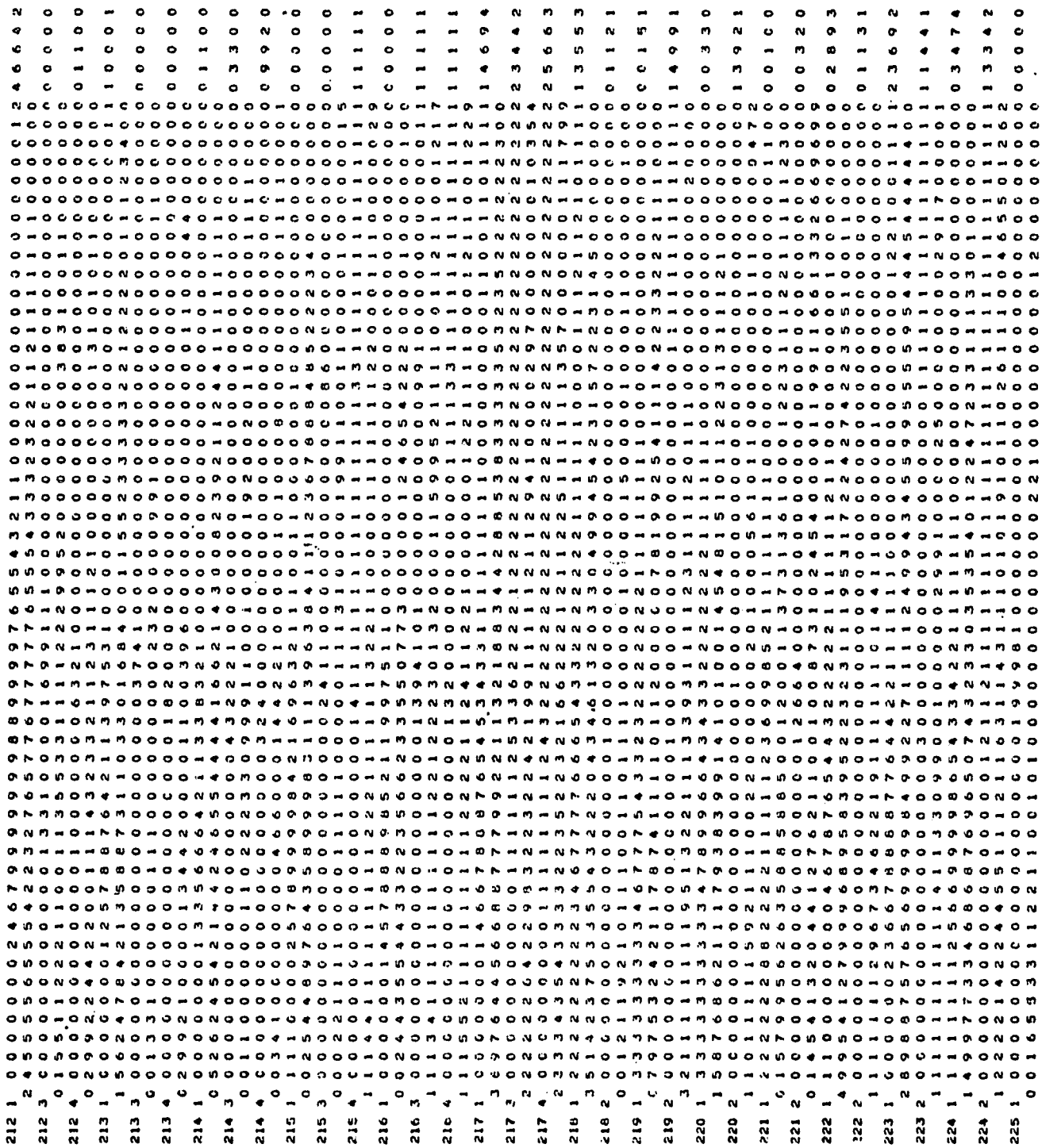


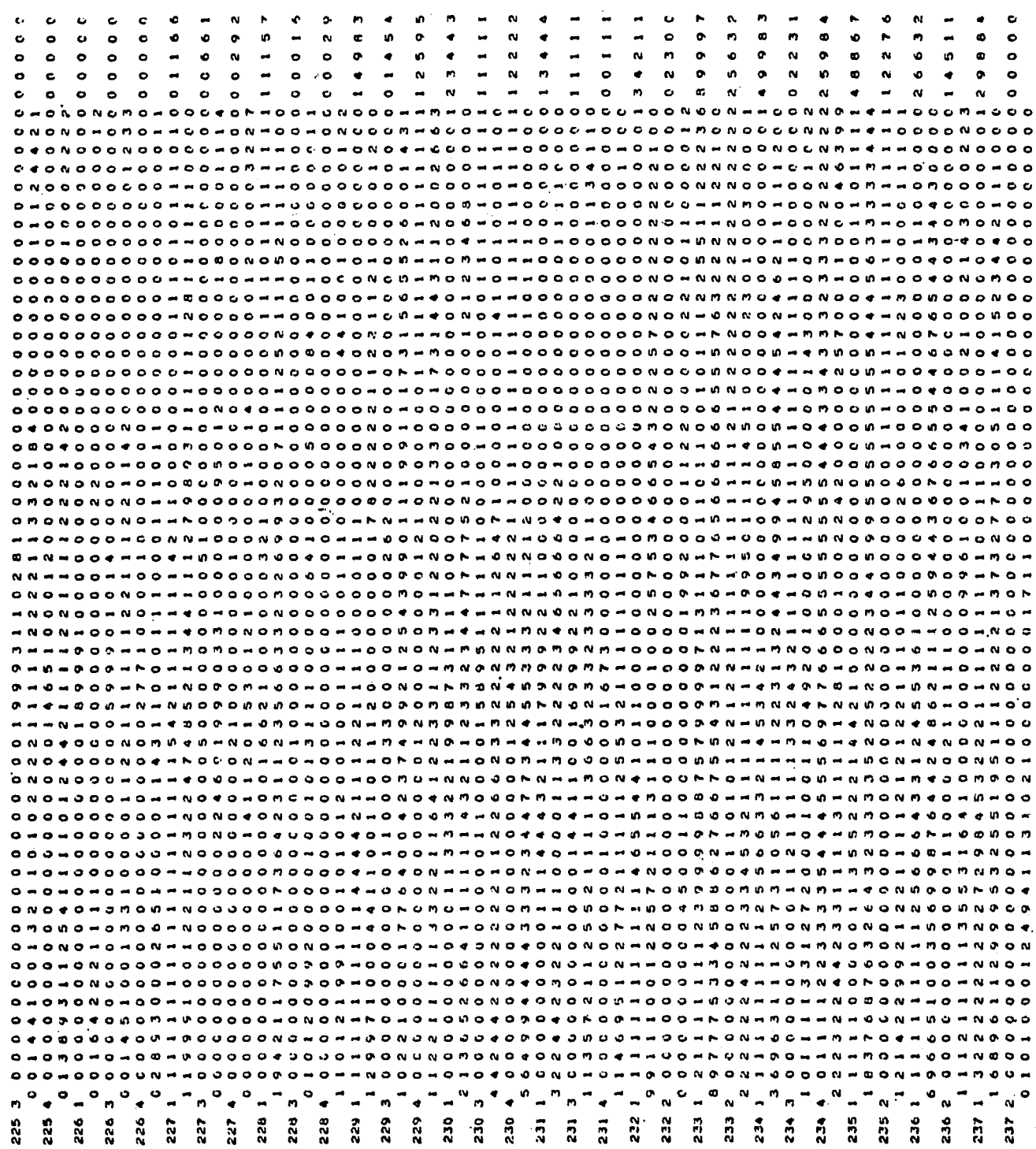


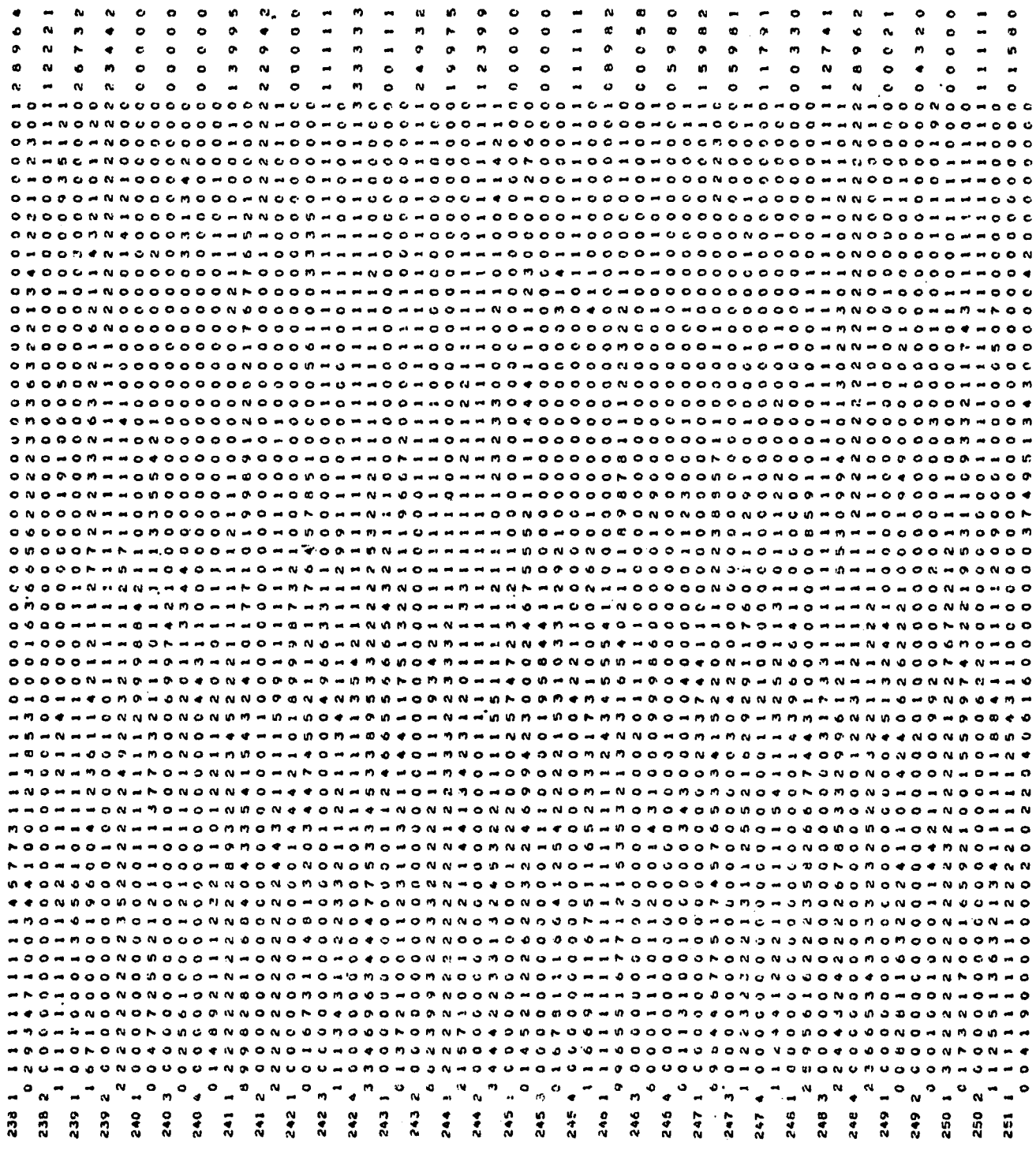


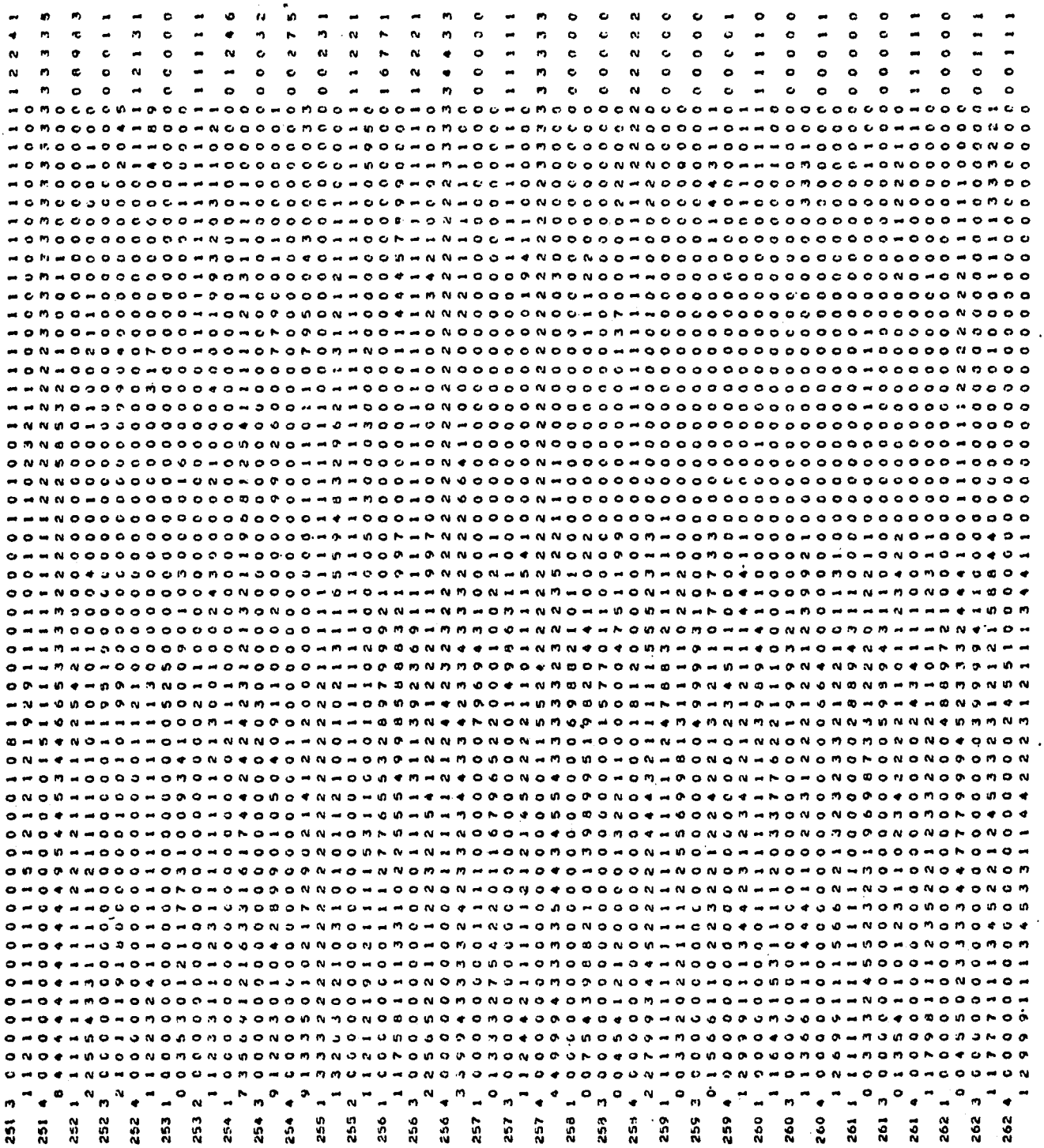


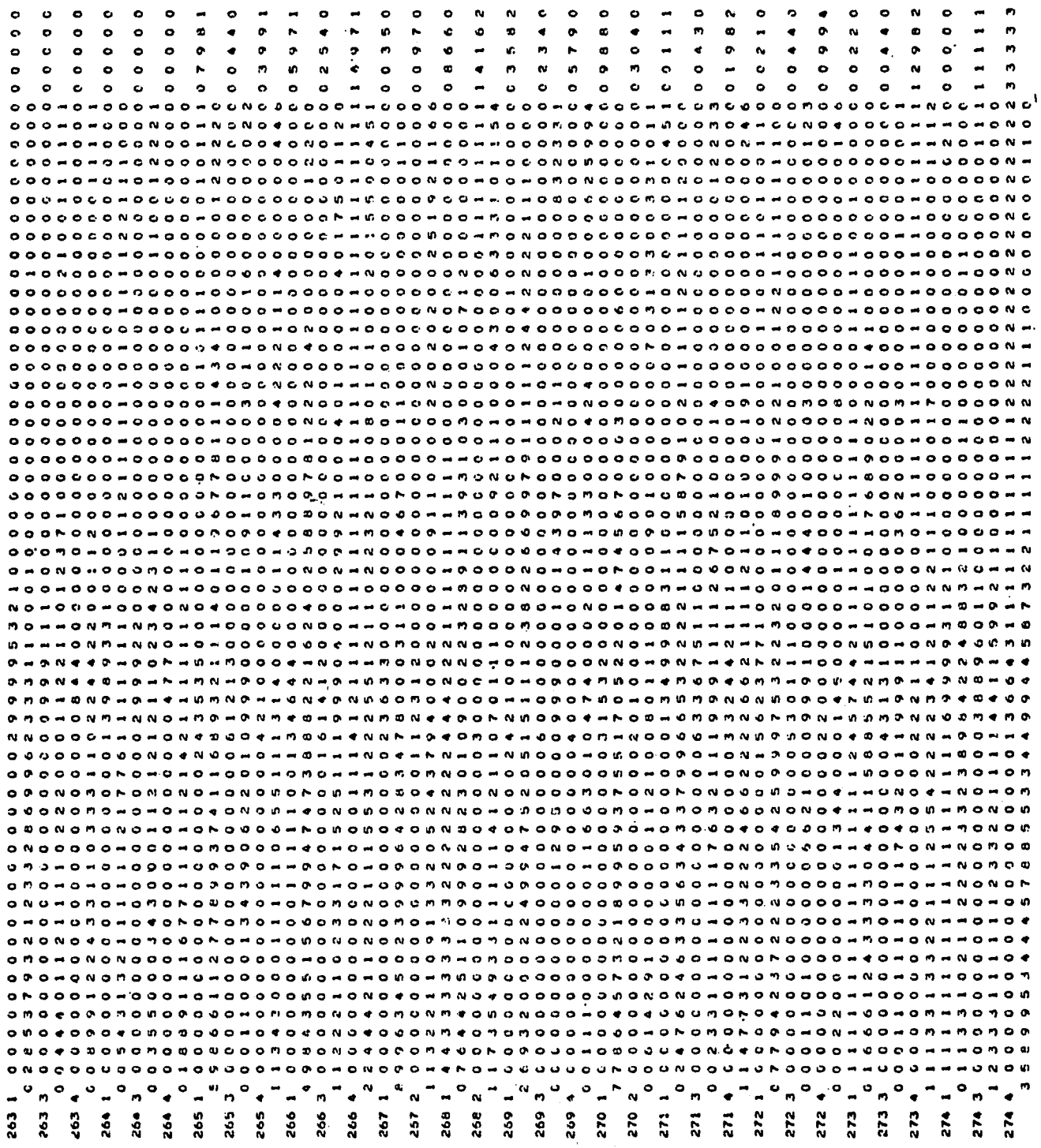


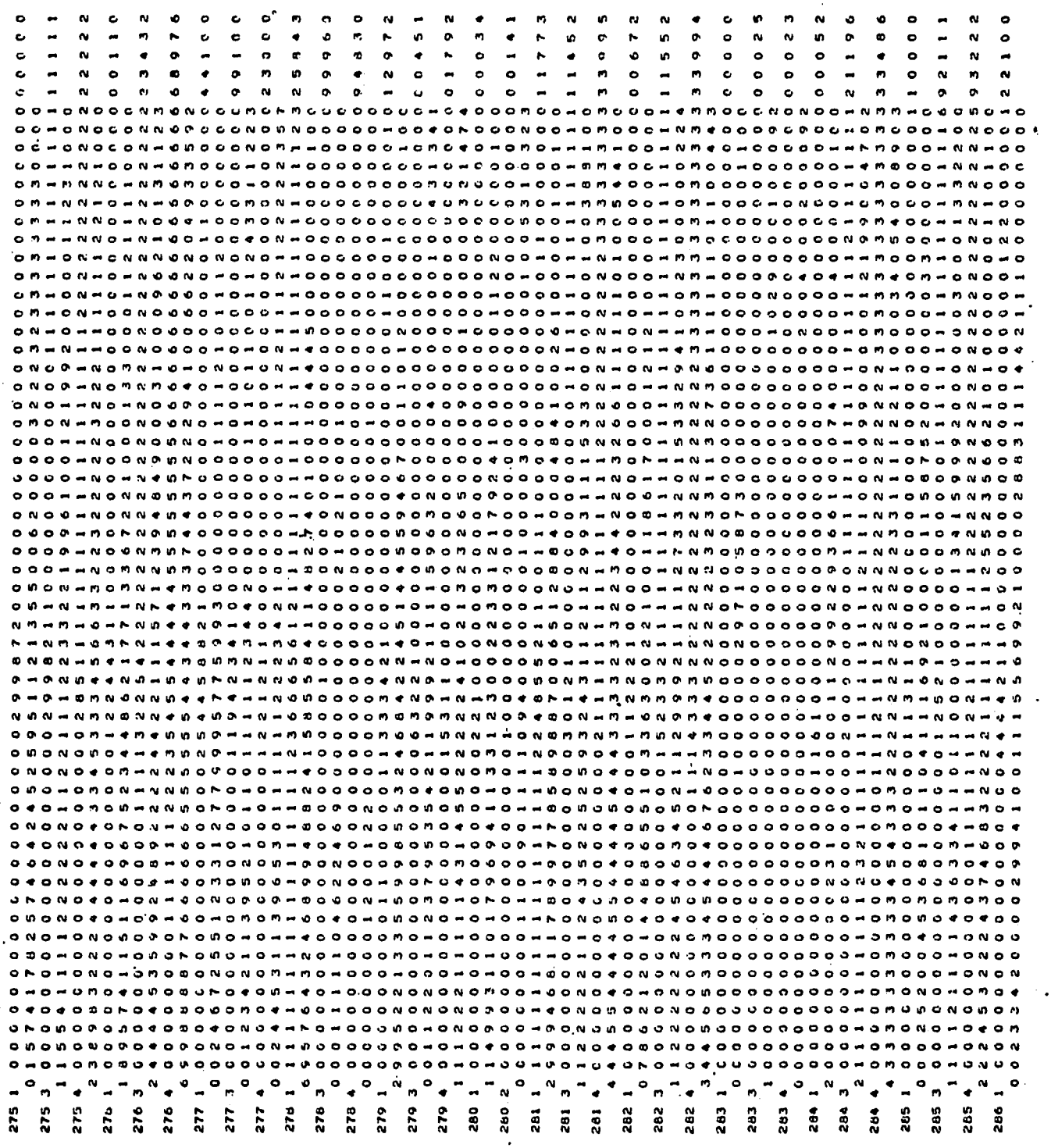


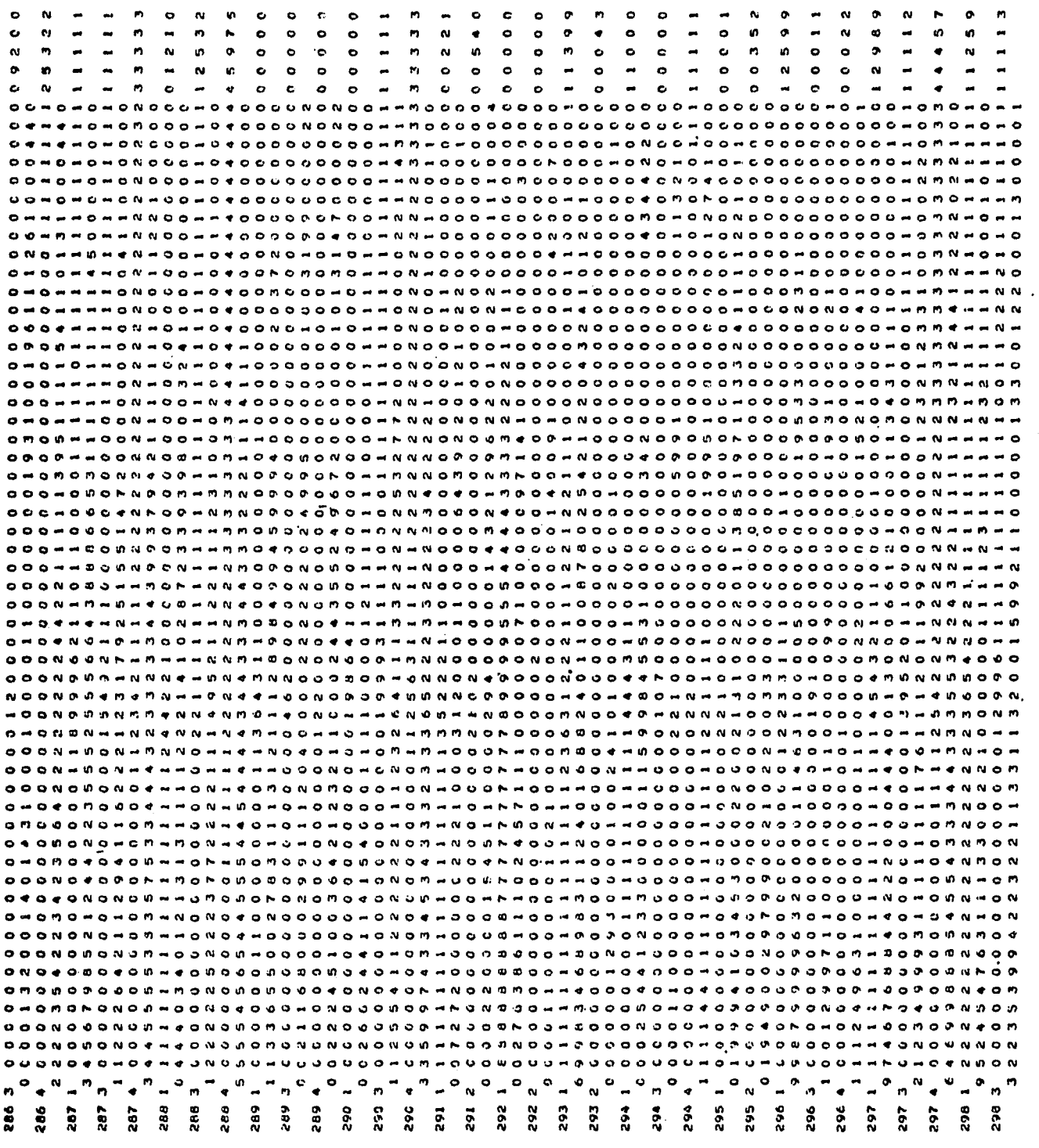


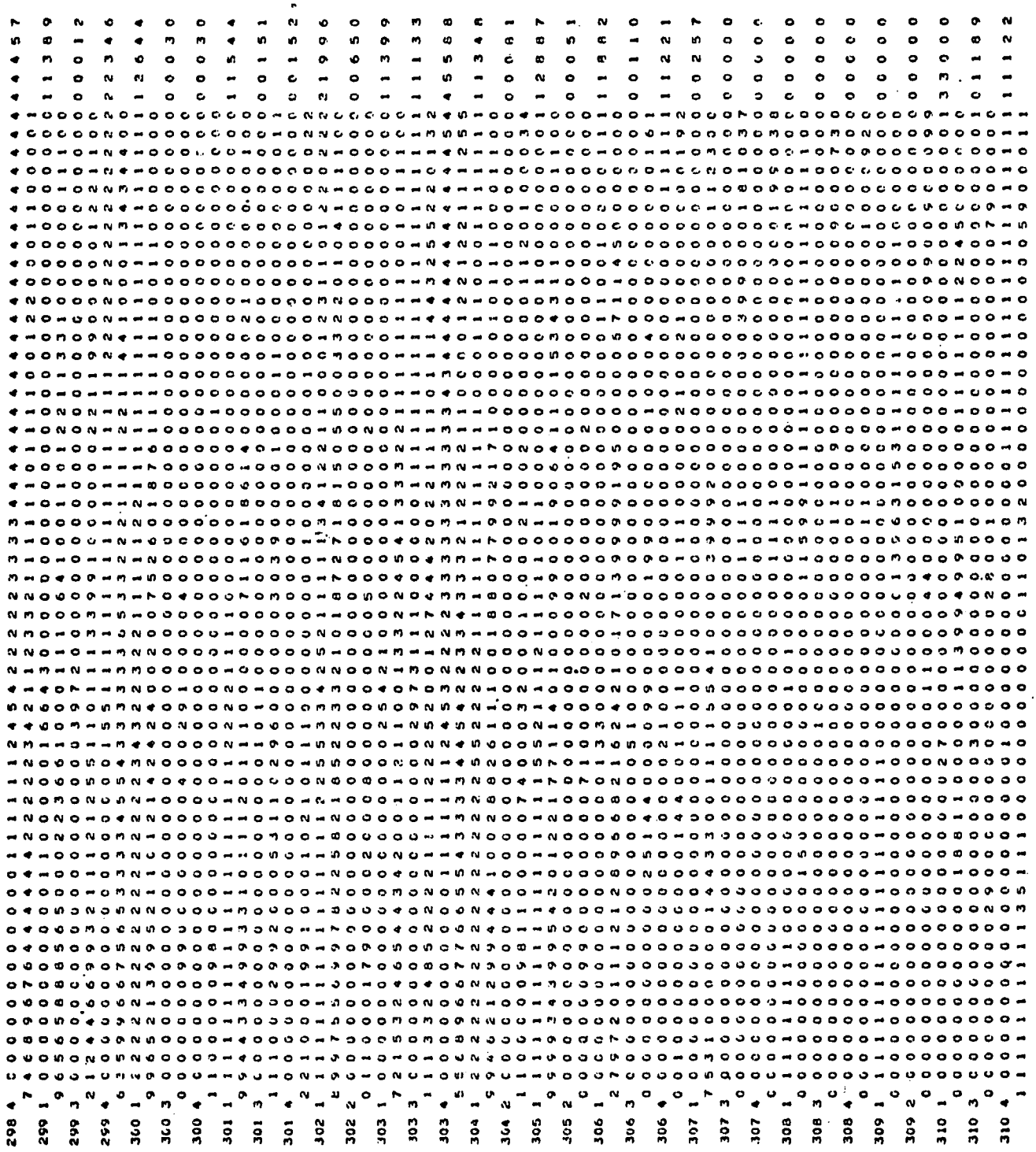


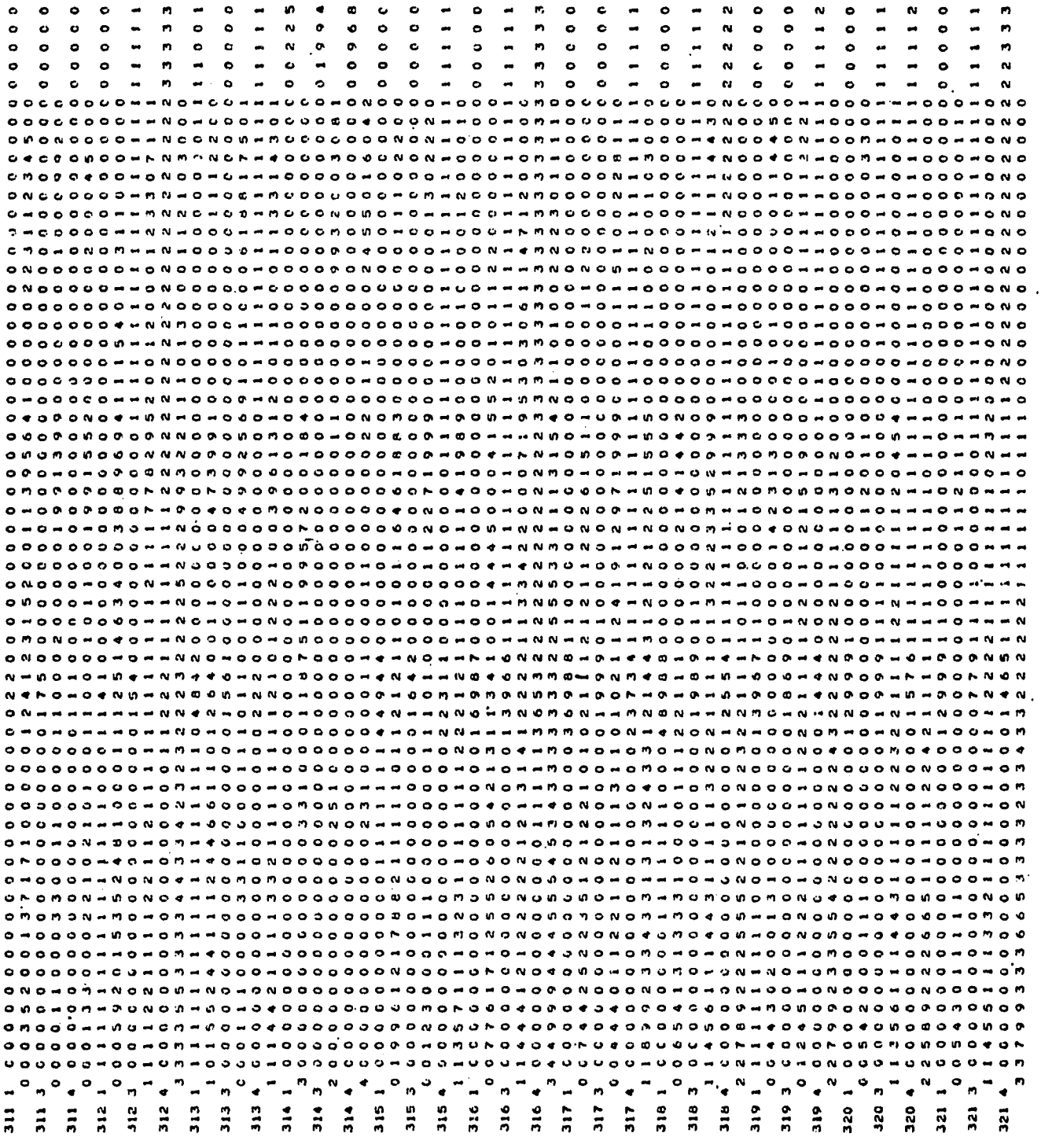


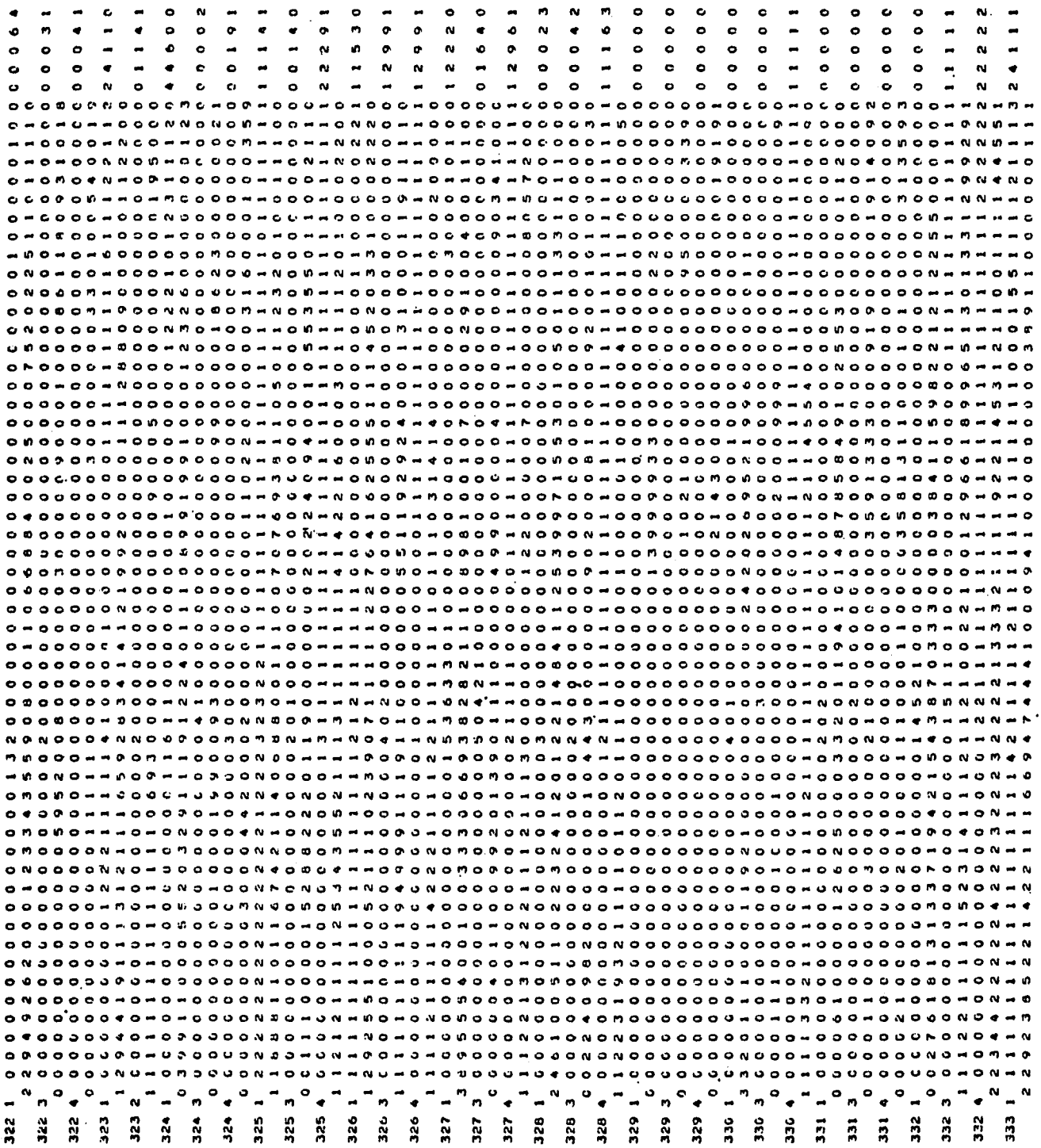


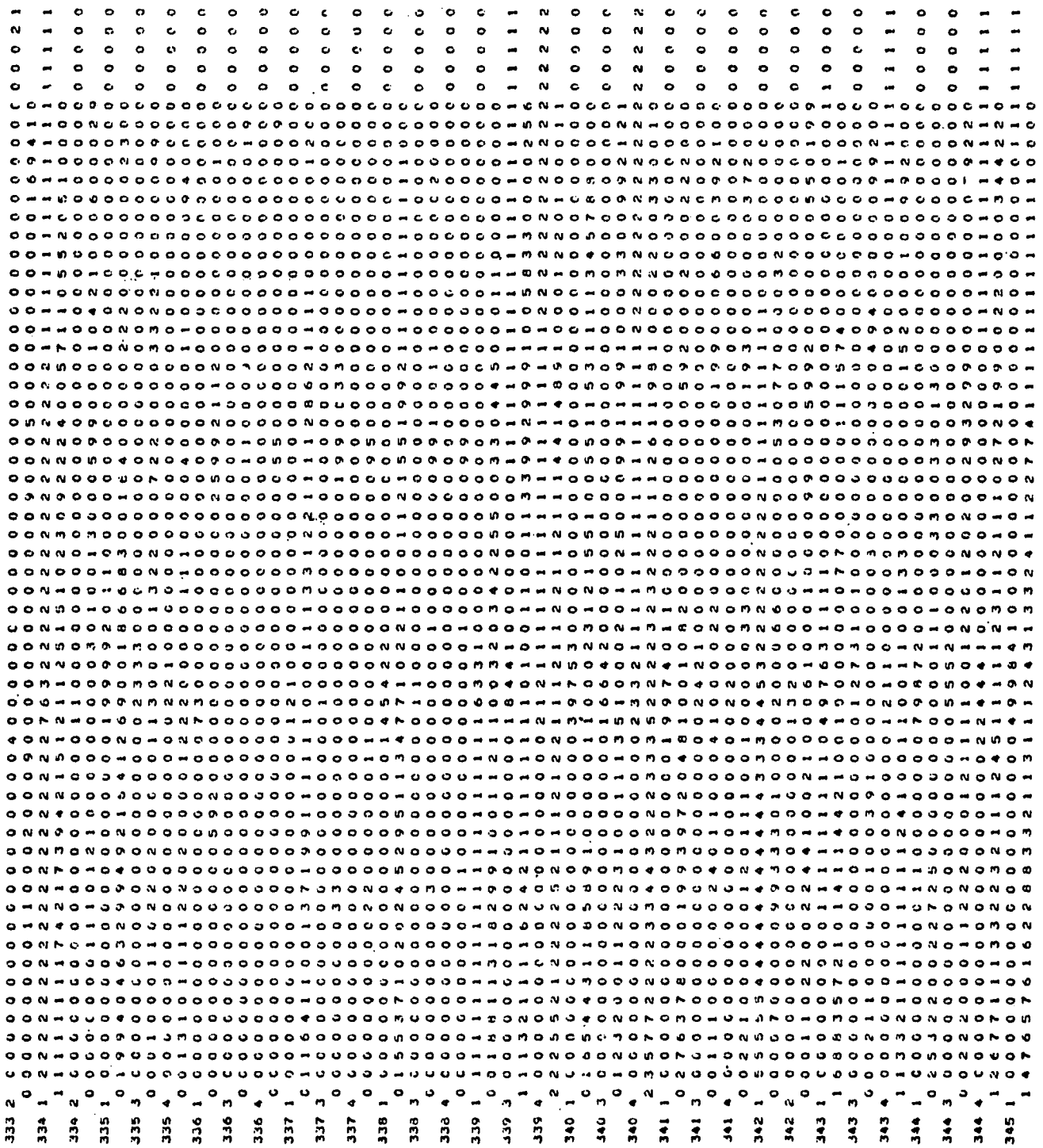


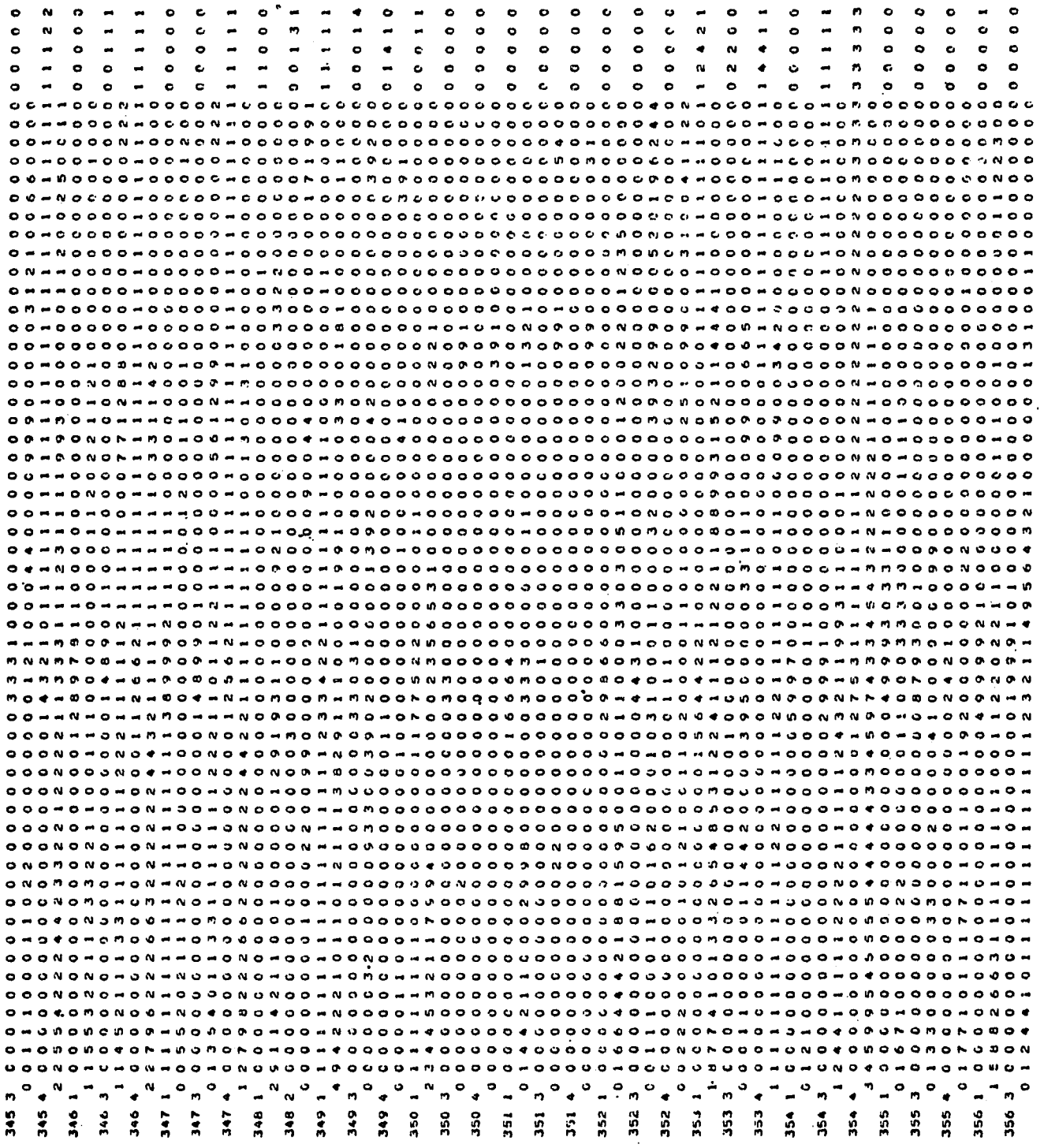


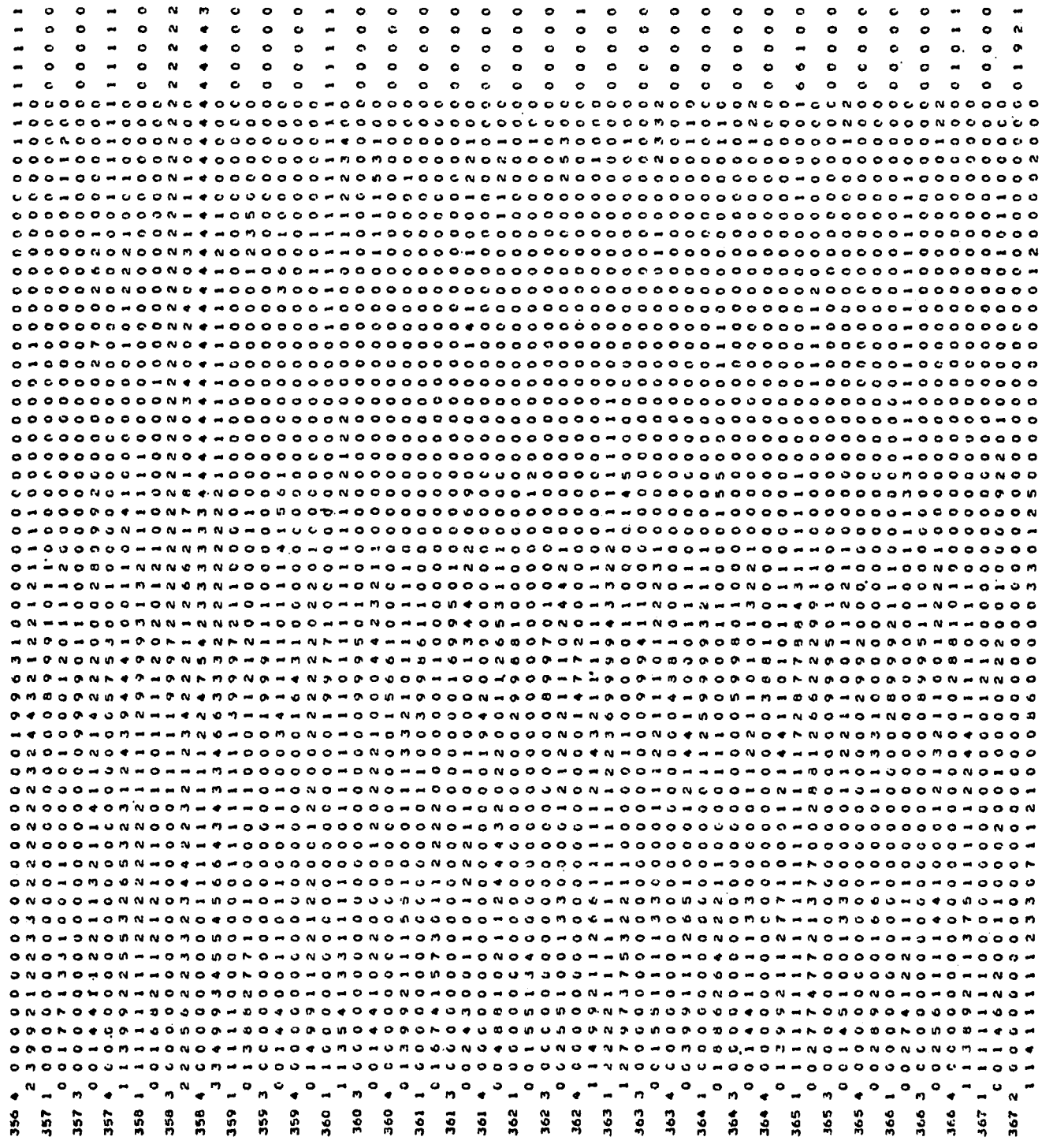


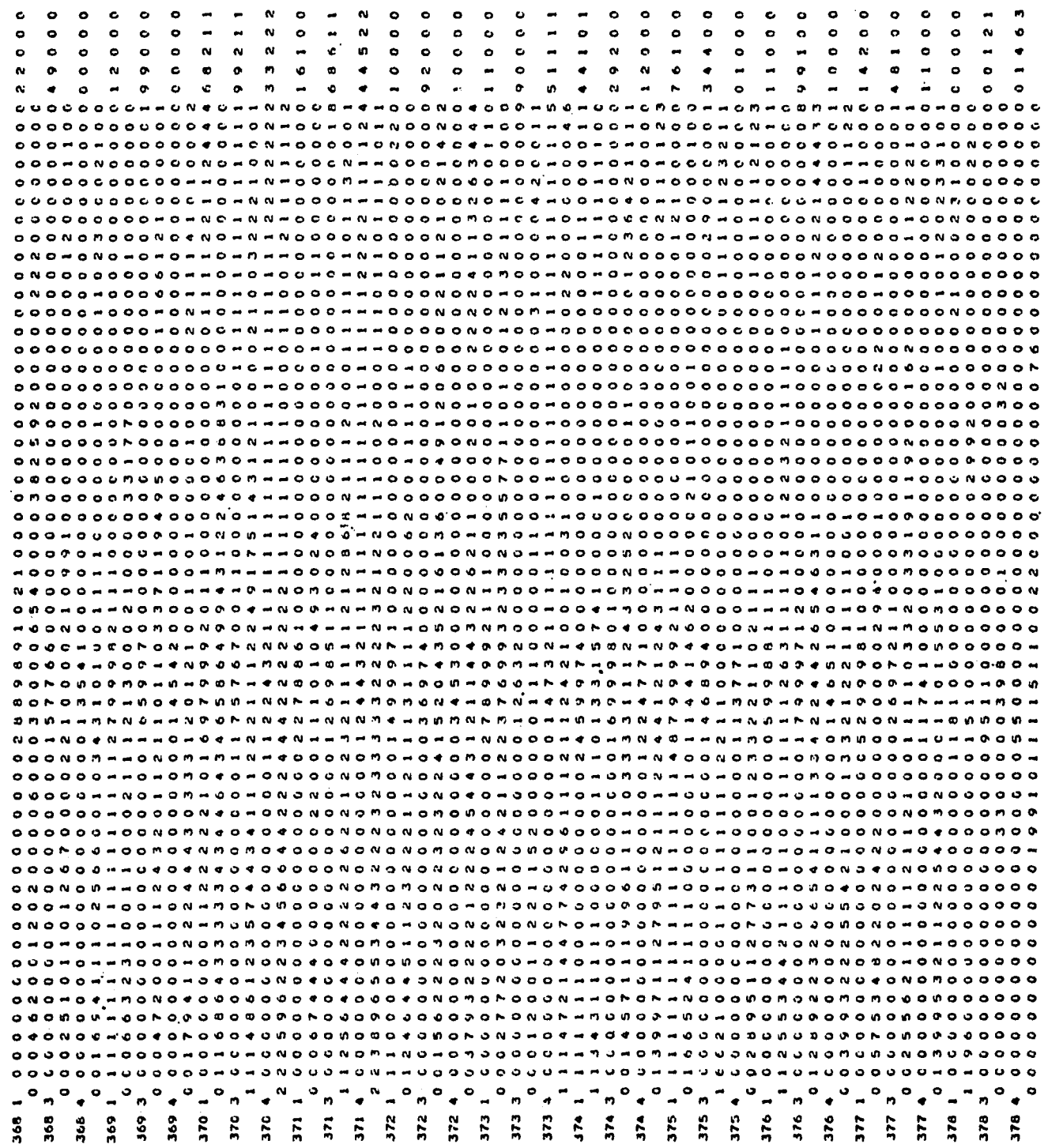


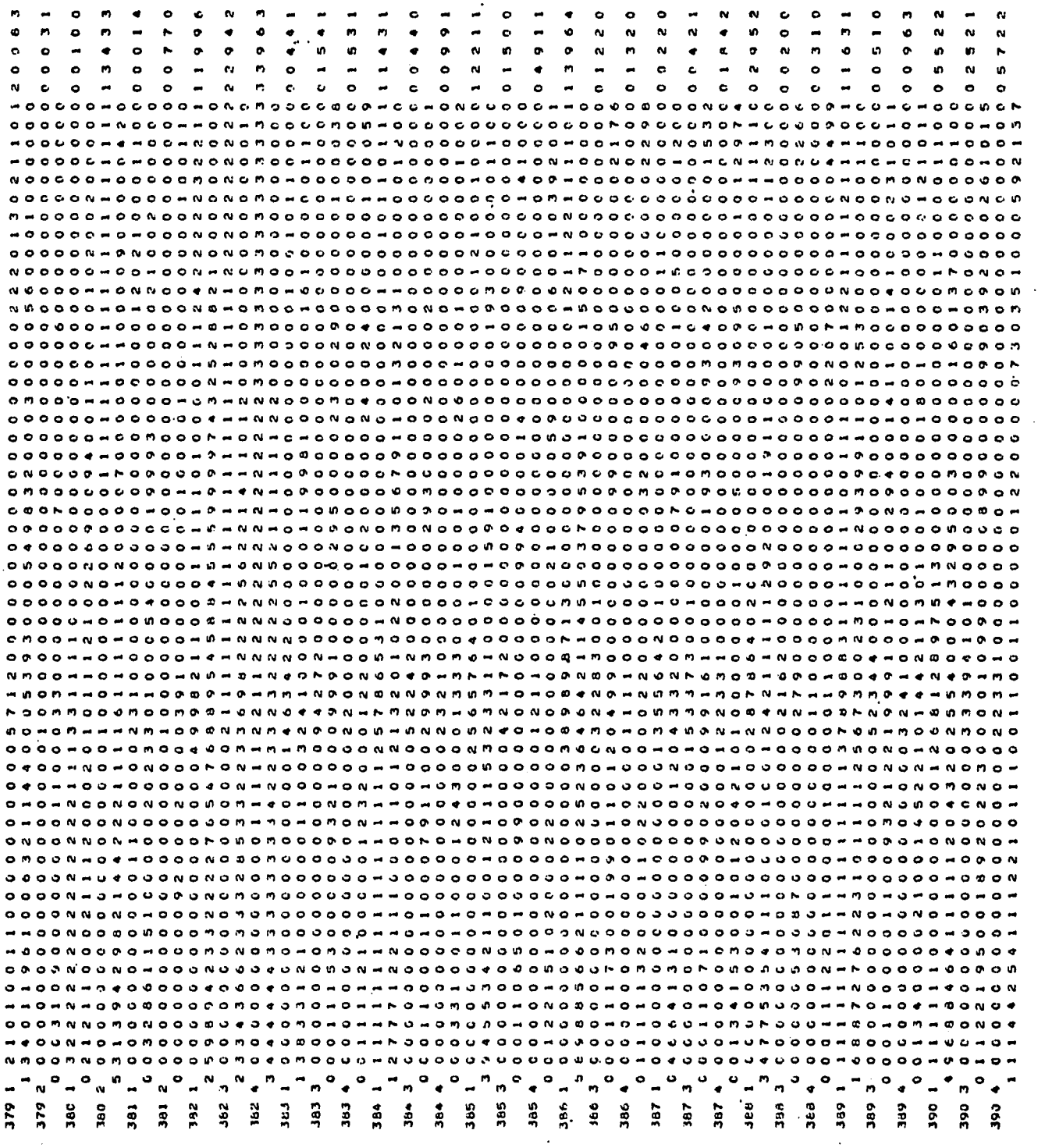


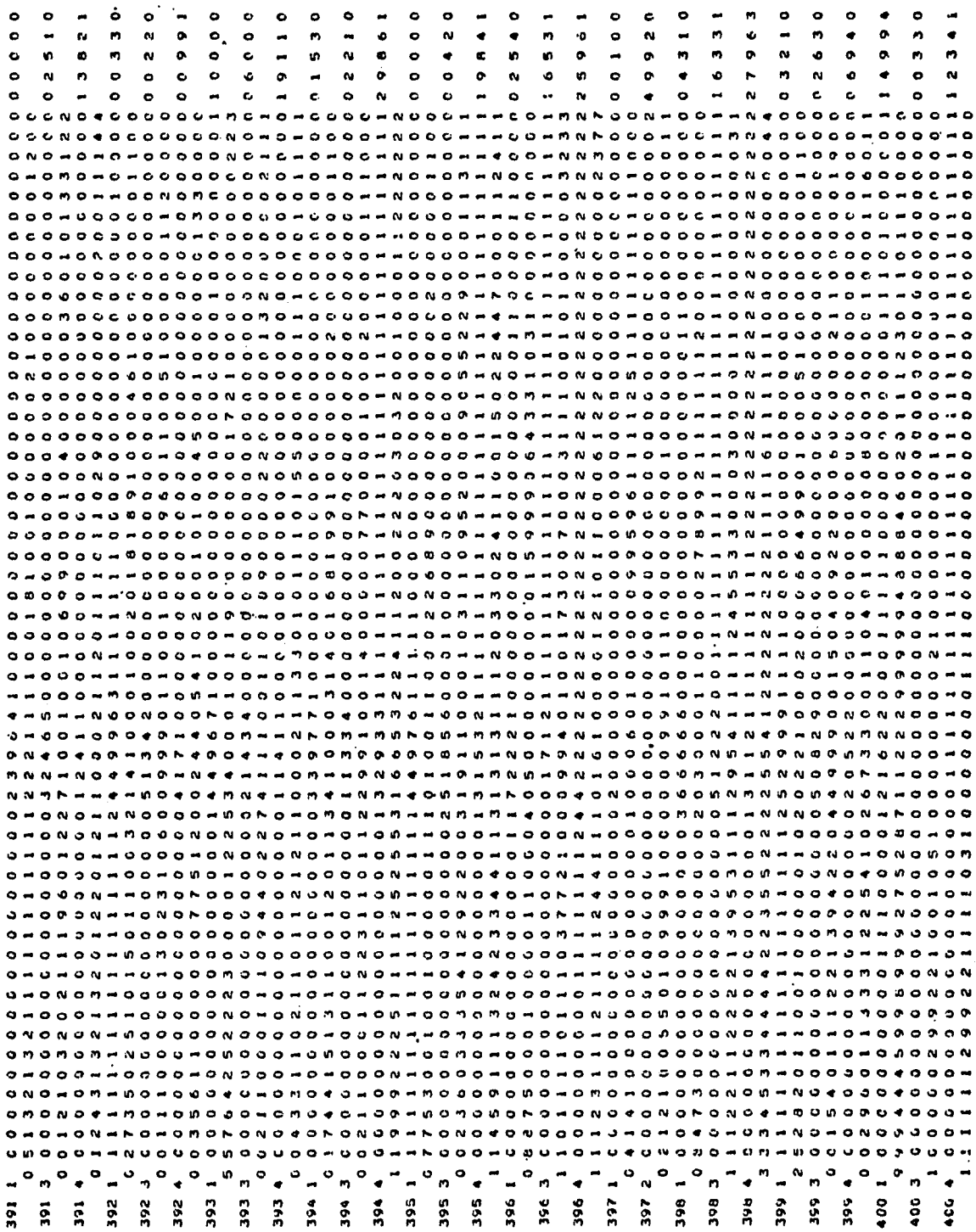












# CHEMICAL CLASS

The following 400 lines of data show the chemical classes to which each of the 400 compounds used belongs. Each line corresponds to a compound, and every 10th line has been numbered to facilitate reading.

The data are written in the format: (4I2). No compound belongs to more than 4 classes. The classes are described in the body of the text. (See Table 1)

In this listing class 15 includes all compounds with a carbonyl group. Compounds 160 through 174 inclusive are the ketones.

```
02 ..... 1
0211
15
15
15
1115
111516
1517
1517
021517 .....10
1315
1315
1315
021115
021115
021117
02111517
1115
1315
1115 .....20
1115
010715
010715
010715
011215
0115
0115
0115
0115
0115 .....30
06
0115
0115
0115
15
0115
0115
0115
0115
0115 ..... 40
0115
0115
0115
0115
010615
0115
0115
0115
0115
0115.....50
0115
0115
010215
0115
0115
0115
0115
0115
0115
0115.....60
```

```
0115
0115
011115
011115
011115
011115
011115
011115
011115
011115 ..........70
011115
011115
011115
011115
011115
011115
011115
011115
011115
011115 ..........80
011115
0115
0115
011115
010315
010415
01020415
010415
01111516
01111516 ..... 90
010615
010615
010615
010615
01061115
01061115
011115
01061115
010615
011315 .....100
011315
011315
1315
011315
011315
011315
011315
011315
011315
010715 ..........110
010715
010915
010915
010915
01091115
01091115
0217
0215
0215
0215 ..........120
0215
0215
021115
021115
02111516
02111517
```

```
021115
0217
021115
111315 ..........130
0717
071117
071117
021117
021117
11
11
02
021115
021117 ..........140
061115
17
00
02
15
15
15
111315
1115
111516 ..........150
021115
061115
061115
02
02
02
0213
0211
0211
15 ..............160
15
15
15
15
15
15
15
15
1115 ............170
15
1115
1115
13
00
03
03
03
03
03 ..............180
03
03
03
03
0311
0311
0311
0304
0304
030411 ..........190
0203
0203
```

```
0203
0203
0203
0203
0203
0306
0306
0306 ••••••••••••••••200
030611
030611
0306
031113
03
0314
0412
04
04
04 ••••••••••••••••••210
0411
0204
05
0511
0514
0514
1116
1116
1116
1116 ••••••••••••••••220
1116
021116
071116
071115
17
17
11
1117
021117
02 ••••••••••••••••••230
02
0211
0211
0211
0211
020611
0211
0211
020711
02 ••••••••••••••••••240
0211
02
0211
020717
0206
02
02
02
02
02 ••••••••••••••••••250
02
02
020617
021117
020911
02
06
06
```

```
06
06 ..................260
06
06
06
06
0611
0611
0611
0611
0611
0611 ................270
0611
0611
0611
06
06
0611
0613
061113
0611
061117 ..............280
060711
060911
0713
13
0713
0913
00
11
07
071-7 ................290
1117
17
021117
17
0211
02
02
02
02
0211 ................300
0211
0211
0211
0207
020711
020711
0811
07
07
07
0713 ................310
07
07
07
0713
07
07
07
07
07
07 ..................320
07
0711
0711
```

```
0711
0711
0711
0711
0711
0709
0709 ••••••••••••330
0709
0709
070911
0912
0912
09
09
09
09
09 ••••••••••••340
09
09
09
09
09
09
09
0911
0911
10 ••••••••••••350
10
10
1011
00
12
00
12
12
00
00 ••••••••••••360
00
00
00
00
00
11
13
13
13 ••••••••••••370
13
13
13
13
13
13
13
11
11
11 ••••••••••••380
11
11
11
11
11
11
11
11
11
```

```
11 ..............390
11
11
11
11
11
11
1113
11
1112
11 ..............400
```

# APPENDIX B

PROGRAM TRAIN

plus

SAMPLE OF INPUT AND OUTPUT

```
LEVEL 21.7 ( JAN 73 )                    OS/360  FORTRAN H

          COMPILER OPTIONS - NAME=  MAIN,OPT=02,LINECNT=58,SIZE=0000K,
                             SOURCE,EBCDIC,NOLIST,NODECK,LOAD,MAP,NOEDIT,ID,NOXREF
 ISN 0002            COMMON FX(400,114),JC(112),T(400),W(113),NERR(400),TERR(400)
 ISN 0003            DIMENSION FY(400),C1(400),S1(400),NECOM(400)
                   C
                   C
                   C
                   C
                   C
                   C
                   C
                   C
                   C
                   C
                   C  ZEROING OF THE MATRICIES IS DONE HERE
                   C
                   C
 ISN 0004        1      DO 2 I=1,400
 ISN 0005               T(I) = 0
 ISN 0006               NECOM(I) = 0
 ISN 0007               FX(I,113) = 1
 ISN 0008               FX(I,114) = 0
 ISN 0009               DO 2 J=1,112
 ISN 0010               W(J) = 0
 ISN 0011        2      FX(I,J) = 0
 ISN 0012               W(113) = 0
                   C
                   C  L COUNTS THE # OF TIMES THE DATA SET IS TREATED (PASS).
                   C  NDC IS THE NUMBER OF THE DATA CARD (SPECTRUM) BEING WORKED ON
                   C  NCARD IS THE TOTAL NUMBER OF CARDS(SPECTRA).
                   C  NSHUT LIMITS THE NUMBER OF PASSES ON THE DATA.
                   C  N EQUALS THE NUMBER OF TERMS IN THE VECTOR BEING TRAINED.
                   C  I.E. IT IS THE DIMENSIONALITY OF THE SPACE WE ARE WORKING IN.
                   C
                   C
 ISN 0013               N=113
 ISN 0014               NT=N-1
 ISN 0015               L=1
 ISN 0016               LJON=1
 ISN 0017               ND5 = 5
 ISN 0018               ND6 = 6
                   C
                   C
                   C
                   C
                   C
                   C
                   C  SUBROUTINE REOP READS THE OPTIONS DESIRED AND
                   C  SUBROUTINE RE71 READS THE MAIN DATA DECK ACCORDING
                   C  TO THE SPECIFIED OPTIONS.  THE MAIN DATA DECK IS STORED
                   C  ON THE DISK IN BINARY FORM.  THE OPTIONS AVAILABLE
                   C  ARE EXPLAINED IN THE SUBROUTINE REOP.
                   C
                   C
 ISN 0019               CALL REOP(NOPT1,NCARD,NOPT2,FIN,NSHUT,LEN,NOPT3,NOPT4,NOPT5,
```

```
     1NOPT6)
ISN 0020      NOP=NOPT6
ISN 0021      CALL RE71(NOPT1,NCARD,NCONT,NOPT2,NOPT3,NOPT4,N30)
C
C
C
C
C
C
C
C
C
C  THE SUM OF SQUARES FOR EACH COMPOUND IN THE TRAINING SET IS CALCULATED
C
C
ISN 0022      DO 10 I=1,NCONT
ISN 0023      FY(I) = 0.0
ISN 0024      DO 10 J=1,N
ISN 0025   10 FY(I) = FY(I) + (FX(I,J)**2)
C
C
C
C
C
C
C
C
C  THE FOLLOWING SECTION INITIALIZES THE VECTOR(W) WHICH IS TO
C  BE TRAINED.  EITHER THE INITIAL VECTOR IS READ FROM A DISK
C  FILE (FIN=0.0) OR IT IS  INITIALIZED AS FIN.
ISN 0026      READ(ND5,13) FKOR
ISN 0027   13 FORMAT(F5.2)
ISN 0028      IF(FIN)21,20,21
ISN 0029   20 READ(ND5,11)NFILE
ISN 0030   11 FORMAT(I4)
ISN 0031      DO 12 J=1,NFILE
ISN 0032   12 READ(73)(W(I),I=1,N)
C
C
C
C
C
C  FKOR IS THE CONSTANT USED TO TO MULTIPLY THE CORRECTION FACTOR.
C
C
ISN 0033      GO TO 30
ISN 0034   21 DO 22 I=1,N
ISN 0035   22 W(I) =FIN
C
C
C
C
C
C
C
C
```

```
                C
                C  THIS SECTION CONTAINS THE MAIN TRAINING ROUTINE.
                C
                C
                C
ISN 0036        30    K=0
ISN 0037              IF(L-NSHUT) 31,31,200
ISN 0038        31    NDC = 0
                C
                C  K IS THE ERROR CHECK TO SEE IF THE VECTOR W IS TRAINED. IF
                C  K IS 0 NO ERROR YET FOUND IN THE L'TH READ OF  THE DATA
                C
ISN 0039        33    IF(NCONT-NDC) 40,40,32
ISN 0040        32    NDC = NDC + 1
ISN 0041              IF(T(NDC) - 9.0 ) 34,33,33
ISN 0042        34    S=0.0
ISN 0043              DO 35 I=1,NT,NOPT6
ISN 0044        35    S=S+(W(I) * FX(NDC,I))
ISN 0045              S=S+ (W(N) *FX(NDC,N))
ISN 0046              IF (T(NDC)) 36,36,37
ISN 0047        36    IF(S) 33,33,38
ISN 0048        37    IF(S) 38,38,33
ISN 0049        38    K = K+1
ISN 0050              NERR(K) = FX(NDC,114)
ISN 0051              NECOM(NDC) = NECOM(NDC) + 1
ISN 0052              TERR(K) = T(NDC)
ISN 0053              C = (-1.0*S)/FY(NDC)
ISN 0054              C=C*FKOP
ISN 0055              S1(K) = S
ISN 0056              C1(K) = C
ISN 0057              DO 39 I=1,NT,NOPT6
ISN 0058        39    W(I) = W(I) + (C*FX(NDC,I))
ISN 0059              W(N) = W(N) + (C * FX(NDC,N))
ISN 0060              GO TO 33
ISN 0061        40    IF(K) 100,100,41
ISN 0062        41    FK = K
ISN 0063              IF(LJON-LEN)50,51,51
                C
                C  THE VALUE OF LEN DETERMINES HOW OFTEN THE UNTRAINED
                C  VECTOR WILL BE PRINTED.  E.G. IF LEN = 10, THE UNTRAINED VECTOR
                C  WILL BE PRINTED EVERY 10TH ITERATION.
                C
ISN 0064        50    L=L+1
ISN 0065              LJON=LJON+1
ISN 0066              GO TO 30
ISN 0067        51    LJON=1
ISN 0068              FNDC = N30
ISN 0069              FPRE = (FK/FNDC)*100.0
ISN 0070              WRITE(NO6,802)L,K,N30,FPRE
ISN 0071        802   FORMAT(1H1,'AFTER ',I5,'   PASSES THERE ARE ',I5,
                     1' ERRORS IN ',I5,'   SPECTRA',/,F6.2,' PERCENT ERRORS')
ISN 0072              WRITE(NO6,4499)
ISN 0073        4499  FORMAT(1H ,'COMPOUND  SIGN     VALUE OF S     VALUE OF C')
ISN 0074              DO 52 I=1,K
ISN 0075        52    WRITE(NO6,907)NERR(I),TERR(I),S1(I),C1(I)
ISN 0076        907   FORMAT(1H ,4X,I3,3X,F5.1,3X,F12.7,3X,F12.7)
```

```
ISN 0077            WRITE(ND6,805)
ISN 0078      805   FORMAT(1H ,'THE FOLLOWING IS THE CURRENT DESCRIPTOR:')
ISN 0079            WRITE(ND6,997)(W(I),I=1,N)
ISN 0080            CALL UNPACK
ISN 0081            L=L+1
ISN 0082            GO TO 30
                    C
                    C
                    C
                    C
                    C
                    C
                    C
                    C
                    C
                    C  THE FOLLOWING IS THE WRITE ROUTINE.
                    C
                    C
ISN 0083      100   WRITE(ND6,995) L
ISN 0084      995   FORMAT(1H1,'THE DATA HAVE BEEN READ',I6,'   TIMES')
ISN 0085            WRITE(ND6,993)
ISN 0086      993   FORMAT(1H ,'THE FOLLOWING IS THE TRAINED VECTOR')
ISN 0087            WRITE(ND6,997) (W(I),I = 1,N)
ISN 0088      997   FORMAT(1H ,10F9.5,/,11(1X,10F9.5,/))
ISN 0089            GO TO 300
                    C
                    C
                    C
                    C
                    C
                    C
                    C
                    C
                    C
                    C  THE FOLLOWING SECTION IS ENTERED ONLY IF THE MAXIMUM
                    C  NUMBER OF ALLOWED ITERATIONS (NSHUT) IS REACHED.
                    C
                    C
ISN 0090      200   WRITE(ND6,899)NSHUT
ISN 0091            L=L-1
ISN 0092            K=FK
ISN 0093            FN30 = NDC
ISN 0094            FPRE = (FK/FN30)*100.0
ISN 0095            WRITE(ND6,802)L,K,N30,FPRE
ISN 0096            WRITE(ND6,4499)
ISN 0097            DO 201 I=1,K
ISN 0098      201   WRITE(ND6,907)NERR(I),TERR(I),SI(I),CI(I)
ISN 0099            WRITE(ND6,1993)
ISN 0100      1993  FORMAT(1H ,'THE FOLLOWING IS THE LAST VECTOR')
ISN 0101            WRITE(ND6,997) (W(I),I = 1,N)
ISN 0102      899   FORMAT(1H ,'PROGRAM HALTED.',I7,' PASSES MADE.')
                    C
                    C
                    C
                    C
                    C
```

```
                C
                C
                C
                C
                C
                C  STATISTICS  CONCERNING THE NUMBER OF COMPOUNDS USED AND THE
                C  PERCENT IN THE CLASS ARE CALCULATED AND PRINTED.
                C
                C
ISN 0103       300   CONTINUE
ISN 0104             FMIN = 0
ISN 0105             FPLUS = 0
ISN 0106             DO 304 I=1,NCONT
ISN 0107             IF(T(I) - 9.0) 301,304,304
ISN 0108       301   IF(T(I)) 302,303,303
ISN 0109       302   FMIN = FMIN+ 1.0
ISN 0110             GO TO 304
ISN 0111       303   FPLUS = FPLUS + 1.0
ISN 0112       304   CONTINUE
ISN 0113             P1=(FMIN*100.0)/(FMIN + FPLUS)
ISN 0114        .    P2=(FPLUS*100)/(FMIN+FPLUS)
ISN 0115             WRITE(ND6,9998)
ISN 0116             WRITE(ND6,9998)
ISN 0117             WRITE(ND6,9998)
ISN 0118             WRITE(ND6,9998)
ISN 0119       9998  FORMAT(1H )
ISN 0120             WRITE(ND6,9999)FMIN,P1
ISN 0121       9999  FORMAT(1H ,'THE DECK WAS COMPOSED OF ',F6.1,'(',F6.2,
                    1' PERCENT ) COMPOUNDS NOT IN THE CLASS. AND')
ISN 0122             WRITE(ND6,9997)FPLUS,P2
ISN 0123       9997  FORMAT(1H ,F6.1,'(',F6.2,' PERCENT ) IN THE CLASS.')
                C
                C
                C
                C
                C
                C
                C
                C
                C
                C  THE VECTOR IS PLOTED.
                C
                C
ISN 0124             WRITE (72) ( W(I), I=1,N)
ISN 0125             WRITE(ND6,350)
ISN 0126       350   FORMAT(1H1,8('COM CLASS ERR   '))
ISN 0127             DO 351 I=1,400,8
ISN 0128             I1 = I+1
ISN 0129             I2 = I+2
ISN 0130             I3 = I+3
ISN 0131             I4 = I+4
ISN 0132             I5 = I+5
ISN 0133             I6 = I+6
ISN 0134             I7=I+7
ISN 0135       351   WRITE(ND6,352)FX(I,114),T(I),NECOM(I),FX(I1,114),T(I1),NECOM(I1),
                    1FX(I2,114),T(I2),NECOM(I2),FX(I3,114),T(I3),NECOM(I3),
```

```
                     1FX(I4,114),T(I4),NECOM(I4),FX(I5,114),T(I5),NFCOM(I5),
                     1FX(I6,114),T(I6),NECOM(I6),FX(I7,114),T(I7),NFCOM(I7)
ISN 0136      352    FORMAT(1H ,8(F5.1,F4.1,1X,I3,3X))
ISN 0137             CALL UNPACK
ISN 0138             IF(NOPT6 .EQ. 1) GO TO 500
ISN 0140             NT2 = NOPT6 - 1
ISN 0141             DO 400 I=1,NT,NOPT6
ISN 0142             DO 400 J=1,NT2
ISN 0143             II = I + J
ISN 0144             IF(II .GE. NT) GO TO 400
ISN 0146             W(II) = W(I)
ISN 0147      400    CONTINUE
ISN 0148             NOPT6 = 1
ISN 0149             L=1
ISN 0150             GO TO 30
ISN 0151      500    CONTINUE
                     C
                     C
                     C
                     C
                     C
                     C
                     C
                     C
                     C
                     C  THE OPTIONS WHICH WERE USED DURING THE RUN ARE LISTED
                     C
                     C
ISN 0152             WRITE(ND6,9999)
ISN 0153             WRITE(ND6,9998)
ISN 0154             WRITE(ND6,9998)
ISN 0155             WRITE(ND6,9998)
ISN 0156             WRITE(ND6,4500)
ISN 0157      4500   FORMAT(1H ,'THE OPTIONS FOLLOWED WERE THE FOLLOWING:')
ISN 0158             WRITE(ND6,4501)NOPT1
ISN 0159      4501   FORMAT(1H ,'OPTION 1 = ',I4)
ISN 0160             WRITE(ND6,4502)NOPT2
ISN 0161      4502   FORMAT(1H ,'OPTION 2 = ',I4)
ISN 0162             WRITE(ND6,4503)NOPT3
ISN 0163      4503   FORMAT(1H ,'OPTION 3 = ',I4)
ISN 0164             WRITE(ND6,4504)NOPT4
ISN 0165      4504   FORMAT(1H ,'OPTION 4 = ',I4)
ISN 0166             WRITE(ND6,4509) NOP
ISN 0167      4509   FORMAT(1H ,'OPTION 6 = ',I4)
ISN 0168             WRITE(ND6,4505)NCARD
ISN 0169      4505   FORMAT(1H ,'TOTAL NUMBER OF SPECTRA CONSIDERED (NCARD) = ',I4)
ISN 0170             WRITE(ND6,4506)NSHUT
ISN 0171      4506   FORMAT(1H ,'MAXIMUM NUMBER OF ITERATIONS (NSHUT) = ',I4)
ISN 0172             WRITE(ND6,4507)LEN
ISN 0173      4507   FORMAT(1H ,'THE UNTRAINED VECTOR WAS PRINTED EVERY ',I4,
                    1'TH ITERATION')
ISN 0174             WRITE(ND6,4508)FIN
ISN 0175      4508   FORMAT(1H ,'THE INITIAL VALUE OF THE VECTOR WAS (FIN) ',F9.4)
ISN 0176             IF(NOPT5)600,600,601
ISN 0177      601    REWIND 71
ISN 0178             GO TO 1
```

```
ISN 0179      600   CONTINUE
ISN 0180            ENDFILE 72
ISN 0181            STOP
ISN 0182            END
```

```
LEVEL 21.7 ( JAN 73 )                    OS/360  FORTRAN H

          COMPILER OPTIONS - NAME=  MAIN,OPT=02,LINECNT=58,SIZE=0000K,
                             SOURCE,EBCDIC,NOLIST,NODECK,LOAD,MAP,NOEDIT,ID,NOXREF
                      C
                      C
                      C
                      C
                      C
                      C
                      C
                      C
                      C
                      C
                      C
 ISN 0002                   SUBROUTINE PE71(NOPT1,NCARD,NCONT,NOPT2,NOPT3,NOPT4,N30)
 ISN 0003                   COMMON FX(400,114),JC(112),T(400),W(113),NERR(400),TERR(400)
                      C
                      C
                      C
                      C
                      C  THIS READ SECTION READS IR AND RAMAN DATA IN THE RANGE FROM
                      C  1900 TO 500 CM-1.  IT CONSIDERS BOTH THE IR AND THE RAMAN AS A
                      C .SINGLE VECTOR.  WHERE THE PARALLEL AND PERPENDICULAR RAMAN DATA
                      C  ARE AVAILABLE, THE PARALLEL SPECTRUM ONLY IS USED.
 ISN 0004                   NCONT = 0
 ISN 0005                   DO 200 KON = 1,NCARD
 ISN 0006                   READ(71) N1,N2,(JC(I),I=1,58)
 ISN 0007                   READ(71)(JC(I),I=59,112)
 ISN 0008                   IF(N2 .GT. 1) GO TO 100
 ISN 0010                   NCONT = NCONT + 1
 ISN 0011                   DO 1 I=1,56
 ISN 0012                   J=I+44
 ISN 0013              1    FX(NCONT,I) = JC(J)
 ISN 0014                   FX(NCONT,114) = N1
 ISN 0015                   GO TO 200
 ISN 0016              100  IF(N2 .EQ. 4) GO TO 200
 ISN 0018                   J=45
 ISN 0019                   DO 101 I=57,112
 ISN 0020                   FX(NCONT,I) = JC(J)
 ISN 0021              101  J=J+1
 ISN 0022              200  CONTINUE
                      C
                      C
                      C
                      C
                      C
                      C
                      C
                      C
                      C
                      C
                      C  THE CLASS IS NOW READ AS PRESCRIBED BY OPTION 2 (NOPT2).
                      C
                      C
                      C  NOPT2 = 1 THRU 17 : CLASS 1 THRU 17
                      C  NOPT2 = 18 : CLASSES 7,8,9AND 10(ALL HALOGENS)
                      C  NOPT2 = 19 : CLASSES 8,9,AND 10 (F,BR AND I)
                      C  NOPT2 = 20 : CLASSES 3,4 AND 5(ALCOHOLS)
```

```
                C  NOPT2 = 21 : CLASSES 3,4,5 AND 16 (ALCOHOLS AND PHENOLS)
ISN 0023     1000  IF(NOPT2 .EQ. 15) GO TO 1081
ISN 0025           DO 1060 J=1,400
ISN 0026           READ(71) (JC(I),I=1,17)
ISN 0027           IF(NOPT2 - 18) 1001,1002,1003
ISN 0028     1001  N = JC(NOPT2)
ISN 0029           GO TO 1050
ISN 0030     1002  N=JC(7) + JC(8) + JC(9) + JC(10)
ISN 0031           GO TO 1050
ISN 0032     1003  IF(NOPT2-20) 1004,1005,1006
ISN 0033     1004  N=JC(8) + JC(9) + JC(10)
ISN 0034           GO TO 1050
ISN 0035     1005  N=JC(3) + JC(4) + JC(5)
ISN 0036           GO TO 1050
ISN 0037     1006  N=JC(3) + JC(4) + JC(5) + JC(16)
ISN 0038     1050  IF(N) 1060,1060,1061
ISN 0039     1061  T(J) = 1
ISN 0040     1060  CONTINUE
ISN 0041           GO TO 1082
ISN 0042     1081  CALL CLASS15
                   C
                   C
                   C
                   C
                   C
                   C  THE FOLLOWING SECTION SETS T(N) TO 9.0 IF THAT
                   C  COMPOUND IS NOT TO BE USED IN THE TRAINING SET BECAUSE
                   C  OPTION 3 OR 4 HAS ALREADY BEEN SATISFIED.
                   C
                   C
ISN 0043     1082  N10=0
ISN 0044           N20=0
ISN 0045           N30=0
ISN 0046           IF(NOPT3 .EQ. 0) GO TO 4000
ISN 0048           DO 3999 I=1,NCONT
ISN 0049           IF(T(I))3001,3001,3002
ISN 0050     3001  IF(N10-NOPT3) 3100,3101,3101
ISN 0051     3100  N10=N10+1
ISN 0052           GO TO 3999
ISN 0053     3101  T(I) = 9.0
ISN 0054           N30=N30+1
ISN 0055           GO TO 3999
ISN 0056     3002  IF(N20-NOPT4) 3200,3201,3201
ISN 0057     3200  N20=N20+1
ISN 0058           GO TO 3999
ISN 0059     3201  T(I) = 9.0
ISN 0060           N30=N30+1
ISN 0061     3999  CONTINUE
ISN 0062     4000  N30=NCONT-N30
ISN 0063           RETURN
ISN 0064           END
```

```
LEVEL 21.7 ( JAN 73 )                    OS/360  FORTRAN H

          COMPILER OPTIONS - NAME=  MAIN,OPT=02,LINECNT=58,SIZE=0000K,
                             SOURCE,EBCDIC,NOLIST,NODECK,LOAD,MAP,NOEDIT,ID,NOXREF
                C
                C
                C
                C
                C
                C
                C
                C
                C
                C
 ISN 0002             SUBROUTINE REOP(NOPT1,NCARD,NOPT2,FIN,NSHUT,LEN,NOPT3,NOPT4,NOPT5,
                     1NOPT6)
 ISN 0003             COMMON FX(400,114),JC(112),T(400),W(113),NERR(400),TERR(400)
                C
                C  NOPT1 = 1 : IR DATA ONLY
                C  NOPT1 = 2 : RAMAN NONPOLARIZED DATA ONLY
                C  NOPT1 = 3 : RAMAN PARALLEL DATA ONLY
                C  NOPT1 = 4 : RAMAN PERPENDICULAR DATA ONLY
                C  NOPT1 = 5 : IR PLUS RAMAN NONPOLARIZED DATA
                C  NOPT1 = 6 : IR PLUS RAMAN POLARIZED (PARALLEL AND PERPENDICULAR)
                C  NOPT1 = 7 : IR PLUS ALL RAMAN DATA
                C  NOPT1 = 8 : ALL RAMAN DATA
                C  NOPT1 = 9 : RAMAN PARALLEL + PERPENDICULAR
                C
                C
                C  NCARD IS THE NUMBER OF SPECTRA IN THE DECK.
                C
                C
                C  NOPT2 = 1 THRU 17 : CLASS 1 THRU 17
                C  NOPT2 = 18 : CLASSES 7,8,9AND 10(ALL HALOGENS)
                C  NOPT2 = 19 : CLASSES 8,9,AND 10 (F,BR AND I)
                C  NOPT2 = 20 : CLASSES 3,4 AND 5(ALCOHOLS)
                C  NOPT2 = 21 : CLASSES 3,4,5 AND 16
                C
                C
                C  FIN IS THE INITIAL VALUE OF THE VECTOR W(I) WHICH IS TO BE TRAINED.
                C  IF FIN IS 0.0, THE PROGRAM WILL READ AN INITIAL
                C  VECTOR FROM FILE 73.  NOTE THAT IF THE PROGRAM IS TO BE USED IN THIS
                C  MODE THE APPROPRIATE FILE CARD TO DEFINE FILE 73 (READ(73)) MUST
                C  BE USED.
                C
                C
                C  NSHUT LIMITS THE NUMBER OF ITERATIONS .
                C
                C
                C  LEN DETERMINES HOW OFTEN THE UNTRAINED VECTOR
                C  WILL BE PRINTED.  IF LEN=10,THE UNTRAINED VECTOR WILL BE
                C  PRINTED EVERY 10TH ITERATION.  IF LEN=1,THE UNTRAINED
                C  VECTOR WILL BE PRINTED AT EVERY ITERATION.
                C
                C
                C  NOPT3 = THE NUMBER OF SPECTRA NOT IN THE CLASS WHICH ARE TO BE
                C  INCLUDED IN THE DECK FOR TRAINING.
                C  NOPT4 = THE NUMBER OF SPECTRA IN THE CLASS WHICH ARE TO BE
```

```
C  INCLUDED IN THE DECK.
C  IF NOPT3 = 00, THE PROGRAM WILL READ THE ENTIRE DATA DECK
C  AND COUNT THE NUMBER OF SPECTRA IN THE CLASS AND THE NUMBER
C  OF SPECTRA NOT IN THE CLASS.
C
C
C  NOPT6 PERMITS A PART OF THE TRAINING VECTORS TO
C  BE USED FOR THE TRAINING ROUTINE.  IF NOPT6 IS 1, EVERY ELEMENT
C  OF THE VECTORS WILL BE USED.  IF,HOWEVER, NOPT6 EQUALS N,
C  EVERY NTH ELEMENT WILL BE USED FOR THE TRAINING ROUTINE.
C  AFTER THE PROGRAM TRAINS W(I) WITH EVERY NTH ELEMENT BEING USED
C  IT WRITES THE VECTOR ON THE DISK AND PLACES THE TRAINED ELEMENTS
C  IN THE N-1 POSITIONS OF W(I) IMMEDIATELY TO THE RIGHT OF THE ELEMENT USED FOR
C  TRAINING.  THE PROGRAM THEN USES THIS NEW VECTOR AS THE STARTING VECTOR FOR
C  THE WHOLE ROUTINE.  IN THE EVENT THAT NSHUT IS REACHED BEFORE ANY VECTOR
C  IS TRAINED (EITHER USING EVERY ELEMENT OR EVERY NTH ELEMENT), THE
C  PROGRAM STOPS TRAINING AND CONTINUES WITH THE PARTIALLY TRAINED VECTOR.
C
C
C  IF NOPT5=0000,THE PROGRAM WILL END
C  IF,HOWEVER,NOPT5 = 1111 THE PROGRAM WILL ZERO ALL MATRICIES AND
C  START OVER AGAIN.  MORE DATA IS EXPECTED.
C
ISN 0004          READ(5,999)NOPT1,NCARD,NOPT2,FIN,NSHUT,LEN,NOPT3,NOPT4,
                 1NOPT6
ISN 0005          READ(5,998)NOPT5
ISN 0006    999   FORMAT(I4,2X,I4,2X,I4,2X,F9.4,2X,I4,2X,I4,2X,I4,2X,I4,
                 12X,I4)
ISN 0007    998   FORMAT(I4)
ISN 0008          RETURN
ISN 0009          END
```

```
LEVEL 21.7 ( JAN 73 )                    OS/360  FORTRAN H

          COMPILER OPTIONS - NAME=  MAIN,OPT=02,LINECNT=58,SIZE=0000K,
                             SOURCE,EBCDIC,NOLIST,NODECK,LOAD,MAP,NOEDIT,ID,NOXREF
 ISN 0002            SUBROUTINE PLOT
 ISN 0003            COMMON FX(400,114),JC(112),T(400),W(113),NERR(400),TERR(400)
 ISN 0004            DIMENSION OUT(50,113)
 ISN 0005            DATA ARF/'**'/,BARF/' '/
 ISN 0006            TOP=0.0
 ISN 0007            BOT=0.0
 ISN 0008            DO 10 I=1,113
 ISN 0009            TOP = AMAX1(W(I),TOP)
 ISN 0010            BOT = AMIN1(W(I),BOT)
 ISN 0011       10   CONTINUE
 ISN 0012            TOPA=TOP
 ISN 0013            DIST=TOP - BOT
 ISN 0014            FIN = DIST/49
 ISN 0015       999  FORMAT(1A1)
 ISN 0016            K=0
 ISN 0017            DO150 I=1,50
 ISN 0018            IF(K) 95,91,95
 ISN 0019       91   IF(TOP)94,94,95
 ISN 0020       94   K=I
 ISN 0021       95   DO 104 J=1,113
 ISN 0022            IF(W(J) - TOP) 24,25,25
 ISN 0023       25   OUT(I,J)=ARF
 ISN 0024            GO TO 104
 ISN 0025       24   OUT(I,J) = BARF
 ISN 0026       104  CONTINUE
 ISN 0027       150  TOP = TOP - FIN
                     C
                     C  THE FOLLOWING IS THE WRITE ALGORITHM.
                     C
 ISN 0028            WRITE(6,899)TOPA,(OUT(1,J),J=1,113)
 ISN 0029            DO 300 I=2,49
 ISN 0030            WRITE(6,897)(OUT(I,J),J=1,113)
 ISN 0031            IF(I-K)300,202,300
 ISN 0032       202  WRITE(6,898)
 ISN 0033       300  CONTINUE
 ISN 0034            WRITE(6,893)BOT
 ISN 0035            WRITE(6,896)
 ISN 0036            WRITE(6,895)
 ISN 0037            WRITE(6,894)
 ISN 0038       899  FORMAT(1H1,5X,F7.3,1H*,112A1,6X,1A1)
 ISN 0039       898  FORMAT(1H+,7X,1H0,124(1H-))
 ISN 0040       897  FORMAT(1H ,12X,1H*,112A1,6X,1A1)
 ISN 0041       896  FORMAT(1H ,12X,1H*,20(1X,1H*),18(3X,1H*),6X,1H*)
 ISN 0042       895  FORMAT(1H ,14X,1H*,7X,1H*,3(9X,1H*),3(19X,1H*),11X,1H*,2X,5HCONST)
 ISN 0043       894  FORMAT(1H ,12X,4H3900,4X,4H3500,6X,4H3000,6X,4H2500,6X,
                    14H2000,16X,4H1500,16X,4H1000,17X,3H500,9X,3H200)
 ISN 0044       893  FORMAT(1H ,5X,F7.3,113(1H*),6X,1H*)
 ISN 0045            WRITE(6,799)TOPA,BOT
 ISN 0046       799  FORMAT(1H1,'THE RANGE IS FROM',F9.5,' TO ',F9.5)
 ISN 0047            WRITE(6,798) FIN
 ISN 0048       798  FORMAT(1H ,'EACH DIVISION ON THE PLOT IS 'F9.5,' UNITS.')
 ISN 0049            WRITE(6,899)TOPA,(OUT(1,J),J=1,113)
 ISN 0050            DO 600 I=2,49
 ISN 0051            IF(I-K) 602,602,600
```

```
ISN 0052      602   WRITE(6,897) (OUT(I,J),J=1,113)
ISN 0053      600   CONTINUE
ISN 0054            WRITE(6,898)
ISN 0055            WRITE(6,796)
ISN 0056      796   FORMAT(1H ,12X,1H*,20(1X,1H*),18(3X,1H*))
ISN 0057            WRITE(6,795)
ISN 0058      795   FORMAT(1H ,14X,1H*,7X,1H*,3(9X,1H*),3(19X,1H*),11X,1H*)
ISN 0059            WRITE(6,894)
ISN 0060      700   FORMAT(I4)
ISN 0061            RETURN
ISN 0062            END
```

```
LEVEL 21.7 ( JAN 73 )                          OS/360   FORTRAN H

          COMPILER OPTIONS - NAME=  MAIN,OPT=02,LINECNT=58,SIZE=0000K,
                             SOURCE,EBCDIC,NOLIST,NODECK,LOAD,MAP,NOEDIT,ID,NOXREF
  ISN 0002            SUBROUTINE UNPACK
  ISN 0003            COMMON FX(400,114),JC(112),T(400),W(113),NERR(400),TERR(400)
  ISN 0004            DIMENSION WA(113)
  ISN 0005            DO 1 I=1,113
  ISN 0006       1    WA(I) = W(I)
  ISN 0007            DO 2 I=1,44
  ISN 0008       2    W(I) = 0
  ISN 0009            J=0
  ISN 0010            DO 3 I=45,100
  ISN 0011            J=J+1
  ISN 0012       3    W(I) = WA(J)
  ISN 0013            DO 4 I=101,112
  ISN 0014       4    W(I) = 0
  ISN 0015            CALL PLOT
  ISN 0016            J=56
  ISN 0017            DO 5 I=45,100
  ISN 0018            J=J+1
  ISN 0019       5    W(I) = WA(J)
  ISN 0020            CALL PLOT
  ISN 0021            DO 7 I=1,113
  ISN 0022       7    W(I) = WA(I)
  ISN 0023            RETURN
  ISN 0024            END
```

```
LEVEL 21.7 ( JAN 73 )                          OS/360  FORTRAN H

          COMPILER OPTIONS - NAME=  MAIN,OPT=02,LINECNT=58,SIZE=0000K,
                             SOURCE,EBCDIC,NOLIST,NODECK,LOAD,MAP,NOEDIT,ID,NOXREF
  ISN 0002            SUBROUTINE CLASS15
  ISN 0003            COMMON FX(400,114),JC(112),T(400),W(113),NERR(400),TERR(400)
  ISN 0004            DO 10 I=160,174
  ISN 0005       10   T(I) = 1.0
  ISN 0006            RETURN
  ISN 0007            END
```

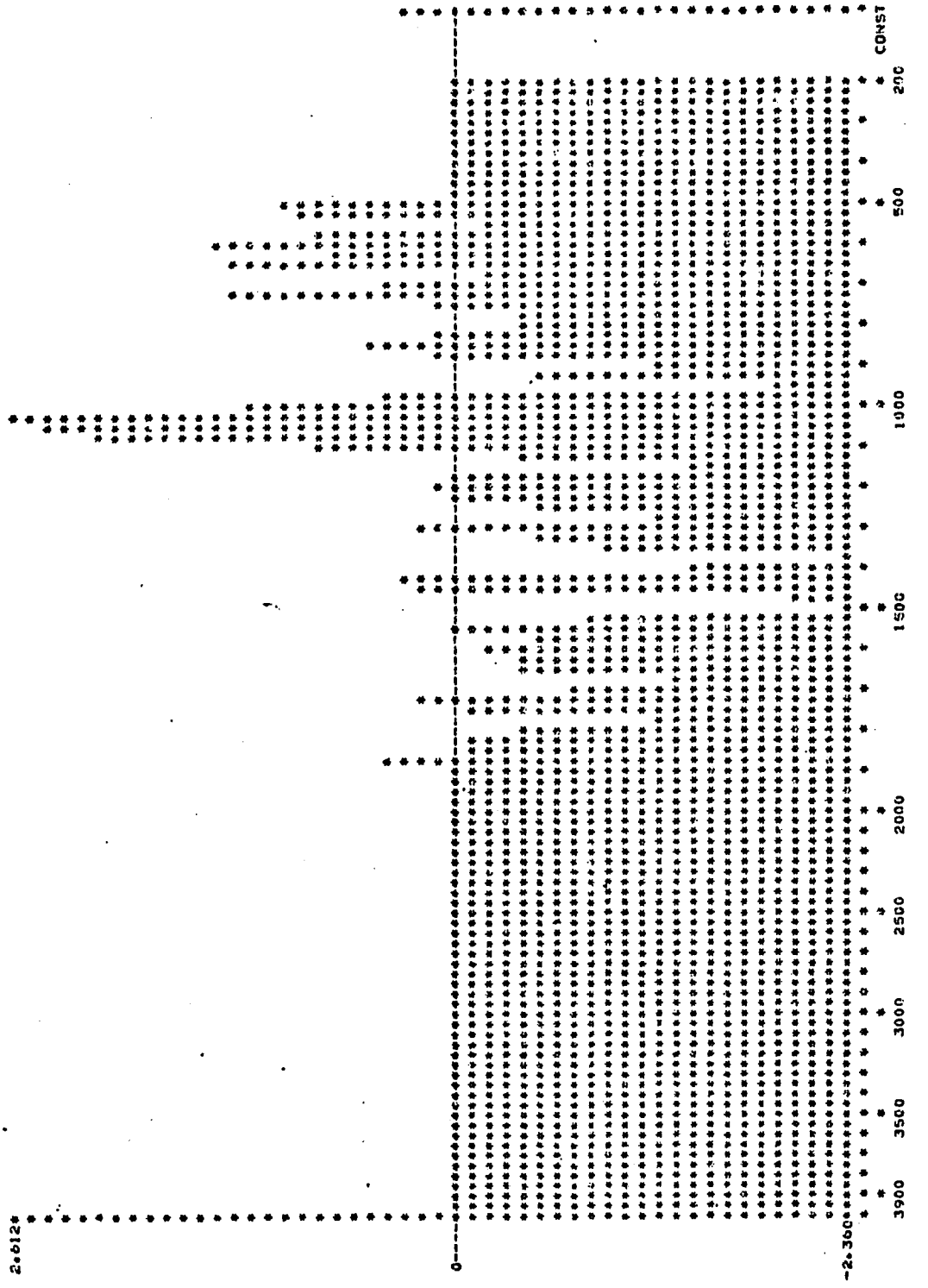
The output shown on the following pages was generated by program TRAIN from the three lines of input shown below. (Each space on the fortran card is shown by a dash.)

```
   1  1117     20     +1.0000    100    200    177     22      1
0000
 2.0
```

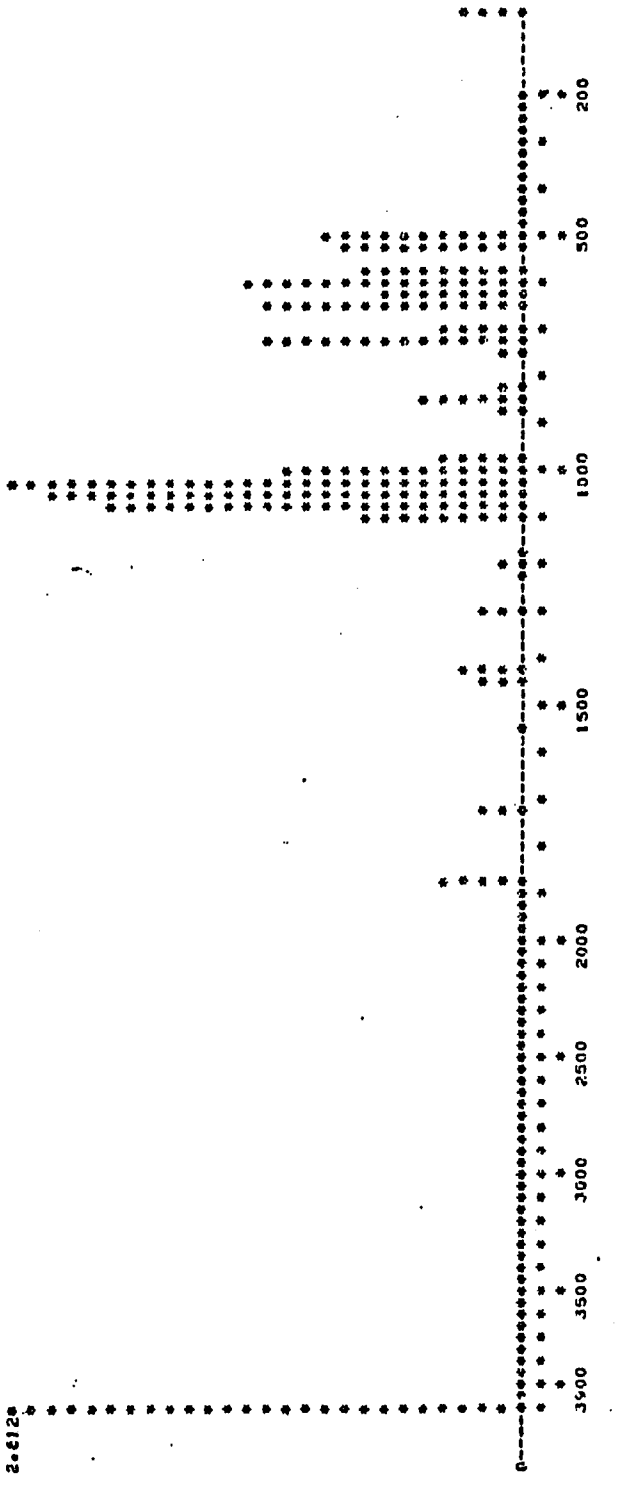
```
THE DATA HAVE BEEN READ    18   TIMES
THE FOLLOWING IS THE TRAINED VECTOR
  0.44463 -0.08075 -0.07861 -0.35825 -1.14690 -0.05585  0.25123 -0.69059 -1.27795 -0.40287
 -0.38603 -0.22513 -0.53108 -0.01810 -0.78272 -2.35986 -2.02481  0.22525  0.30821 -1.40743
 -2.28163 -0.87661 -0.52956  0.27117 -1.19552 -0.46860  0.00326  0.15155 -0.00935 -1.40902
 -0.41946  0.88202  2.11065  2.41291  2.61166  1.23642  0.39444 -1.89345 -0.47185 -1.16293
  0.13072  0.55187  0.16916 -0.37205 -0.35397  0.14132  1.31463  0.47219 -0.20140  1.34689
  0.77279  1.45850  0.86057 -0.13755  0.89046  1.01151  0.66423  0.20154 -0.15637  0.51626
  0.31259  1.36355  1.25963  0.48859 -0.28144 -0.43815 -0.94010 -1.03402 -1.43875  0.07842
  0.29638  0.93576 -1.45690  0.96986  0.50012 -0.95333 -1.32729 -0.09421 -0.62487  0.25418
 -2.21942 -0.61693  0.24854  0.52467 -0.63685 -0.39788 -0.21255 -0.33621 -0.07330  0.07126
  0.62020 -1.39361 -0.52905 -2.04557 -0.32956 -0.68659  1.76589  1.89221 -1.92398 -1.13274
 -0.09665 -1.65210  0.75314 -0.64107 -0.50927 -1.22728 -1.82229 -0.37610 -0.53357  0.50679
 -0.32441  0.42915  0.30108


THE DECK WAS COMPOSED OF  177.0( 88.94 PERCENT ) COMPOUNDS NOT IN THE CLASS, AND
  22.0( 11.06 PERCENT ) IN THE CLASS.
```

| COM | CLASS | ERR | COM | CLASS | ERR | COM | CLASS | ERR | COM | CLASS | ERR | COM | CLASS | ERR | COM | CLASS | ERR | COM | CLASS | ERR | COM | CLASS | ERR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.0 | 0.0 | 1 | 2.0 | 0.0 | 0 | 3.0 | 0.0 | 1 | 4.0 | 0.0 | 2 | 5.0 | 0.0 | 0 | 6.0 | 0.0 | 1 | 7.0 | 0.0 | 0 | 8.0 | 0.0 | 1 |
| 9.0 | 0.0 | 0 | 10.0 | 0.0 | 0 | 11.0 | 0.0 | 0 | 12.0 | 0.0 | 0 | 13.0 | 0.0 | 0 | 14.0 | 0.0 | 0 | 15.0 | 0.0 | 0 | 16.0 | 0.0 | 0 |
| 17.0 | 0.0 | 3 | 18.0 | 0.0 | 1 | 19.0 | 0.0 | 4 | 20.0 | 0.0 | 1 | 21.0 | 0.0 | 1 | 22.0 | 0.0 | 2 | 23.0 | 0.0 | 0 | 24.0 | 0.0 | 0 |
| 25.0 | 0.0 | 0 | 26.0 | 0.0 | 0 | 27.0 | 0.0 | 0 | 28.0 | 0.0 | 0 | 29.0 | 0.0 | 0 | 30.0 | 0.0 | 0 | 31.0 | 0.0 | 10 | 32.0 | 0.0 | 0 |
| 33.0 | 0.0 | 0 | 34.0 | 0.0 | 19 | 35.0 | 0.0 | 0 | 36.0 | 0.0 | 0 | 37.0 | 0.0 | 0 | 38.0 | 0.0 | 7 | 39.0 | 0.0 | 0 | 40.0 | 0.0 | 0 |
| 41.0 | 0.0 | 0 | 42.0 | 0.0 | 0 | 43.0 | 0.0 | 0 | 44.0 | 0.0 | 0 | 45.0 | 0.0 | 14 | 46.0 | 0.0 | 0 | 47.0 | 0.0 | 0 | 48.0 | 0.0 | 0 |
| 49.0 | 0.0 | 0 | 50.0 | 0.0 | 0 | 51.0 | 0.0 | 0 | 52.0 | 0.0 | 0 | 53.0 | 0.0 | 0 | 54.0 | 0.0 | 0 | 55.0 | 0.0 | 0 | 56.0 | 0.0 | 0 |
| 57.0 | 0.0 | 0 | 58.0 | 0.0 | 0 | 59.0 | 0.0 | 0 | 60.0 | 0.0 | 0 | 61.0 | 0.0 | 0 | 62.0 | 0.0 | 0 | 63.0 | 0.0 | 0 | 64.0 | 0.0 | 0 |
| 65.0 | 0.0 | 0 | 66.0 | 0.0 | 0 | 67.0 | 0.0 | 0 | 68.0 | 0.0 | 0 | 69.0 | 0.0 | 0 | 70.0 | 0.0 | 8 | 71.0 | 0.0 | 0 | 72.0 | 0.0 | 0 |
| 73.0 | 0.0 | 0 | 74.0 | 0.0 | 0 | 75.0 | 0.0 | 0 | 76.0 | 0.0 | 0 | 77.0 | 0.0 | 0 | 78.0 | 0.0 | 0 | 79.0 | 0.0 | 0 | 80.0 | 0.0 | 0 |
| 81.0 | 0.0 | 0 | 82.0 | 0.0 | 0 | 83.0 | 0.0 | 0 | 84.0 | 0.0 | 0 | 85.0 | 1.0 | 4 | 86.0 | 1.0 | 5 | 87.0 | 1.0 | 7 | 88.0 | 1.0 | 15 |
| 89.0 | 0.0 | 1 | 90.0 | 0.0 | 1 | 91.0 | 0.0 | 0 | 92.0 | 0.0 | 9 | 93.0 | 0.0 | 0 | 94.0 | 0.0 | 0 | 95.0 | 0.0 | 0 | 96.0 | 0.0 | 0 |
| 97.0 | 0.0 | 0 | 98.0 | 0.0 | 0 | 99.0 | 0.0 | 0 | 100.0 | 0.0 | 5 | 101.0 | 0.0 | 0 | 102.0 | 0.0 | 0 | 103.0 | 0.0 | 1 | 104.0 | 0.0 | 0 |
| 105.0 | 0.0 | 0 | 106.0 | 0.0 | 0 | 107.0 | 0.0 | 0 | 108.0 | 0.0 | 0 | 109.0 | 0.0 | 0 | 110.0 | 0.0 | 4 | 111.0 | 0.0 | 0 | 112.0 | 0.0 | 1 |
| 113.0 | 0.0 | 0 | 114.0 | 0.0 | 1 | 115.0 | 0.0 | 0 | 116.0 | 0.0 | 0 | 117.0 | 0.0 | 0 | 118.0 | 0.0 | 1 | 119.0 | 0.0 | 0 | 120.0 | 0.0 | 0 |
| 121.0 | 0.0 | 0 | 122.0 | 0.0 | 0 | 123.0 | 0.0 | 0 | 124.0 | 0.0 | 0 | 125.0 | 0.0 | 0 | 126.0 | 0.0 | 0 | 127.0 | 0.0 | 0 | 128.0 | 0.0 | 9 |
| 129.0 | 0.0 | 0 | 130.0 | 0.0 | 0 | 131.0 | 0.0 | 3 | 132.0 | 0.0 | 0 | 133.0 | 0.0 | 0 | 134.0 | 0.0 | 0 | 135.0 | 0.0 | 0 | 136.0 | 0.0 | 8 |
| 137.0 | 0.0 | 0 | 138.0 | 0.0 | 0 | 139.0 | 0.0 | 0 | 140.0 | 0.0 | 0 | 141.0 | 0.0 | 0 | 142.0 | 0.0 | 0 | 143.0 | 0.0 | 0 | 144.0 | 0.0 | 0 |
| 145.0 | 0.0 | 0 | 146.0 | 0.0 | 0 | 147.0 | 0.0 | 0 | 148.0 | 0.0 | 0 | 149.0 | 0.0 | 0 | 150.0 | 0.0 | 0 | 151.0 | 0.0 | 0 | 152.0 | 0.0 | 0 |
| 153.0 | 0.0 | 0 | 154.0 | 0.0 | 0 | 155.0 | 0.0 | 0 | 156.0 | 0.0 | 0 | 157.0 | 0.0 | 0 | 158.0 | 0.0 | 0 | 159.0 | 0.0 | 0 | 160.0 | 0.0 | 0 |
| 161.0 | 0.0 | 0 | 162.0 | 0.0 | 0 | 163.0 | 0.0 | 0 | 164.0 | 0.0 | 0 | 165.0 | 0.0 | 0 | 166.0 | 0.0 | 0 | 167.0 | 0.0 | 0 | 168.0 | 0.0 | 0 |
| 169.0 | 0.0 | 0 | 170.0 | 0.0 | 0 | 171.0 | 0.0 | 0 | 172.0 | 0.0 | 0 | 173.0 | 0.0 | 0 | 174.0 | 0.0 | 0 | 175.0 | 0.0 | 0 | 176.0 | 1.0 | 2 |
| 177.0 | 1.0 | 0 | 178.0 | 1.0 | 3 | 179.0 | 1.0 | 3 | 180.0 | 1.0 | 3 | 181.0 | 1.0 | 0 | 182.0 | 1.0 | 0 | 183.0 | 1.0 | 0 | 184.0 | 1.0 | 0 |
| 185.0 | 1.0 | 5 | 186.0 | 1.0 | 0 | 187.0 | 1.0 | 0 | 188.0 | 1.0 | 0 | 189.0 | 1.0 | 0 | 190.0 | 1.0 | 0 | 191.0 | 1.0 | 12 | 192.0 | 1.0 | 0 |
| 193.0 | 1.0 | 7 | 194.0 | 9.0 | 0 | 195.0 | 9.0 | 0 | 196.0 | 9.0 | 0 | 197.0 | 9.0 | 0 | 198.0 | 9.0 | 0 | 199.0 | 9.0 | 0 | 200.0 | 9.0 | 0 |
| 201.0 | 9.0 | 0 | 202.0 | 9.0 | 0 | 203.0 | 9.0 | 0 | 204.0 | 9.0 | 0 | 205.0 | 9.0 | 0 | 206.0 | 9.0 | 0 | 207.0 | 9.0 | 0 | 208.0 | 9.0 | 0 |
| 209.0 | 9.0 | 0 | 210.0 | 9.0 | 0 | 211.0 | 9.0 | 0 | 212.0 | 9.0 | 0 | 213.0 | 9.0 | 0 | 214.0 | 9.0 | 0 | 215.0 | 9.0 | 0 | 216.0 | 9.0 | 0 |
| 217.0 | 0.0 | 1 | 218.0 | 0.0 | 2 | 219.0 | 0.0 | 0 | 220.0 | 0.0 | 0 | 221.0 | 0.0 | 0 | 222.0 | 0.0 | 1 | 223.0 | 9.0 | 0 | 224.0 | 9.0 | 0 |
| 225.0 | 9.0 | 0 | 226.0 | 9.0 | 0 | 227.0 | 9.0 | 0 | 228.0 | 9.0 | 0 | 229.0 | 9.0 | 0 | 230.0 | 9.0 | 0 | 231.0 | 9.0 | 0 | 232.0 | 9.0 | 0 |
| 233.0 | 9.0 | 0 | 234.0 | 9.0 | 0 | 235.0 | 9.0 | 0 | 236.0 | 9.0 | 0 | 237.0 | 9.0 | 0 | 238.0 | 9.0 | 0 | 239.0 | 9.0 | 0 | 240.0 | 9.0 | 0 |
| 241.0 | 9.0 | 0 | 242.0 | 9.0 | 0 | 243.0 | 9.0 | 0 | 244.0 | 9.0 | 0 | 245.0 | 9.0 | 0 | 246.0 | 9.0 | 0 | 247.0 | 9.0 | 0 | 248.0 | 9.0 | 0 |
| 249.0 | 9.0 | 0 | 250.0 | 9.0 | 0 | 251.0 | 9.0 | 0 | 252.0 | 9.0 | 0 | 253.0 | 9.0 | 0 | 254.0 | 9.0 | 0 | 255.0 | 9.0 | 0 | 256.0 | 9.0 | 0 |
| 257.0 | 9.0 | 0 | 258.0 | 9.0 | 0 | 259.0 | 9.0 | 0 | 260.0 | 9.0 | 0 | 261.0 | 9.0 | 0 | 262.0 | 9.0 | 0 | 263.0 | 9.0 | 0 | 264.0 | 9.0 | 0 |
| 265.0 | 9.0 | 0 | 266.0 | 9.0 | 0 | 267.0 | 9.0 | 0 | 268.0 | 9.0 | 0 | 269.0 | 9.0 | 0 | 270.0 | 9.0 | 0 | 271.0 | 9.0 | 0 | 272.0 | 9.0 | 0 |
| 273.0 | 9.0 | 0 | 274.0 | 9.0 | 0 | 275.0 | 9.0 | 0 | 276.0 | 9.0 | 0 | 277.0 | 9.0 | 0 | 278.0 | 9.0 | 0 | 279.0 | 9.0 | 0 | 280.0 | 9.0 | 0 |
| 281.0 | 9.0 | 0 | 282.0 | 9.0 | 0 | 283.0 | 9.0 | 0 | 284.0 | 9.0 | 0 | 285.0 | 9.0 | 0 | 286.0 | 9.0 | 0 | 287.0 | 9.0 | 0 | 288.0 | 9.0 | 0 |
| 289.0 | 9.0 | 0 | 290.0 | 9.0 | 0 | 291.0 | 9.0 | 0 | 292.0 | 9.0 | 0 | 293.0 | 9.0 | 0 | 294.0 | 9.0 | 0 | 295.0 | 9.0 | 0 | 296.0 | 9.0 | 0 |
| 297.0 | 9.0 | 0 | 298.0 | 9.0 | 0 | 299.0 | 9.0 | 0 | 300.0 | 9.0 | 0 | 301.0 | 9.0 | 0 | 302.0 | 9.0 | 0 | 303.0 | 9.0 | 0 | 304.0 | 9.0 | 0 |
| 305.0 | 9.0 | 0 | 306.0 | 9.0 | 0 | 307.0 | 9.0 | 0 | 308.0 | 9.0 | 0 | 309.0 | 9.0 | 0 | 310.0 | 9.0 | 0 | 311.0 | 9.0 | 0 | 312.0 | 9.0 | 0 |
| 313.0 | 9.0 | 0 | 314.0 | 9.0 | 0 | 315.0 | 9.0 | 0 | 316.0 | 9.0 | 0 | 317.0 | 9.0 | 0 | 318.0 | 9.0 | 0 | 319.0 | 9.0 | 0 | 320.0 | 9.0 | 0 |
| 321.0 | 9.0 | 0 | 322.0 | 9.0 | 0 | 323.0 | 9.0 | 0 | 324.0 | 9.0 | 0 | 325.0 | 9.0 | 0 | 326.0 | 9.0 | 0 | 327.0 | 9.0 | 0 | 328.0 | 9.0 | 0 |
| 329.0 | 9.0 | 0 | 330.0 | 9.0 | 0 | 331.0 | 9.0 | 0 | 332.0 | 9.0 | 0 | 333.0 | 9.0 | 0 | 334.0 | 9.0 | 0 | 335.0 | 9.0 | 0 | 336.0 | 9.0 | 0 |
| 337.0 | 9.0 | 0 | 338.0 | 9.0 | 0 | 339.0 | 9.0 | 0 | 340.0 | 9.0 | 0 | 341.0 | 9.0 | 0 | 342.0 | 9.0 | 0 | 343.0 | 9.0 | 0 | 344.0 | 9.0 | 0 |
| 345.0 | 9.0 | 0 | 346.0 | 9.0 | 0 | 347.0 | 9.0 | 0 | 348.0 | 9.0 | 0 | 349.0 | 9.0 | 0 | 350.0 | 9.0 | 0 | 351.0 | 9.0 | 0 | 352.0 | 9.0 | 0 |
| 353.0 | 9.0 | 0 | 354.0 | 9.0 | 0 | 355.0 | 9.0 | 0 | 356.0 | 9.0 | 0 | 357.0 | 9.0 | 0 | 358.0 | 9.0 | 0 | 359.0 | 9.0 | 0 | 360.0 | 9.0 | 0 |
| 361.0 | 9.0 | 0 | 362.0 | 9.0 | 0 | 363.0 | 9.0 | 0 | 364.0 | 9.0 | 0 | 365.0 | 9.0 | 0 | 366.0 | 9.0 | 0 | 367.0 | 9.0 | 0 | 368.0 | 9.0 | 0 |
| 369.0 | 9.0 | 0 | 370.0 | 9.0 | 0 | 371.0 | 9.0 | 0 | 372.0 | 9.0 | 0 | 373.0 | 9.0 | 0 | 374.0 | 9.0 | 0 | 375.0 | 9.0 | 0 | 376.0 | 9.0 | 0 |
| 377.0 | 9.0 | 0 | 378.0 | 9.0 | 0 | 379.0 | 9.0 | 0 | 380.0 | 9.0 | 0 | 381.0 | 9.0 | 0 | 382.0 | 9.0 | 0 | 383.0 | 9.0 | 0 | 384.0 | 9.0 | 0 |
| 385.0 | 9.0 | 0 | 386.0 | 9.0 | 0 | 387.0 | 9.0 | 0 | 388.0 | 9.0 | 0 | 389.0 | 9.0 | 0 | 390.0 | 9.0 | 0 | 391.0 | 9.0 | 0 | 392.0 | 9.0 | 0 |
| 393.0 | 9.0 | 0 | 394.0 | 9.0 | 0 | 395.0 | 9.0 | 0 | 396.0 | 9.0 | 0 | 397.0 | 9.0 | 0 | 398.0 | 9.0 | 0 | 399.0 | 9.0 | 0 | 400.0 | 9.0 | 0 |

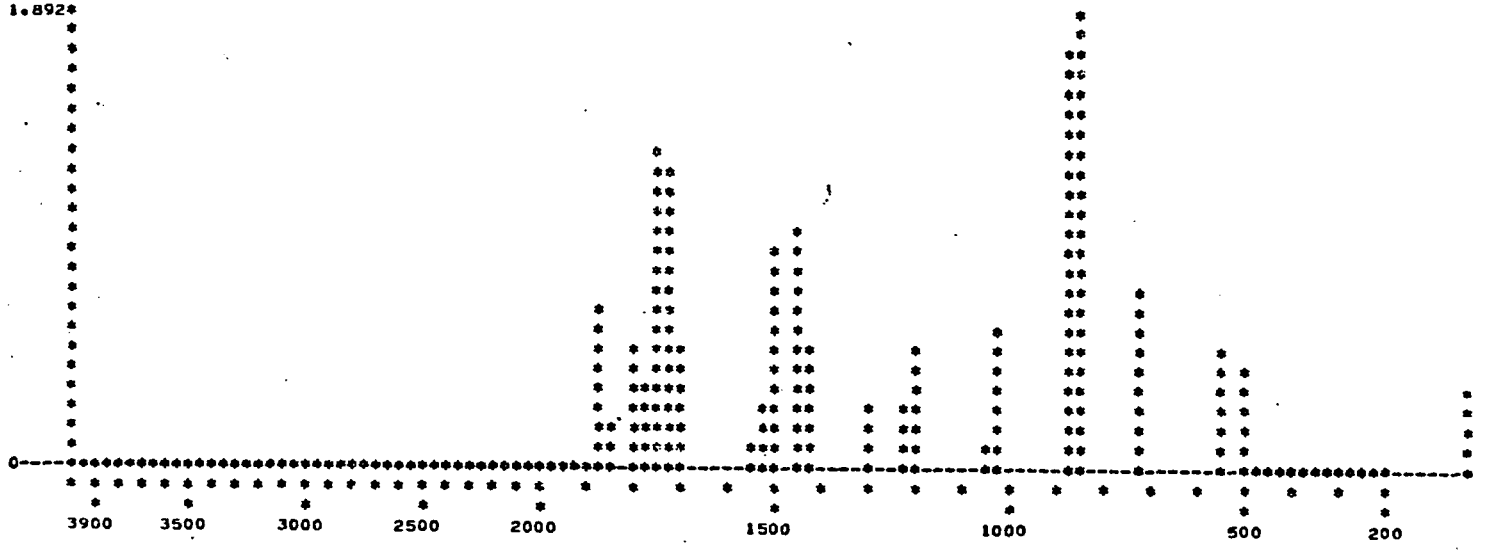

2.612E
0
-2.300E
3900
3500
3000
2500
2000
1500
1000
500
200
CONST

THE RANGE IS FROM  2.61166 TO  -2.35986
EACH DIVISION ON THE PLOT IS  0.10146 UNITS.


2.E12
0
500
1000
1500
2000
2500
3000
3500
4000


1.892+
0
-2.219+
CONST
200
500
1000
1500
2000
2500
3000
3500
3900

THE RANGE IS FROM  1.89221 TO  -2.21942
EACH DIVISION ON THE PLOT IS  0.08391 UNITS.




THE OPTIONS FOLLOWED WERE THE FOLLOWING:
OPTION 1 = 1
OPTION 2 = 20
OPTION 3 = 177
OPTION 4 = 22
OPTION 6 = 1
TOTAL NUMBER OF SPECTRA CONSIDERED (NCARD) = 1117
MAXIMUM NUMBER OF ITERATIONS (NSHUT) = 100
THE UNTRAINED VECTOR WAS PRINTED EVERY 200TH ITERATION
THE INITIAL VALUE OF THE VECTOR WAS (FIN) 1.0000

# PROGRAM TEST

```
      COMMON FX(400,114),JC(112),T(400),W(113),NERR(400),TERR(400)
      DIMENSION CI(400)
      N6 = 1
      NOPT6 = 1
1     FNT1 = 0
      FNT2=0
      FNT3=0
      FNT4=0
      K=0
      IF(NOPT6 .EQ. 1) GO TO 2
      IF(N6 .EQ. 1) GO TO 2
      N6=1
      GO TO 3
2     DO 8000 I=1,400
      T(I) = 0
      DO 8000 J=1,114
8000  FX(I,J) = 0.0
      CALL READP(NOPT1,NCARD,NOPT2,FIN,NSHUT,LEN,NOPT3,NOPT4,NOPT5,NOPT6)
C
C
C
C  NOPT7 IS EITHER 1 OF 2.  1 IF SUBROUTINE RE71 IS TO BE USED. ELSE 2.
C  NOPT8 GIVES THE POSITION ON THE FILE WHERE THE TEST VECTOR IS TO BE FOUND.
C  NOPT9 GIVES THE FILE NUMBER WHICH IS TO BE READ.
C
C
      READ(5,500) NOPT7,NOPT8,NOPT9
500   FORMAT(I4,2X,I4,2X,I4)
      N1=NOPT3
      N2=NOPT4
      GO TO (4,5),NOPT7
4     CALL RE71(NOPT1,NCARD,NCONT,NOPT2,NOPT3,NOPT4,N30)
      GO TO 6
5     CALL RE74(NOPT1,NCARD,NCONT,NOPT2,NOPT3,NOPT4,N30)
6     N6 = 2
3     DO 7 KNOP=1,NOPT8
7     READ(NOPT9)(W(I),I=1,113)
      DO 100 I=1,NCONT
      F=0.0
      DO 10 J=1,113
10    F=(W(J)*FX(I,J))+F
      IF(T(I)) 20,20,30
20    IF(F) 21,21,25
21    FNT1=FNT1+1
      GO TO 100
25    FNT2=FNT2+1
      K=K+1
      NERR(K) = FX(I,114)
      TERR(K) = T(I)
      CI(K) = F
      GO TO 100
30    IF(F) 31,31,35
31    FNT3 = FNT3 + 1
      K=K+1
      NERR(K) = FX(I,114)
      TERR(K) = T(I)
      CI(K) = F
      GO TO 100
35    FNT4 = FNT4 + 1
100   CONTINUE
      F1=FNT1
      F2=(FNT1/(FNT1+FNT2))*100.0
      F3=FNT2
      F4=(FNT2/(FNT1+FNT2))*100.0
```

```
      F5=FNT4
      F6=(FNT4/(FNT4+FNT3))*100.0
      F7=FNT3
      F8=(FNT3/(FNT4+FNT3))*100.0
      F9=FNT4+FNT1
      F11=FNT1+FNT2+FNT3+FNT4
      F10=((FNT4+FNT1)/F11)*100.0
      FN=N1+N2
      F12=F9-FN
      F14=F11-FN
      F13=(F12/F14)*100.0
      ND6 = 6
      WRITE(ND6,999)
  999 FORMAT(1H1,'      COMPOUNDS NOT IN THE CLASS')
      WRITE(ND6,998)F1,F2
  998 FORMAT(1H ,F6.1,'(',F6.2,' PERCENT) WERE CORRECTLY CLASSIFIED')
      WRITE(ND6,997)F3,F4
  997 FORMAT(1H ,F6.1,'(',F6.2,' PERCENT) WERE NOT CORRECTLY CLASSIFIED'
     1)
      N21 = IFIX(F1 + F3 ) - NOPT3
      N20 = IFIX(F1) - NOPT3
      F20 = (FLOAT(N20)/FLOAT(N21))*100.0
      WRITE(ND6,799) NOPT3
  799 FORMAT(1H ,'SINCE',I5,' COMPOUNDS WERE USED IN THE TRAINING SET.')
      WRITE(ND6,798) N20,F20,N21
  798 FORMAT(1H ,'THE PREDICTIVE ABILITY WAS',I4,'(',F6.2,' PERCENT) OUT
     1 OF',I4,' COMPOUNDS')
      WRITE(ND6,996)
      WRITE(ND6,996)
      WRITE(ND6,996)
      WRITE(ND6,996)
  996 FORMAT(1H )
      WRITE(ND6,995)
  995 FORMAT(1H ,'      COMPOUNDS IN THE CLASS')
      WRITE(ND6,998)F5,F6
      WRITE(ND6,997)F7,F8
      N24 = IFIX(F5 + F7) - NOPT4
      N23=IFIX(F5) - NOPT4
      F23 = (FLOAT(N23)/FLOAT(N24)) * 100.0
      WRITE(ND6,799) NOPT4
      WRITE(ND6,798) N23,F23,N24
      WRITE(ND6,996)
      WRITE(ND6,996)
      WRITE(ND6,996)
      WRITE(ND6,996)
      WRITE(ND6,994)
  994 FORMAT(1H ,4X,'ERROR LIST')
      WRITE(ND6,993)
  993 FORMAT(1H ,'COMPOUND   CLASSIFICATION')
      DO 40 I=1,K
   40 WRITE(ND6,992)NERR(I),TERR(I),C1(I)
  992 FORMAT(1H ,2X,I4,9X,F5.1,4X,E15.8)
      WRITE(ND6,990)
  990 FORMAT(1H1,'OVERALL RESULTS')
      WRITE(ND6,899)F9,F10,F11
  899 FORMAT(1H ,F6.2,'(',F6.2,' PERCENT) CORRECT OUT OF ',F6.2)
      WRITE(ND6,898)FN
  898 FORMAT(1H ,'THE TRAINING SET USED ',F6.2,' COMPOUNDS')
      WRITE(ND6,897)F12,F13,F14
  897 FORMAT(1H ,'THUS THE PREDICTIVE ABILITY WAS ',F6.2,'(',
     1F6.2,' PERCENT) OUT OF ',F6.2,' COMPOUNDS')
      REWIND 71
      REWIND NOPT9
      IF(NOPT6.EQ.1) GO TO 1001
      IF(N6 .EQ. 2) GO TO 1
 1001 IF(NOPT5)1000,1000,1
```

```
1000  CONTINUE
      STOP
      END
      SUBROUTINE RE71(NOPT1,NCARD,NCONT,NOPT2,NOPT3,NOPT4,N30)
      COMMON FX(400,114),JC(112),T(400),W(113),NERR(400),TERR(400)
C   NOPT1 = 1 : IR DATA ONLY
C   NOPT1 = 2 : RAMAN NONPOLARIZED DATA ONLY
C   NOPT1 = 3 : RAMAN PARALLEL DATA ONLY
C   NOPT1 = 4 : RAMAN PERPENDICULAR DATA ONLY
C   NOPT1 = 5 : IR PLUS RAMAN NONPOLARIZED DATA
C   NOPT1 = 6 : IR PLUS RAMAN POLARIZED (PARALLEL AND PERPENDICULAR)
C   NOPT1 = 7 : IR PLUS ALL RAMAN DATA
C   NOPT1 = 8 : ALL RAMAN DATA
C   NOPT1 = 9 : RAMAN PARALLEL + PERPENDICULAR
      NSKIP = 0
      NCONT=0
      GO TO (100,100,100,100,500,500,500,800,900),NOPT1
100   DO 199 J=1,NCARD
      READ(71)N1,N2,(JC(I),I=1,58)
      READ(71)(JC(I),I=59,112)
      IF(N2-NOPT1)199,101,199
101   NCONT=NCONT+1
      DO 102 I=1,112
102   FX(NCONT,I) = JC(I)
      FX(NCONT,113) = N1
199   CONTINUE
      GO TO 1000
500   DO 599 J=1,NCARD
      READ(71) N1,N2,(JC(I),I=1,58)
      READ(71) (JC(I),I=59,112)
      IF(N2-1) 501,501,510
501   NCONT=NCONT+1
      DO 502 I=1,112
502   FX(NCONT,I) = JC(I)
      FX(NCONT,113) = N1
      GO TO 599
510   GO TO (599,599,599,599,511,520,530,599,599),NOPT1
511   IF(N2-2)598,512,598
598   NSKIP = NSKIP +1
      IF(NSKIP .EQ. 1) GO TO 599
      NCONT = NCONT-1
      NSKIP = 0
      GO TO 599
512   DO 513 I=1,112
      FIX=JC(I)
513   FX(NCONT,I) = (FX(NCONT,I) + FIX)/2
      GO TO 599
520   IF(N2-2)599,599,521
521   NSKIP = NSKIP + 1
      DO 522 I=1,112
      FIX=JC(I)
522   FX(NCONT,I) = FX(NCONT,I) + FIX
      IF(NSKIP-2) 599,523,599
523   DO 524 I=1,112
524   FX(NCONT,I) = FX(NCONT,I) / 3
      NSKIP = 0
      GO TO 599
530   DO 531 I=1,112
      FIX=JC(I)
531   FX(NCONT,I) = FX(NCONT,I) + FIX
      IF(N2-3)532,594,533
532   DO 534 I=1,112
534   FX(NCONT,I) = FX(NCONT,I) / 2
      GO TO 599
533   DO 535 I=1,112
535   FX(NCONT,I) = FX(NCONT,I) / 3
```

```
      GO TO 1000
  800 DO 899 J=1,NCARD
      READ(71) N1,N2,(JC(I),I=1,58)
      READ(71)(JC(I),I=59,112)
      IF(N2-2) 899,801,802
  801 NCONT = NCONT+1
      DO 803 I=1,112
  803 FX(NCONT,I) = JC(I)
      FX(NCONT,113) = N1
      GO TO 899
  802 NCONT=NCONT+1
      DO 810 I=1,112
      FIX=JC(I)
  810 FX(NCONT,I) = FX(NCONT,I) + FIX
      FX(NCONT,113) = N1
      IF(N2-3) 899,898,811
  811 DO 812 I=1,112
  812 FX(NCONT,I) = FX(NCONT,I) / 2
      GO TO 899
  898 NCONT=NCONT-1
  899 CONTINUE
      GO TO 1000
  900 DO 999 J=1,NCARD
      READ(71) N1,N2,(JC(I),I=1,58)
      READ(71) (JC(I),I=59,112)
      IF(N2-3)999,901,902
  901 NCONT=NCONT+1
  902 DO 903 I=1,112
      FIX=JC(I)
  903 FX(NCONT,I) = FX(NCONT,I) + FIX
      IF(N2-3) 999,999,910
  910 DO911 I=1,112
  911 FX(NCONT,I) = FX(NCONT,I) / 2
  999 CONTINUE
C  NOPT2 = 1 THRU 17 : CLASS 1 THRU 17
C  NOPT2 = 18 : CLASSES 7,8,9 AND 10(ALL HALOGENS)
C  NOPT2 = 19 : CLASSES 8,9,AND 10 (F,BR AND I)
C  NOPT2 = 20 : CLASSES 3,4 AND 5(ALCOHOLS)
C  NOPT2 = 21 : CLASSES 3,4,5 AND 16
 1000 CONTINUE
      DO 1060 J=1,400
      READ(71) (JC(I),I=1,17)
      IF(NOPT2 - 18) 1001,1002,1003
 1001 N = JC(NOPT2)
      GO TO 1050
 1002 N=JC(7) + JC(8) + JC(9) + JC(10)
      GO TO 1050
 1003 IF(NOPT2-20) 1004,1005,1006
 1004 N=JC(8) + JC(9) + JC(10)
      GO TO 1050
 1005 N=JC(3) + JC(4) + JC(5)
      GO TO 1050
 1006 N=JC(3) + JC(4) + JC(5) + JC(16)
 1050 FT=N
      IF(FT) 1060,1060,1061
 1061 T(J) = 1
 1060 CONTINUE
      DO 2000 I=1,NCONT
      N=FX(I,113)
 2000 FX(I,114)=T(N)
      DO 2001 I=1,NCONT
      T(I) = FX(I,114)
      FX(I,114) = FX(I,113)
 2001 FX(I,113) = 1
      N30=0
```

```
      RETURN
      END
      SUBROUTINE REOP(NOPT1,NCARD,NOPT2,FIN,NSHUT,LEN,NOPT3,NOPT4,NOPT5)
      COMMON FX(400,114),JC(112),T(400),W(113),NERR(400),TERR(400)
C
C  NOPT1 = 1 : IR DATA ONLY
C  NOPT1 = 2 : RAMAN NONPOLARIZED DATA ONLY
C  NOPT1 = 3 : RAMAN PARALLEL DATA ONLY
C  NOPT1 = 4 : RAMAN PERPENDICULAR DATA ONLY
C  NOPT1 = 5 : IR PLUS RAMAN NONPOLARIZED DATA
C  NOPT1 = 6 : IR PLUS RAMAN POLARIZED (PARALLEL AND PERPENDICULAR)
C  NOPT1 = 7 : IR PLUS ALL RAMAN DATA
C  NOPT1 = 8 : ALL RAMAN DATA
C  NOPT1 = 9 : RAMAN PARALLEL + PERPENDICULAR
C
C
C  NCARD IS THE NUMBER OF SPECTRA IN THE DECK.
C
C
C  NOPT2 = 1 THRU 17 : CLASS 1 THRU 17
C  NOPT2 = 19 : CLASSES 7,8,9AND 10(ALL HALOGENS)
C  NOPT2 = 19 : CLASSES 8,9,AND 10 (F,BR AND I)
C  NOPT2 = 20 : CLASSES 3,4 AND 5(ALCOHOLS)
C  NOPT2 = 21 : CLASSES 3,4,5 AND 16
C
C
C  FIN IS THE INITIAL VALUE OF THE VECTOR W(I) WHICH IS TO BE TRAINED.
C  IF FIN IS 0.0, THE PROGRAM WILL READ AN INITIAL
C  VECTOR FROM FILE'73.  NOTE THAT IF THE PROGRAM IS TO BE USED IN THIS
C  MODE THE APPROPRIATE FILE CARD TO DEFINE FILE 73 (DSET73) MUST
C  BE USED.
C
C
C  NSHUT LIMITS THE NUMBER OF ITERATIONS .
C
C
C  LEN DETERMINES HOW OFTEN THE UNTRAINED VECTOR
C  WILL BE PRINTED.  IF LEN=10,THE UNTRAINED VECTOR WILL BE
C  PRINTED EVERY 10TH ITERATION.  IF LEN=1,THE UNTRAINED
C  VECTOR WILL BE PRINTED AT EVERY ITERATION.
C
C
C  NOPT3 = THE NUMBER OF SPECTRA NOT IN THE CLASS WHICH ARE TO BE
C  INCLUDED IN THE DECK FOR TRAINING.
C  NOPT4 = THE NUMBER OF SPECTRA IN THE CLASS WHICH ARE TO BE
C  INCLUDED IN THE DECK.
C  IF NOPT3 = 00, THE PROGRAM WILL READ THE ENTIRE DATA DECK
C  AND COUNT THE NUMBER OF SPECTRA IN THE CLASS AND THE NUMBER
C  OF SPECTRA NOT IN THE CLASS.
C  A READ ERROR WILL OCCUR IF THE PROGRAM IS INSTRUCTED TO
C  READ MORE COMPOUNDS EITHER IN THE CLASS OR NOT IN THE CLASS THAN
C  ACTUALLY EXIST IN THE DECK.
C
C
C  IF NOPT5=0000,THE PROGRAM WILL END
C  IF,HOWEVER,NOPT5 = 1111 THE PROGRAM WILL ZERO ALL MATRICIES AND
C  START OVER AGAIN.  MORE DATA IS EXPECTED.
C
      READ(5,999)NOPT1,NCARD,NOPT2,FIN,NSHUT,LEN,NOPT3,NOPT4
      READ(5,998)NOPT5
999   FORMAT(I4,2X,I4,2X,I4,2X,F9.4,2X,I4,2X,I4,2X,I4,2X,I4)
998   FORMAT(I4)
      RETURN
      END
      SUBROUTINE RE74(NOPT1,NCARD,NCONT,NOPT2,NOPT3,NOPT4,N30)
      COMMON FX(400,114),JC(112),T(400),W(113),NERR(400),TERR(400)
```

```
      NCONT = 0
      DO 200 KON = 1,NCARD
      READ(71) N1,N2,(JC(I),I=1,58)
      READ(71)(JC(I),I=59,112)
      IF(N2 .GT. 1) GO TO 100
      NCONT = NCONT + 1
      DO 1 I=1,56
      J=I+44
1     FX(NCONT,I) = JC(J)
      FX(NCONT,114) = N1
      GO TO 200
100   IF(N2 .EQ. 4) GO TO 200
      J=45
      DO 101 I=57,112
      FX(NCONT,I) = JC(J)
101   J=J+1
200   CONTINUE
1000  CONTINUE
      DO 1060 J=1,400
      READ(71) (JC(I),I=1,17)
      IF(NOPT2 - 18) 1001,1002,1003
1001  N = JC(NOPT2)
      GO TO 1050
1002  N=JC(7) + JC(8) + JC(9) + JC(10)
      GO TO 1050
1003  IF(NOPT2-20) 1004,1005,1006
1004  N=JC(8) + JC(9) + JC(10)
      GO TO 1050
1005  N=JC(3) + JC(4) + JC(5)
      GO TO 1050
1006  N=JC(3) + JC(4) + JC(5) + JC(16)
1050  IF(N) 1060,1060,1061
1061  T(J) = 1
1060  CONTINUE
      N30 = 0
      RETURN
      END
// THOR END
```

PROGRAM SNOOP

Subroutines RE71 and PLOT are not included in this listing. They were included in the listing of earlier programs.

```
      COMMON FX(400,114),JC(112),T(400),W(113)
      DIMENSION W1(113)
C  IF N1 = 1, THE PROGRAM AVERAGES VECTORS
C  WHEN USED IN THIS MODE N2 EQUALS THE NUMBER OF VECTORS TO BE AVERAGED.
C  N5 EQUALS 00 IF THE VECTOR CALCULATED (AVERAGED) IS NOT TO BE WRITTEN ON
C  A DISK FILE.
C  N3 AND N4 GIVE THE FILE NUMBER AND THE POSITION ON THE FILE TO BE READ.
C  THERE SHOULD BE N2 CARDS SHOWING N3,N4 DATA.
C
C
C  IF N1 EQUALS 2 THE PROGRAM WILL CALCULATE THE AVERAGE REPRESENTATION
C  (A VECTOR) FOR A PARTICULAR CLASS.
C  IF THE PROGRAM IS USED IN THIS MODE, N2 IS 00.  N5 AGAIN IS 00 IF IT IS NOT
C  DESIRED TO WRITE THE CALCULATED VECTOR ON A DISK FILE.
C  IF THE PROGRAM IS USED WITH N2 SET AT 2 THE SECOND CARD MUST GIVE
C  NOPT1,NOPT2,NOPT3,AND NOPT4.
C
C
C  IF N1 IS 3 OR GREATER THE PROGRAM WILL HALT.
C
C
C
1003  DO 1001 I=1,400
      T(I) = 0
      DO 1001 J=1,114
1001  FX(I,J) = 0
      DO 1002 I=1,112
      JC(I) = 0
      W1(I) = 0
1002  W(I) = 0
      W(113) = 0
      W1(113) = 0
C
C
C
C
C
      ND5 = 5
      ND6 = 6
      READ(ND5,9999) N1,N2,N5
9999  FORMAT(I2,2X,I2,2X,I2)
      WRITE(ND6,9999) N1
9999  FORMAT(1H1,'THE VALUE OF N1 IS',1X,I4)
      WRITE(ND6,9998) N2
9998  FORMAT(1H ,'THE VALUE OF N2 IS',1X,I4)
      WRITE(ND6,9997)N5
9997  FORMAT(1H ,'IF N5 >0,THE VECTOR WILL BE SAVED.  N5 =',I5)
      WRITE(ND6,9996)
      WRITE(ND6,9996)
      WRITE(ND6,9996)
      WRITE(ND6,9996)
      WRITE(ND6,9996)
9996  FORMAT(1H )
      IF(N1 - 2) 1,100,1000
C
C
C
C
C
1     DO 3 I=1,N2
      READ(ND5,9998) N3,N4
9998  FORMAT(I4,2X,I4)
      WRITE(ND6,9995) N3,N4
```

```
 9985  FORMAT(1H ,'FILE ',I4,'WAS READ.  VECTOR',I5,'WAS USED.')
       IF(N4 .EQ. 1 ) GO TO 13
       DO 2 J=1,N4
 2     READ(N3) (T(K),K=1,113)
       GO TO 14
 13    READ(N3) (T(K),K=1,113)
 14    REWIND N3
       DO 3 K=1,113
 3     W(K) = W(K) + T(K)
       FN2 = N2
       DO 4 K=1,113
 4     W(K) = W(K) / FN2
       GO TO 200
 C
 C
 C
 C
 C
 C
 100   READ(ND5,9996) NOPT1,NOPT2,NOPT3,NOPT4
 9996  FORMAT(I4,2X,I4,2X,I4,2X,I4)
       WRITE(ND6,9984) NOPT1
 9984  FORMAT(1H ,'OPTION 1 =',I5)
       WRITE(ND6,9983) NOPT2
 9983  FORMAT(1H ,'OPTION 2 =',I5)
       WRITE(ND6,9982) NOPT3
 9982  FORMAT(1H ,'OPTION 3 =',I5)
       WRITE(ND6,9981) NOPT4
 9981  FORMAT(1H ,'OPTION 4 =',I5)
       NCARD = 1117
       CALL RE71(NOPT1,NCARD,NCONT,NOPT2,NOPT3,NOPT4,N30)
       REWIND 71
       K = 0
       K1 = 0
       DO 101 I=1,NCONT
       IF(T(I) .EQ. 0) GO TO 103
       IF(T(I) .GT. 200) GO TO 101
       K = K+1
       DO 104 J=1,113
 104   W(J) = W(J) + FX(I,J)
       GO TO 101
 103   K1 = K1 + 1
       DO 501 J=1,113
 501   W1(J) = W1(J) + FX(I,J)
 101   CONTINUE
       FK = K
       FK1 = K1
       DO 102 I=1,113
       W(I) = W(I)/FK
 102   W1(I) = W1(I) / FK1
       DO 500 I=1,113
 500   W(I) = W(I) - W1(I)
 C
 C
 C
 C
 C
 200   IF(N5 .LT. 1) GO TO 201
       WRITE(75) (W(K),K=1,113)
 201   WRITE(ND6,9997)( W(K),K=1,113)
 9997  FORMAT(1H ,10F9.5,/,11(1X,10F9.5,/))
       CALL PLOT
       GO TO 1003
 1000  CONTINUE
       STOP
       END
```

# APPENDIX C

# SIMPLE AVERAGE REPRESENTATIONS

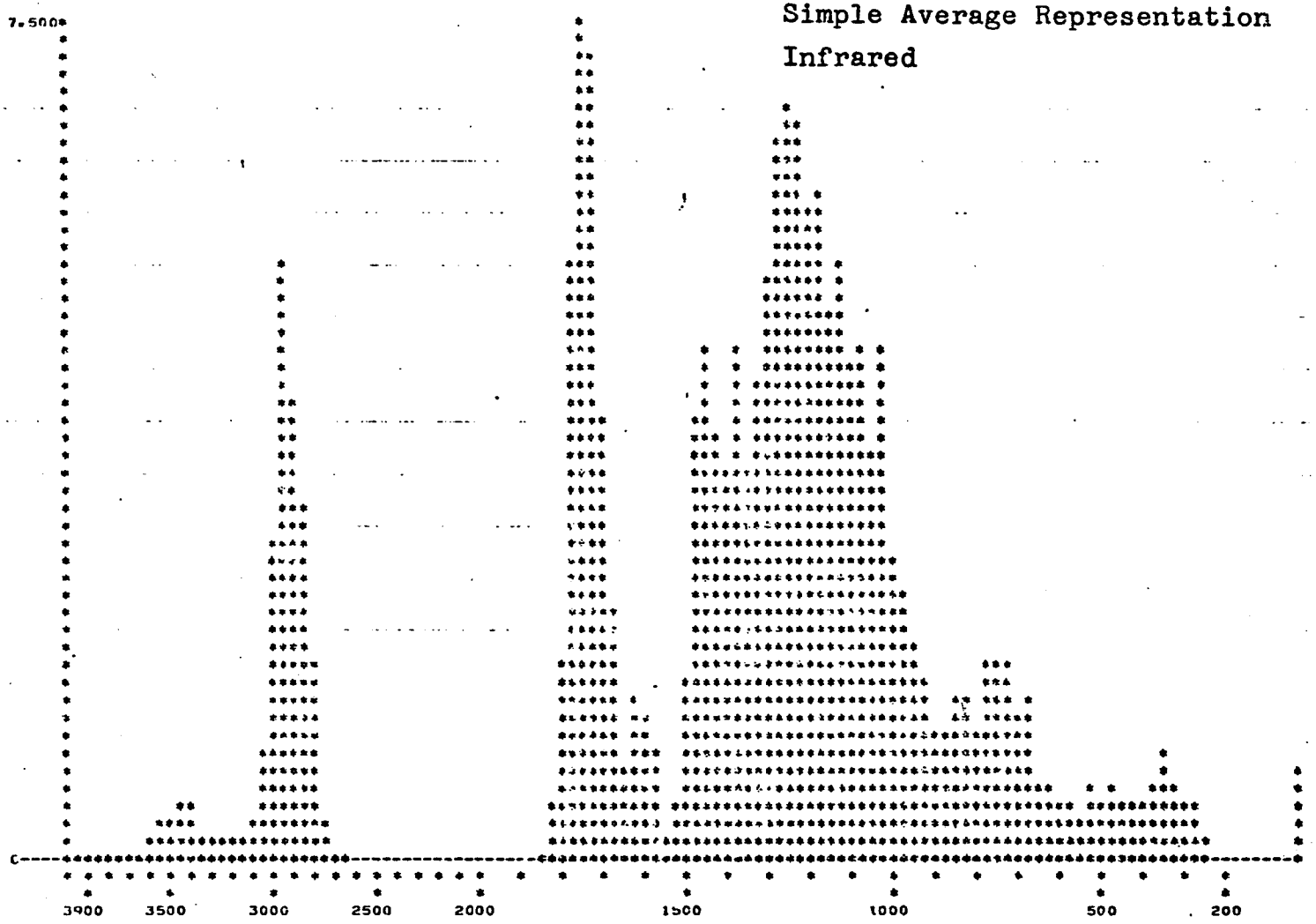
# and

# MUTED AVERAGE REPRESENTATIONS

ESTERS

Simple Average Representation

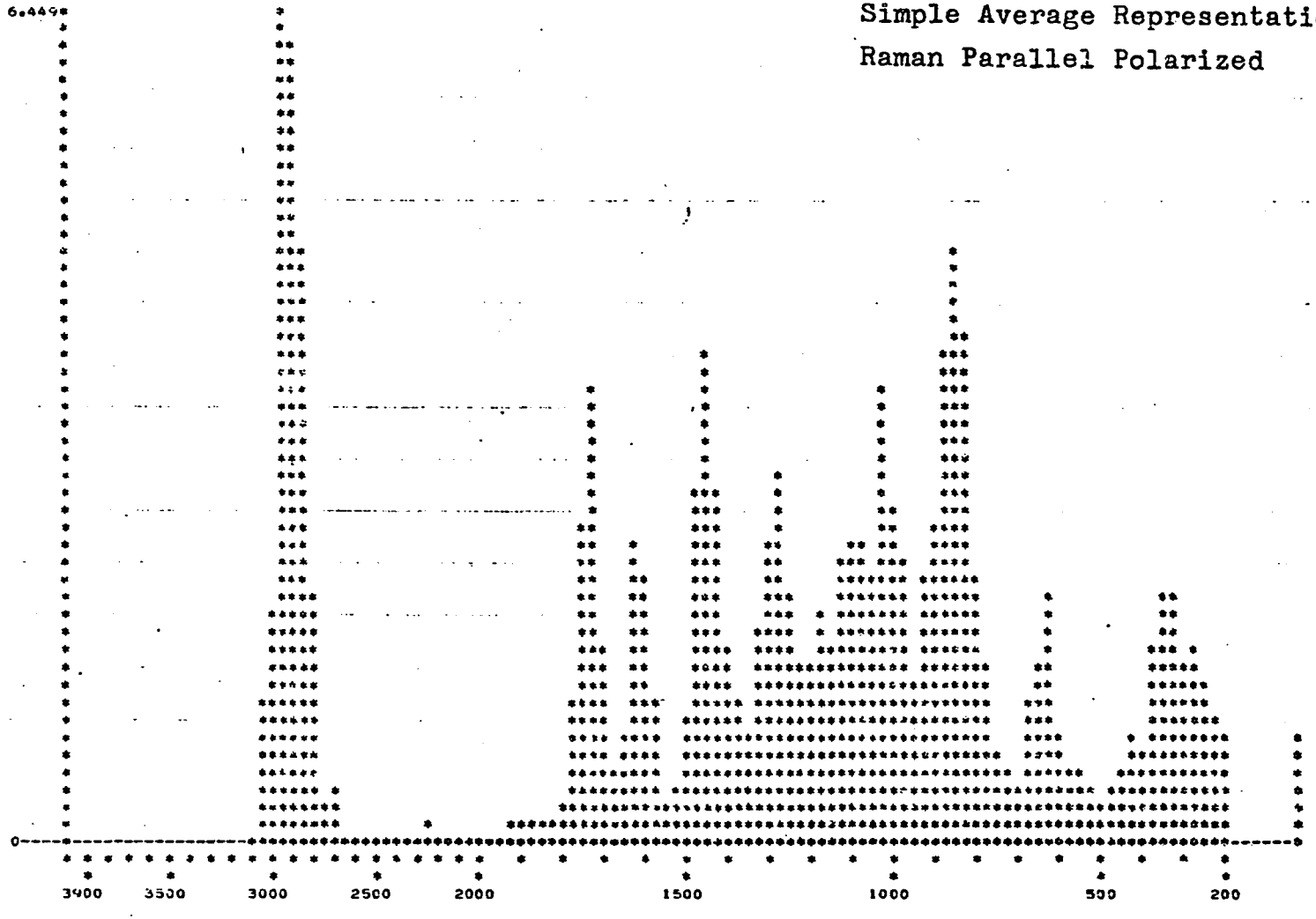
Infrared



ESTERS

Simple Average Representation

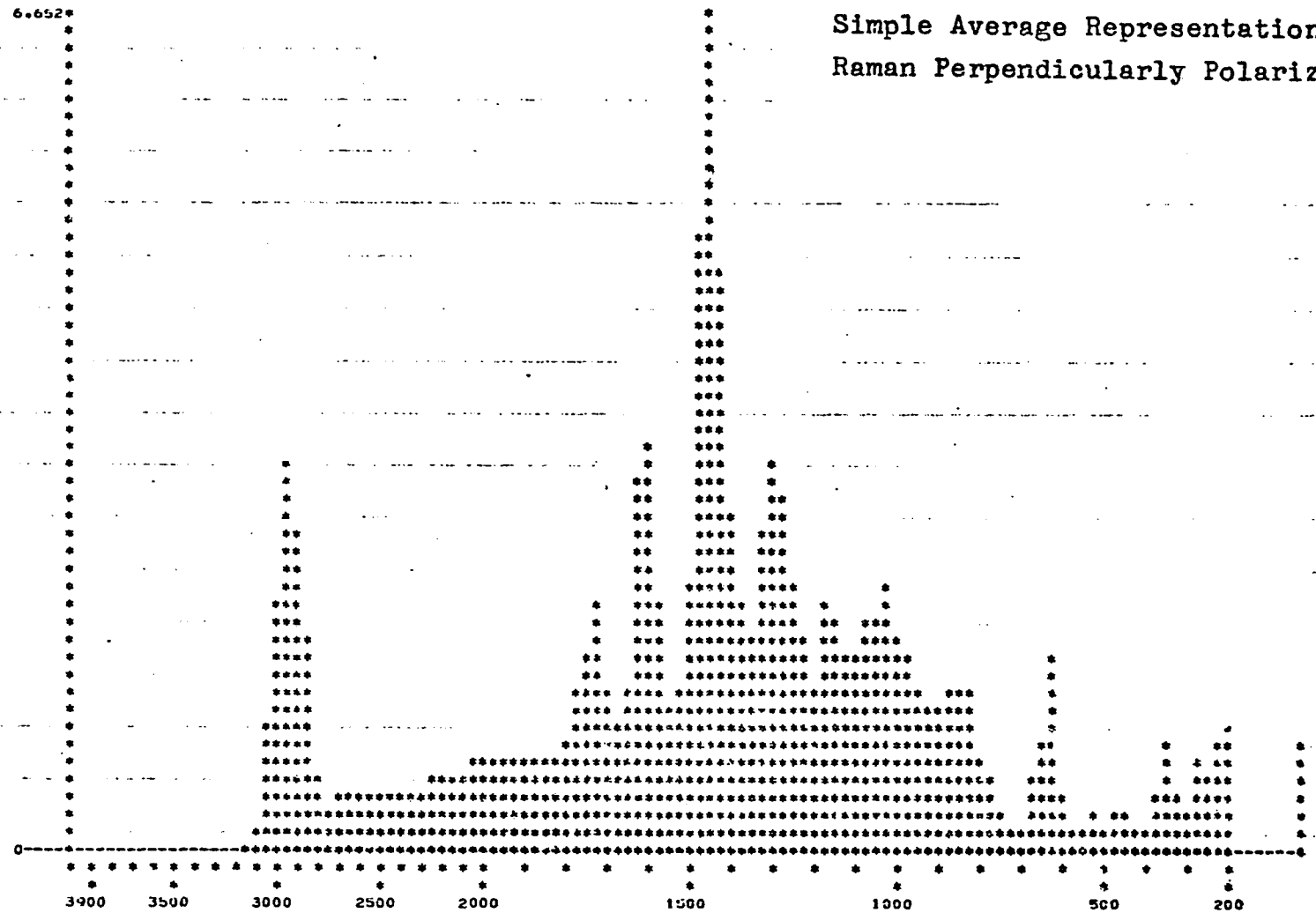
Raman Parallel Polarized

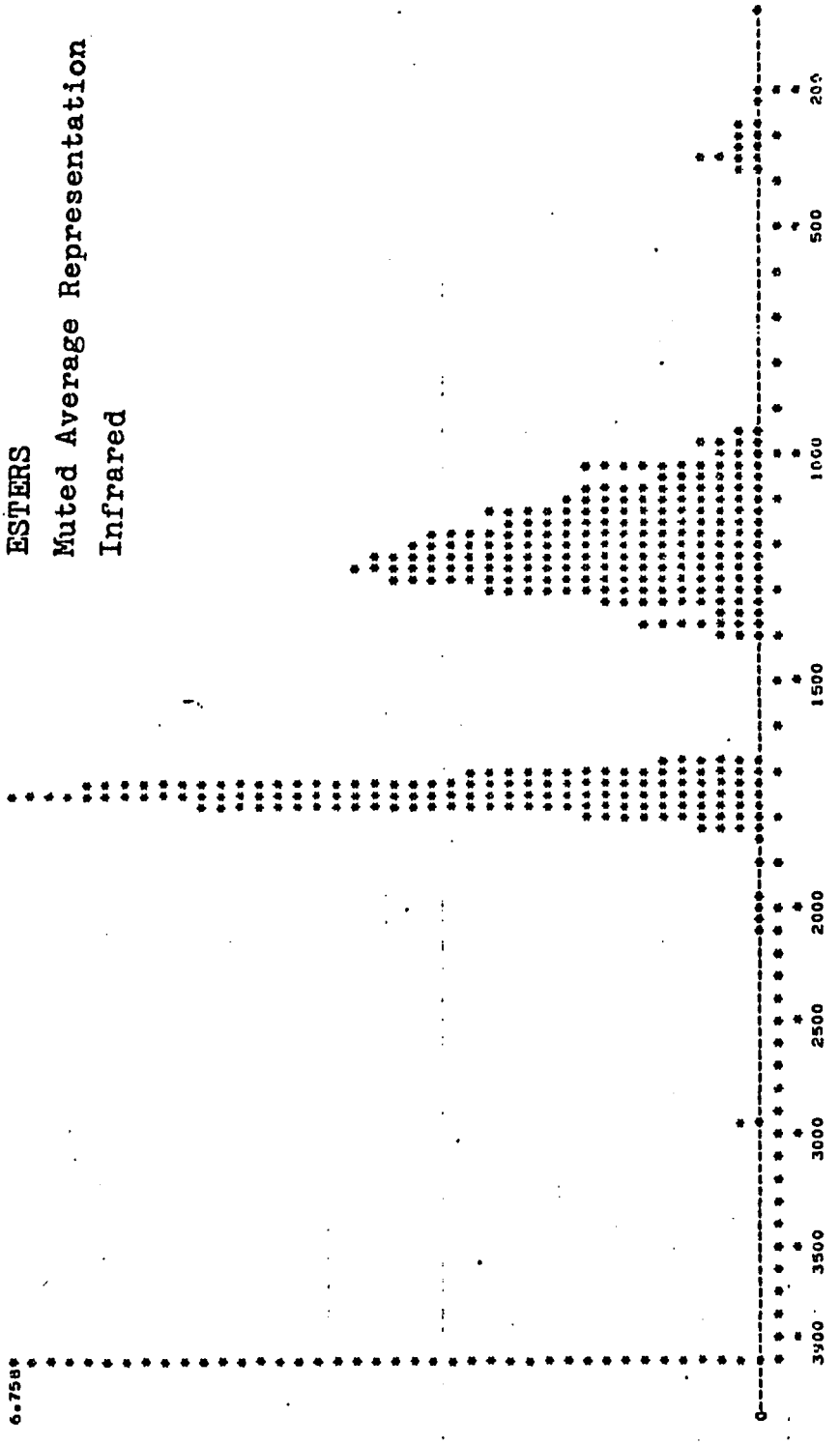


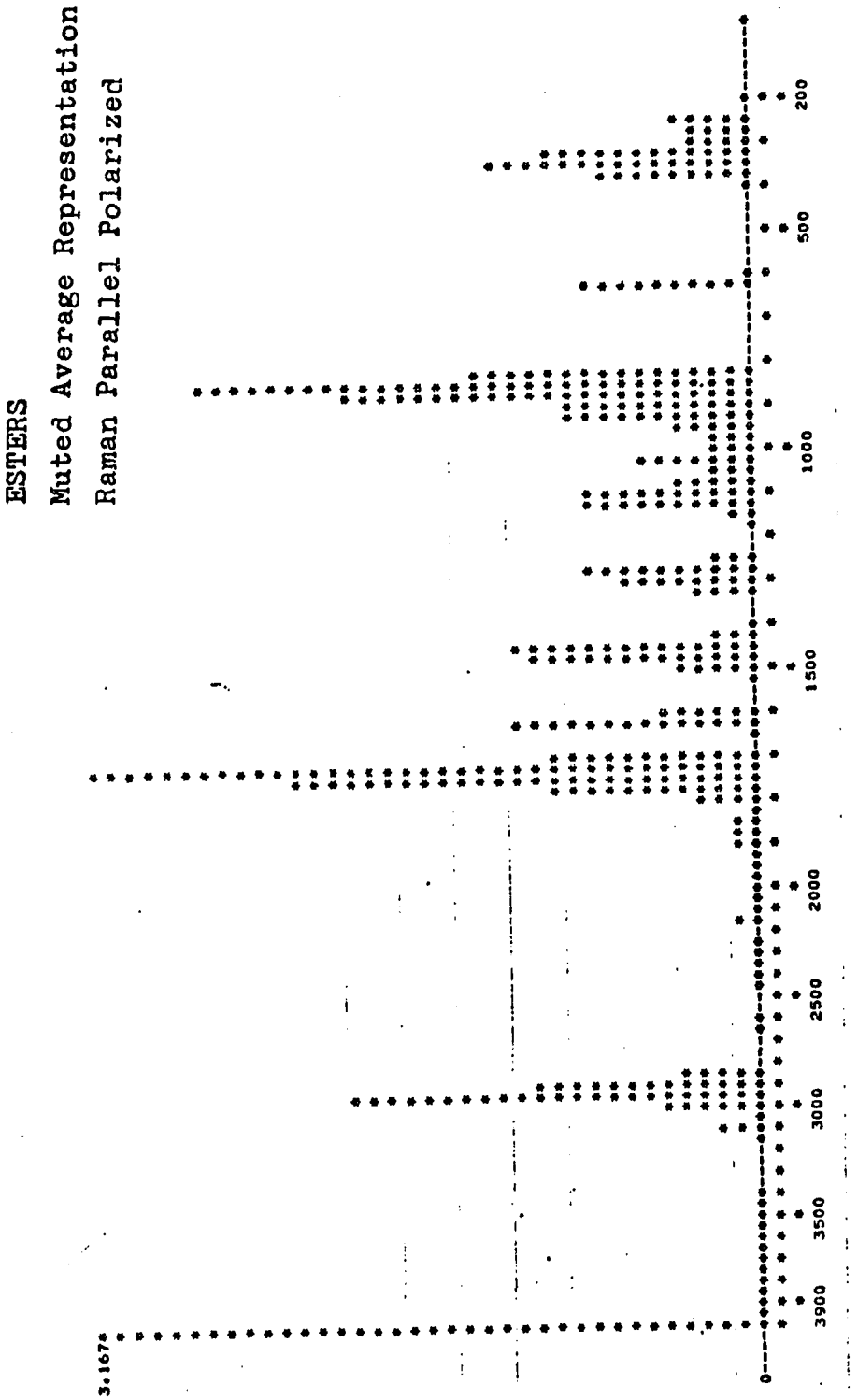
ESTERS

Simple Average Representation

Raman Perpendicularly Polarized



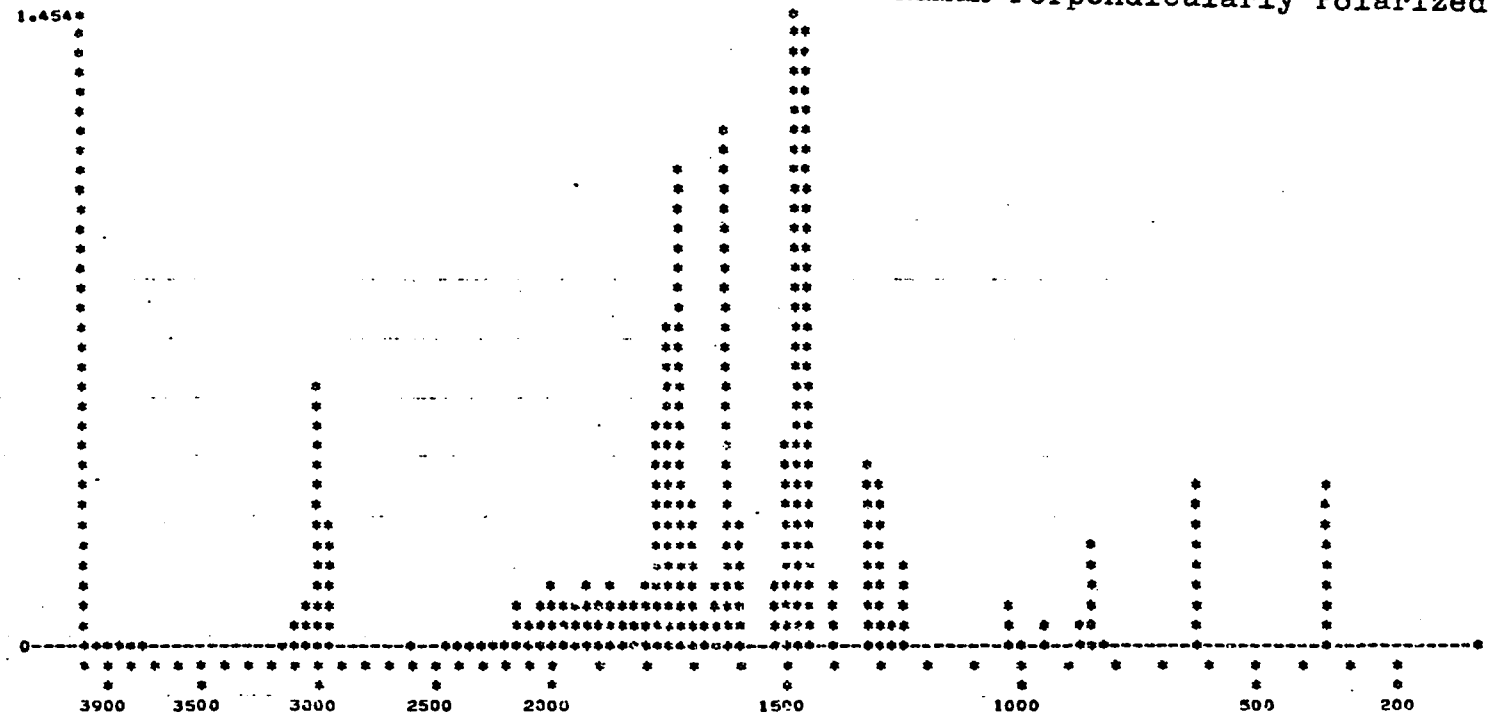

ESTERS
Muted Average Representation
Infrared
6.758
0
500
1000
1500
2000
2500
3000
3500
4000

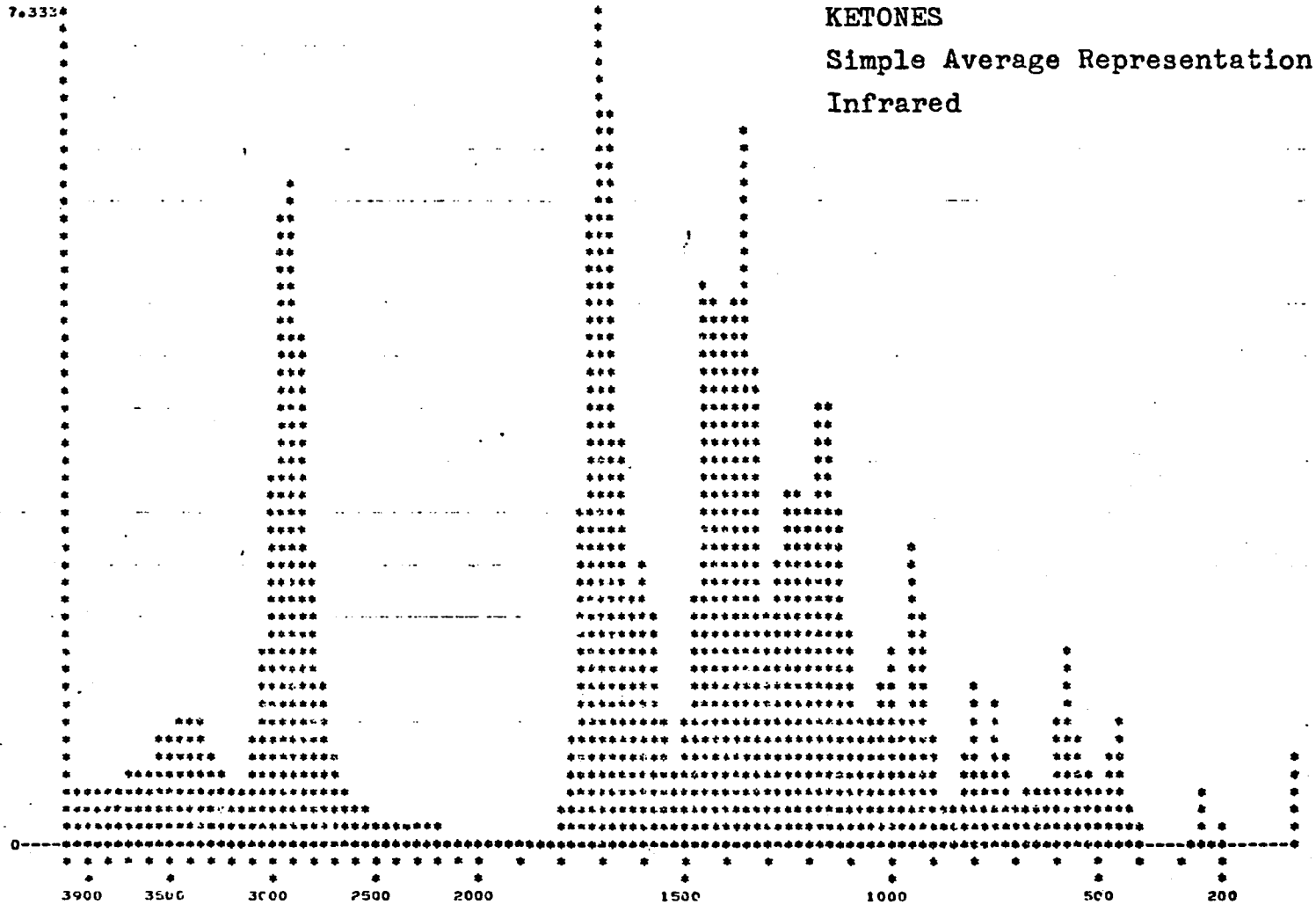

ESTERS
Muted Average Representation
Raman Parallel Polarized
3.167
0
3900
3500
3000
2500
2000
1500
1000
500
200

ESTERS

Muted Average Representation

Raman Perpendicularly Polarized

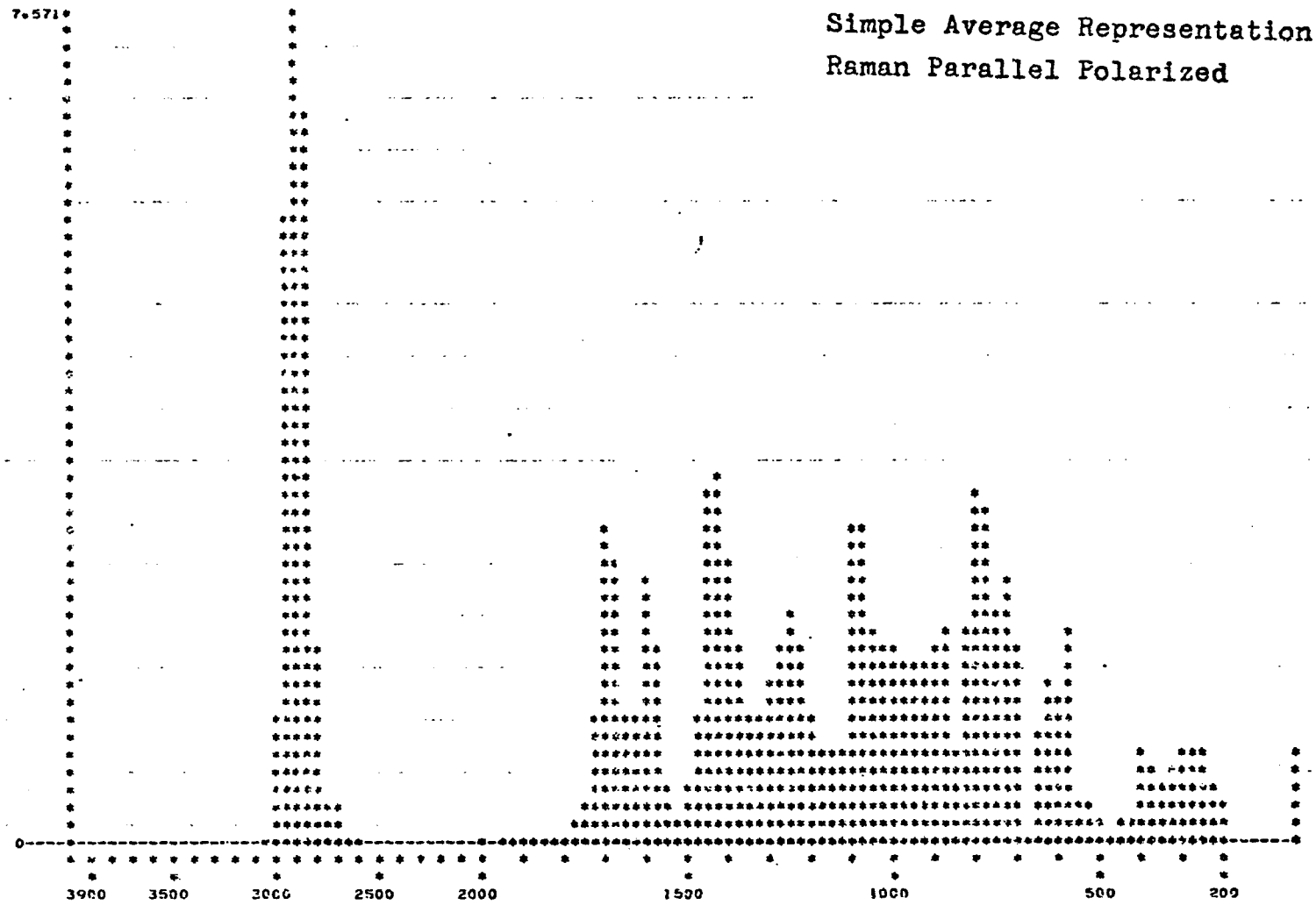



KETONES
Simple Average Representation
Infrared
7.3334
0
3900
3500
3000
2500
2000
1500
1000
500
200

KETONES

Simple Average Representation

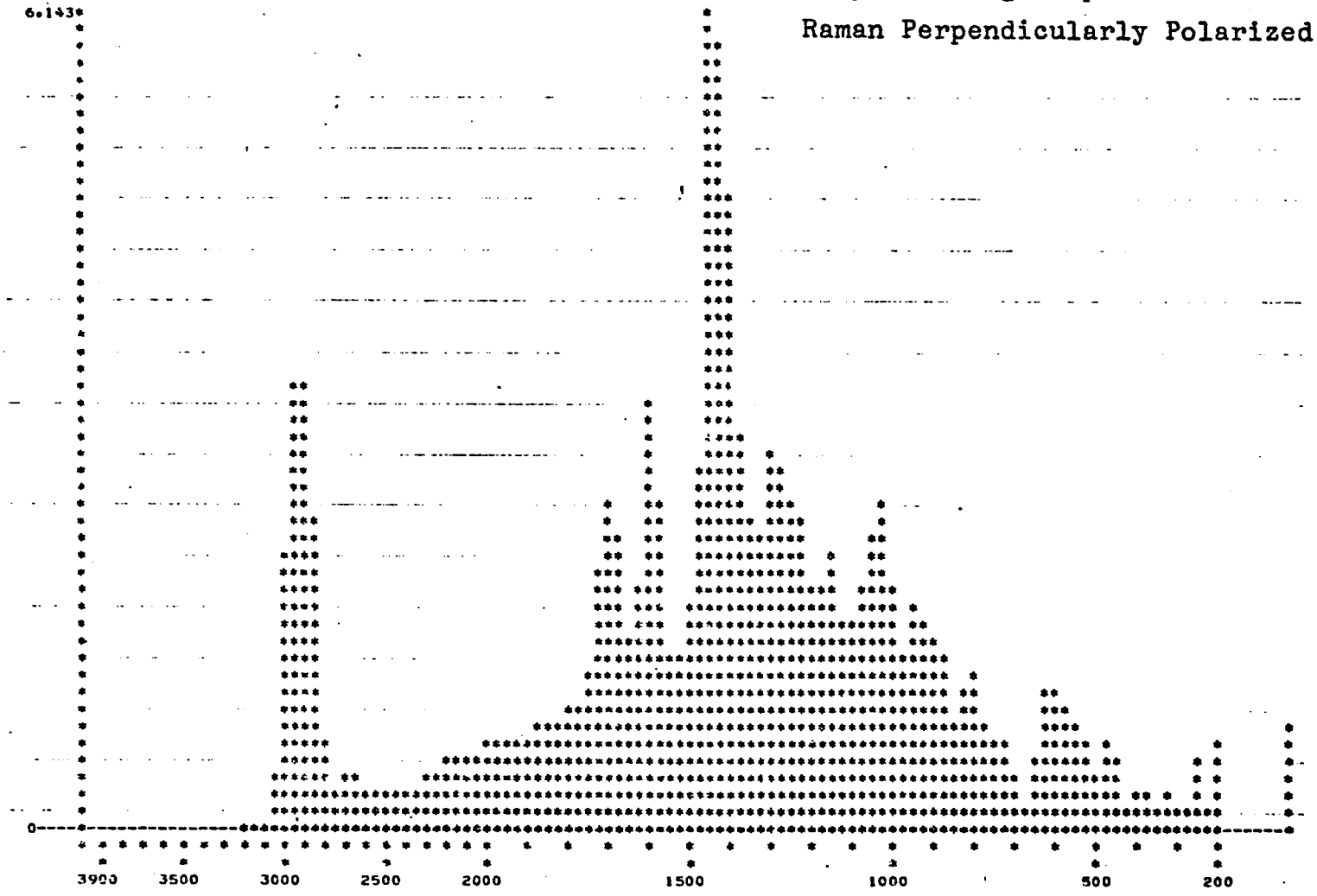
Raman Parallel Polarized



KETONES

Simple Average Representation

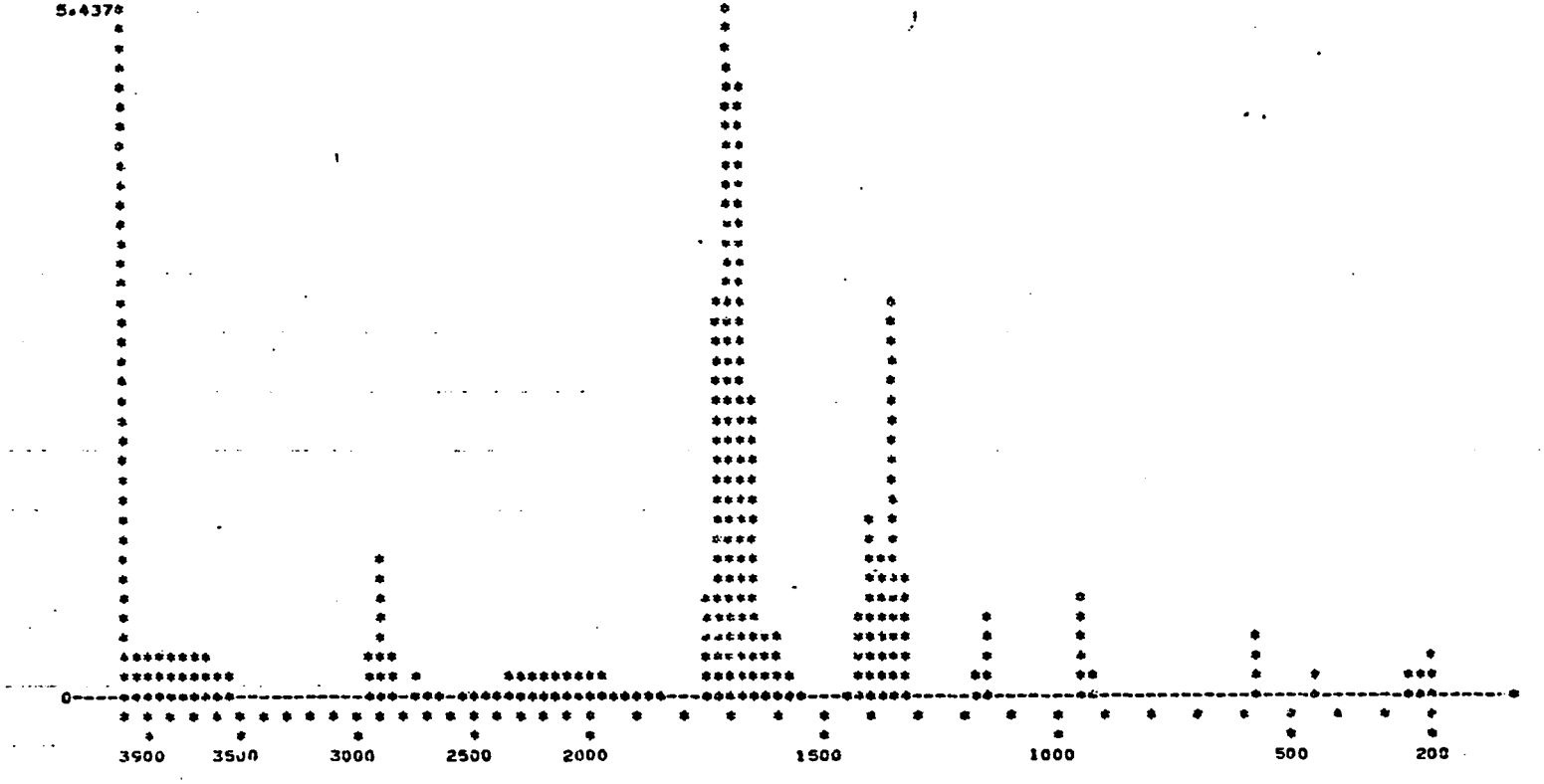
Raman Perpendicularly Polarized



KETONES

Muted Average Representation

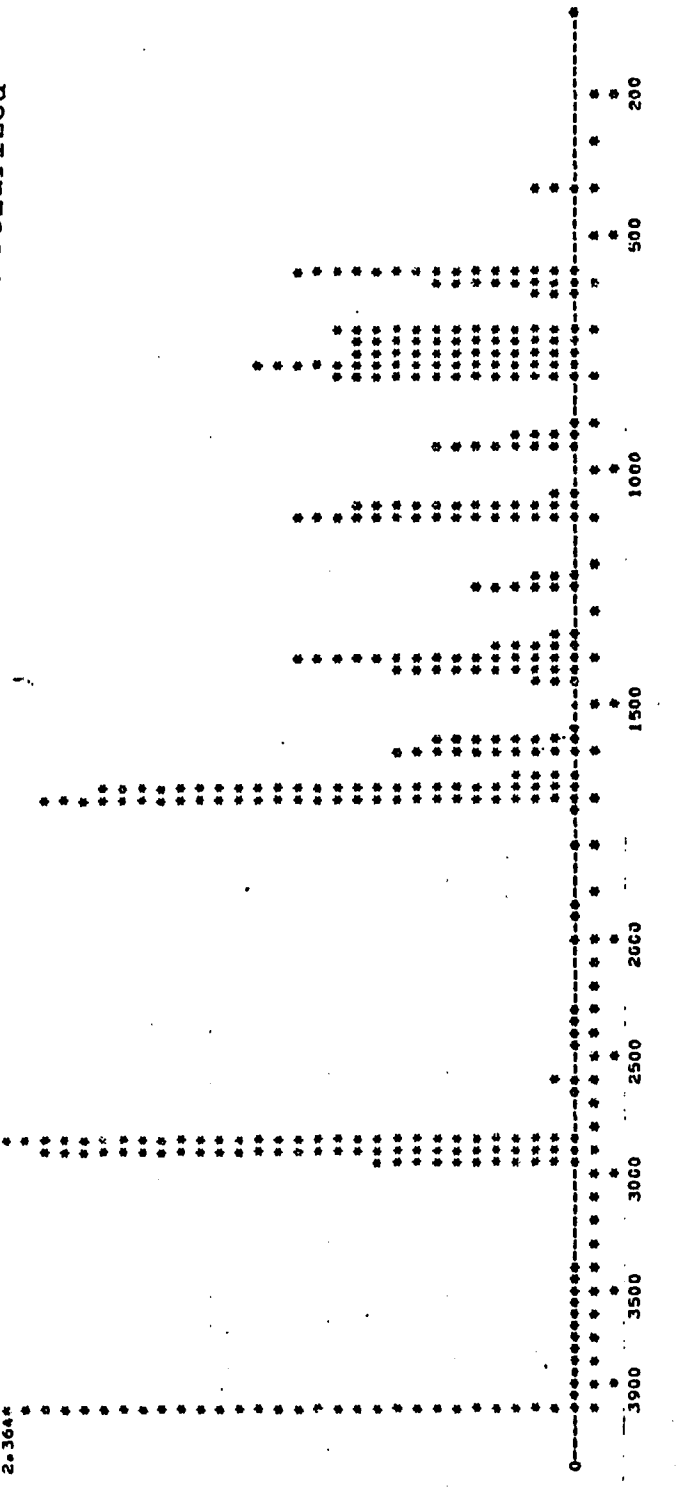
Infrared



KETONES

Muted Average Representation

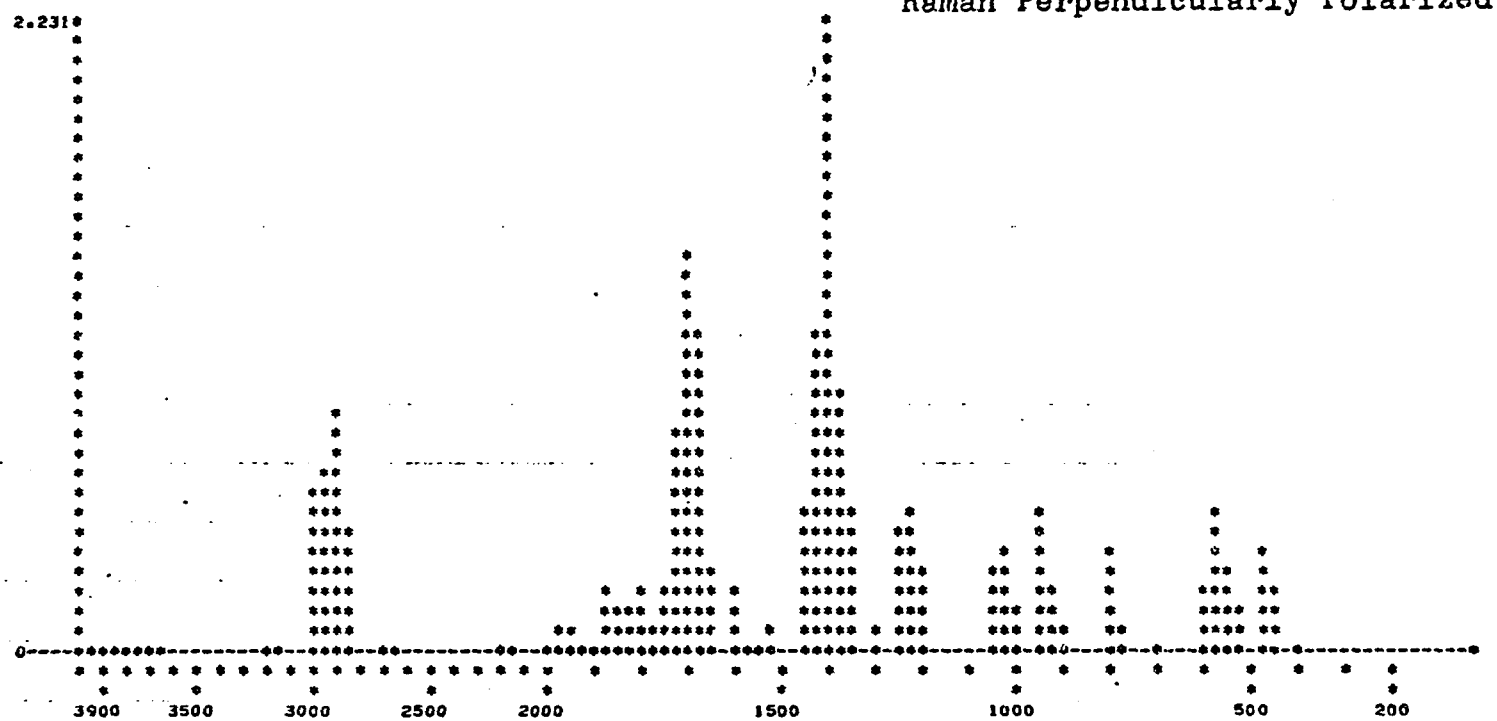
Raman Parallel Polarized



KETONES

Muted Average Representation

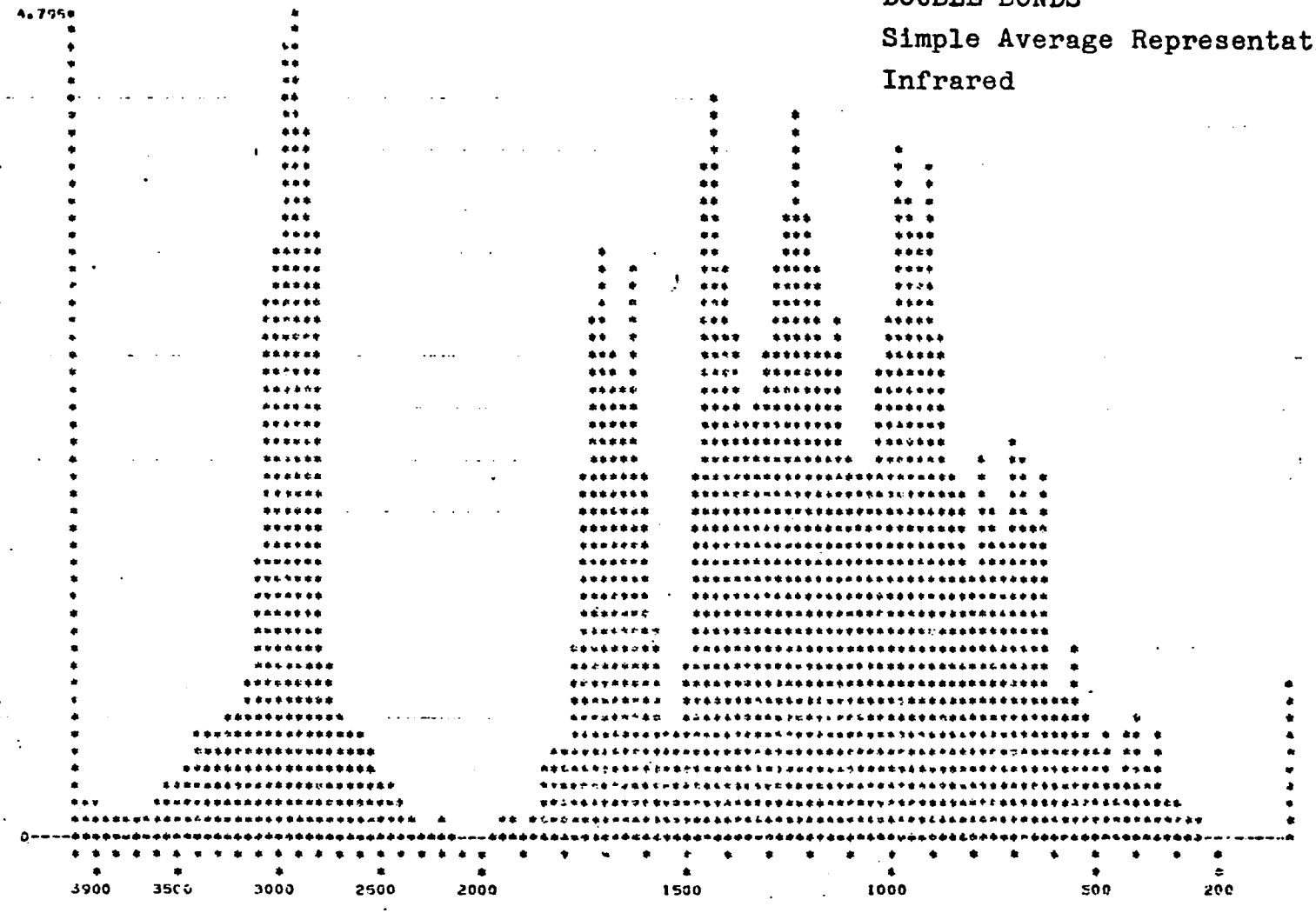
Raman Perpendicularly Polarized



DOUBLE BONDS

Simple Average Representation

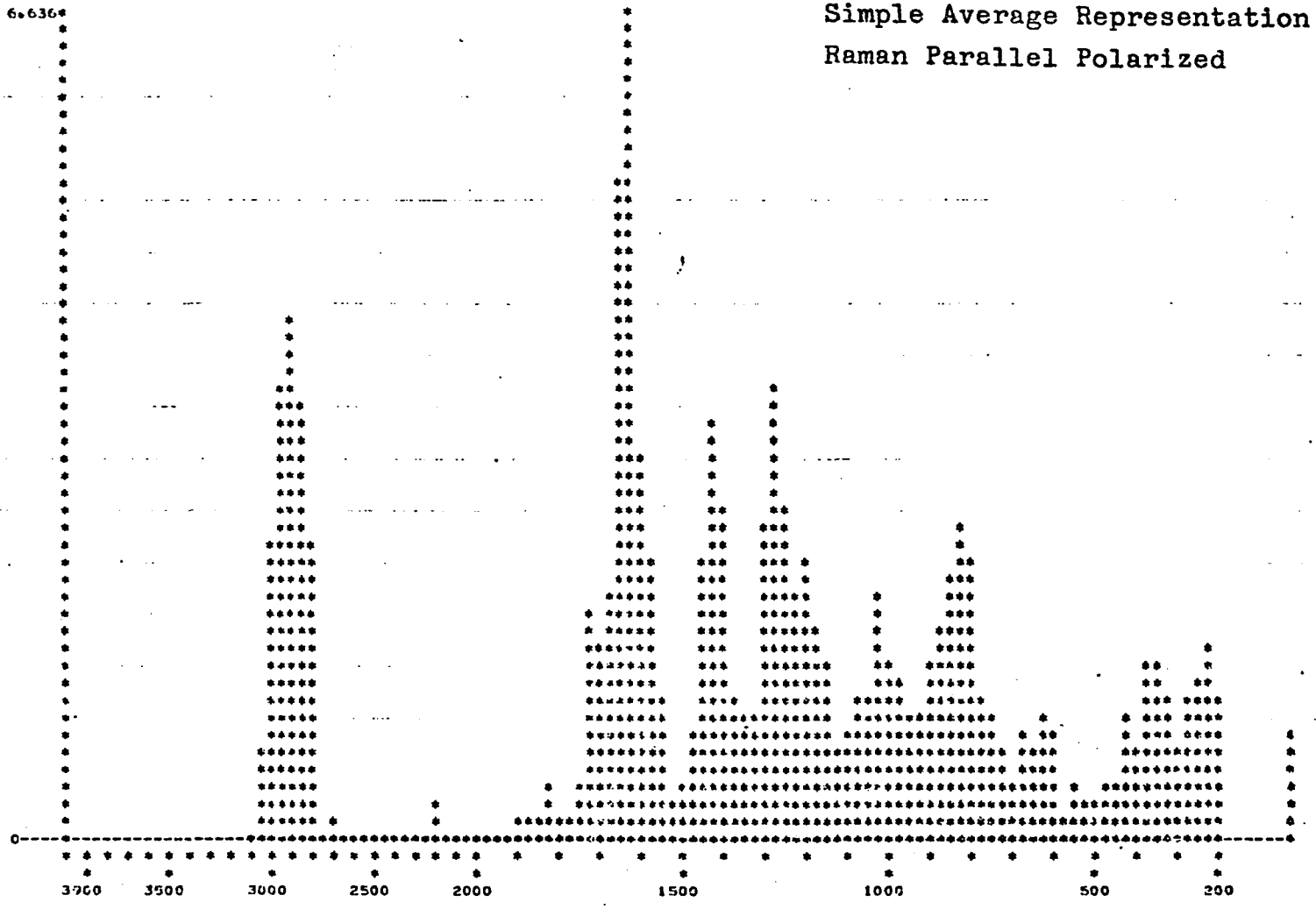
Infrared



DOUBLE BONDS

Simple Average Representation

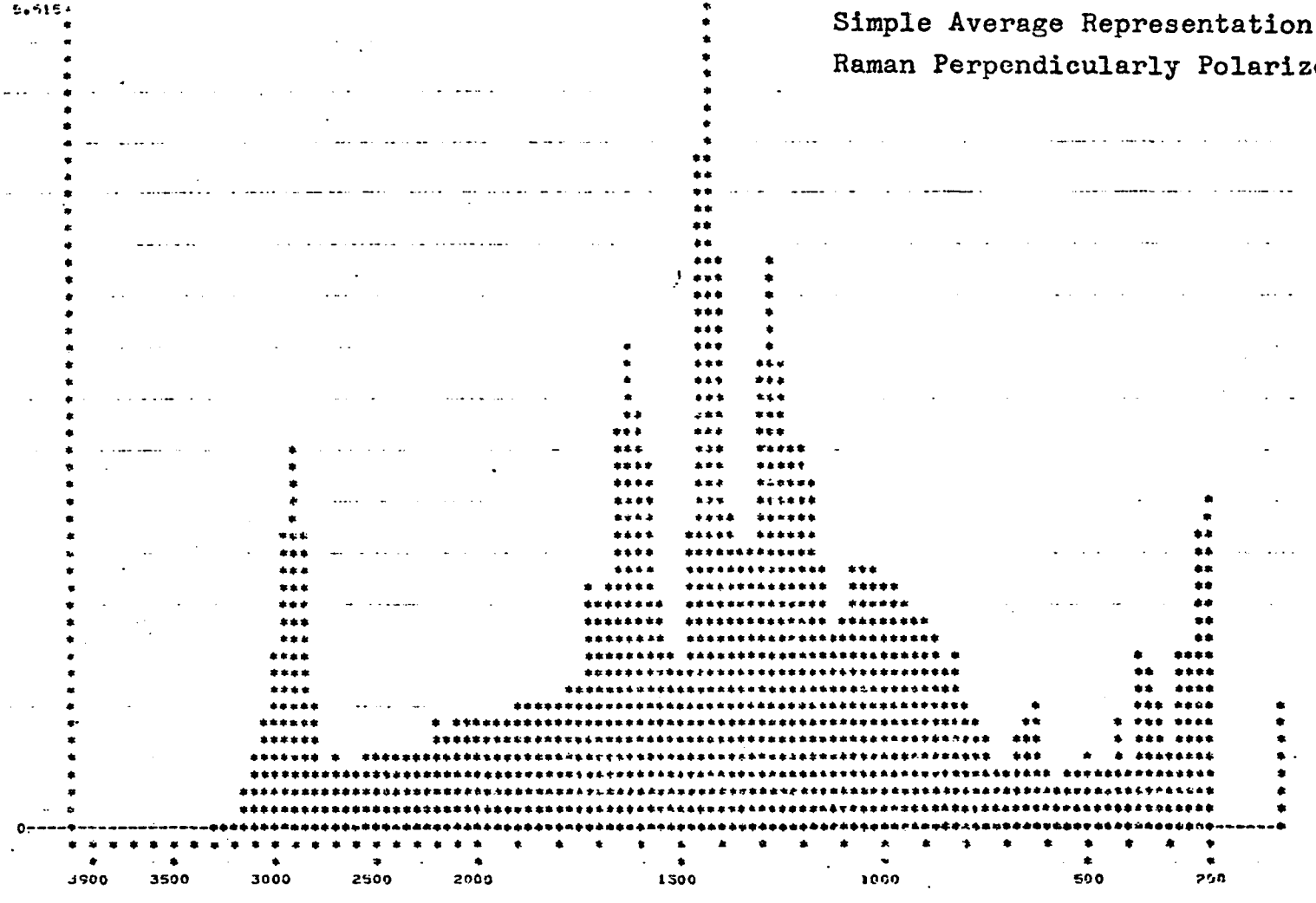
Raman Parallel Polarized



DOUBLE BONDS

Simple Average Representation

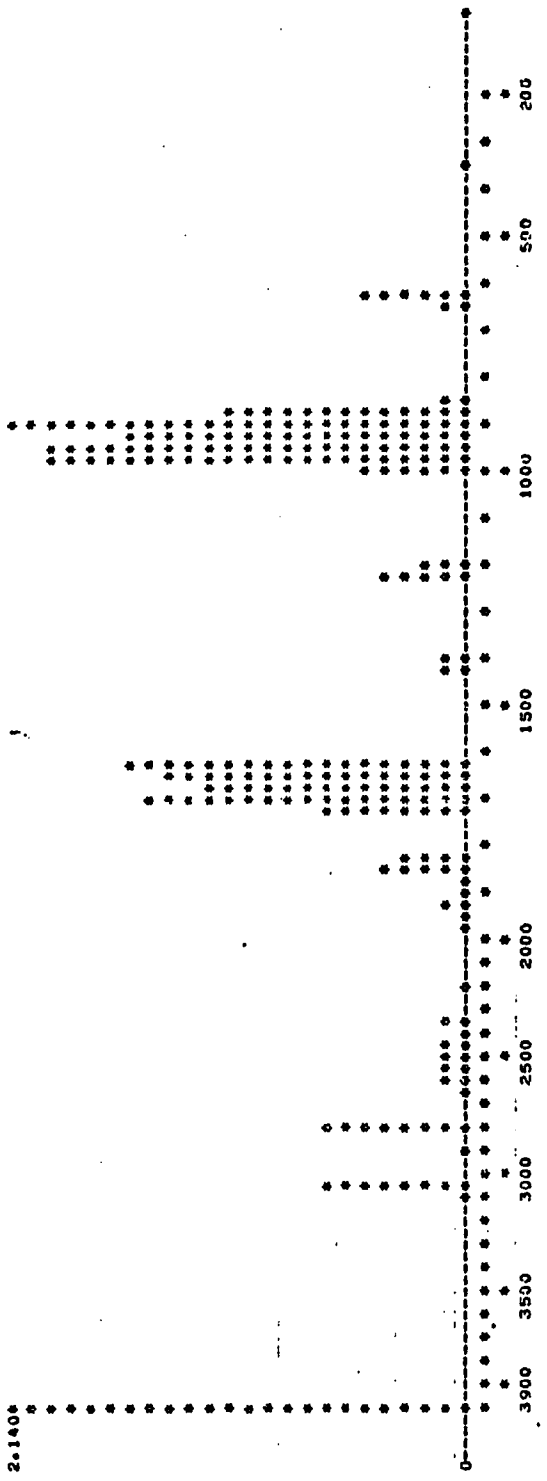
Raman Perpendicularly Polarized



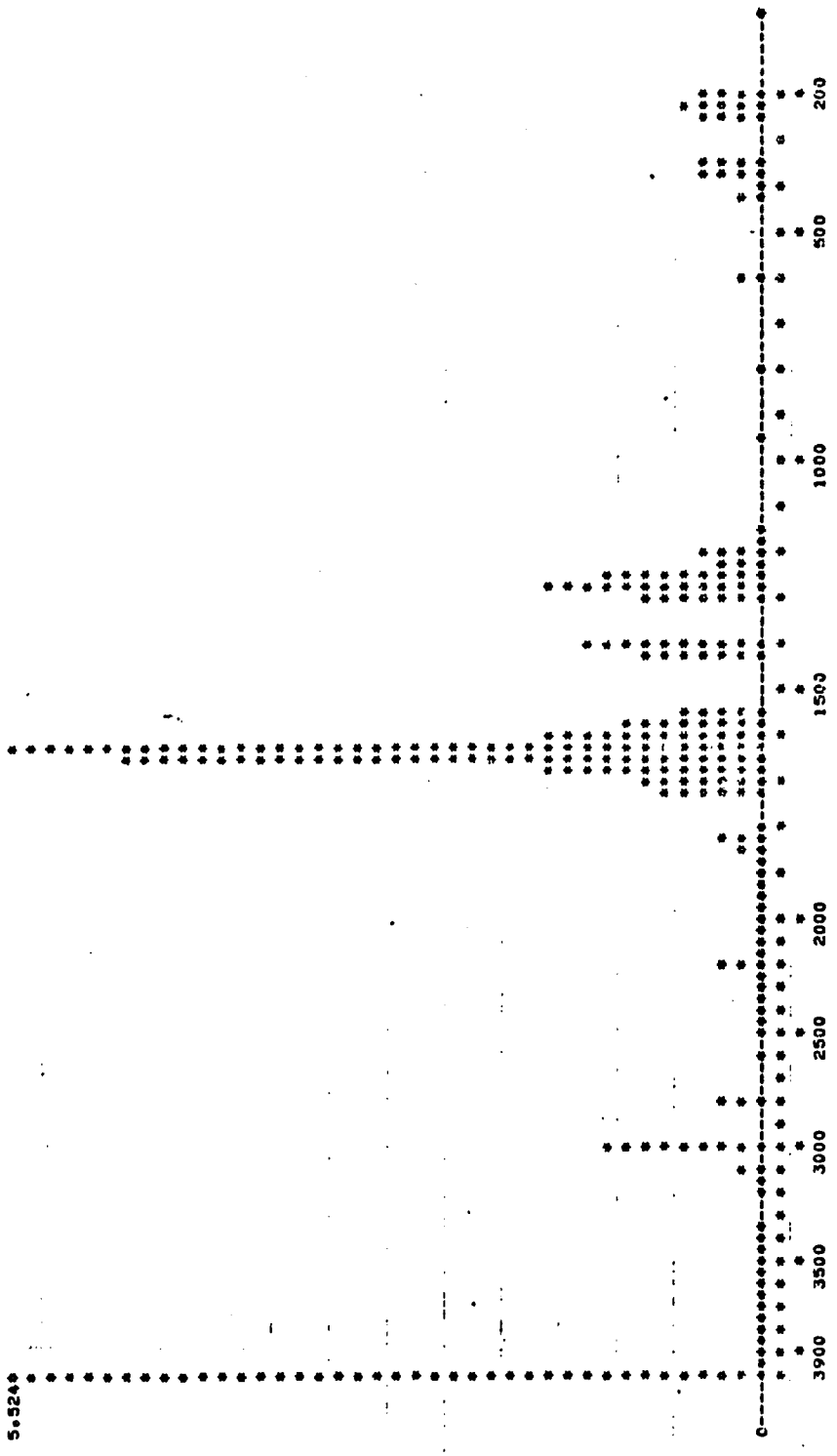
DOUBLE BONDS

Muted Average Representation

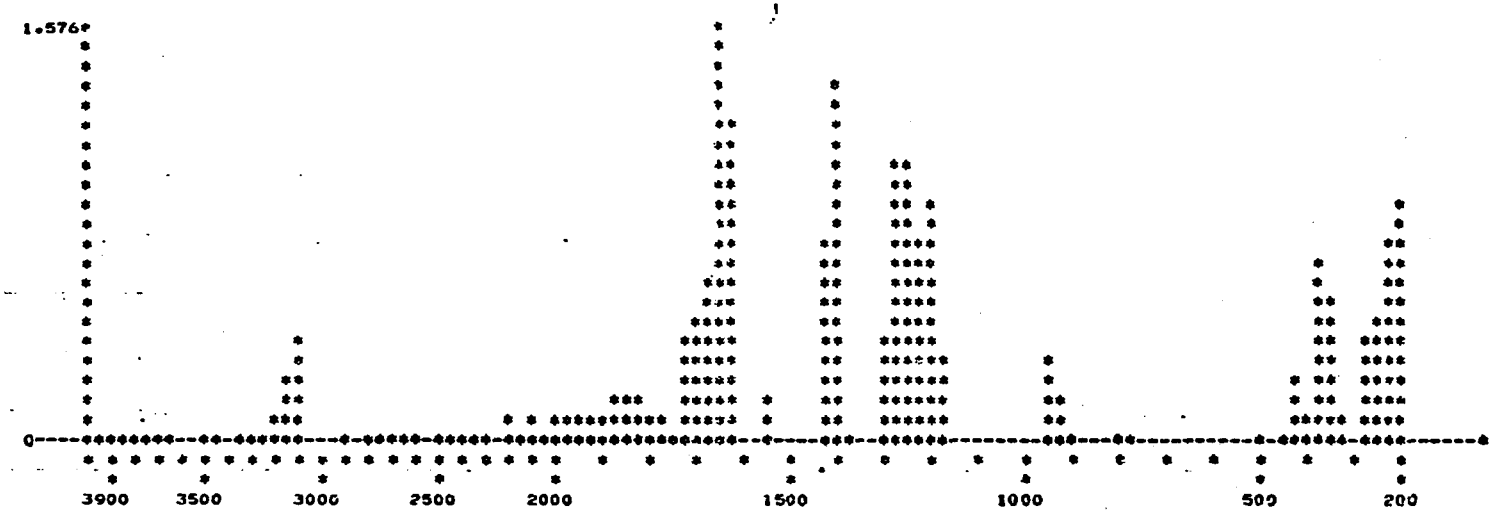
Infrared



DOUBLE BONDS
Muted Average Representation
Raman Parallel Polarized



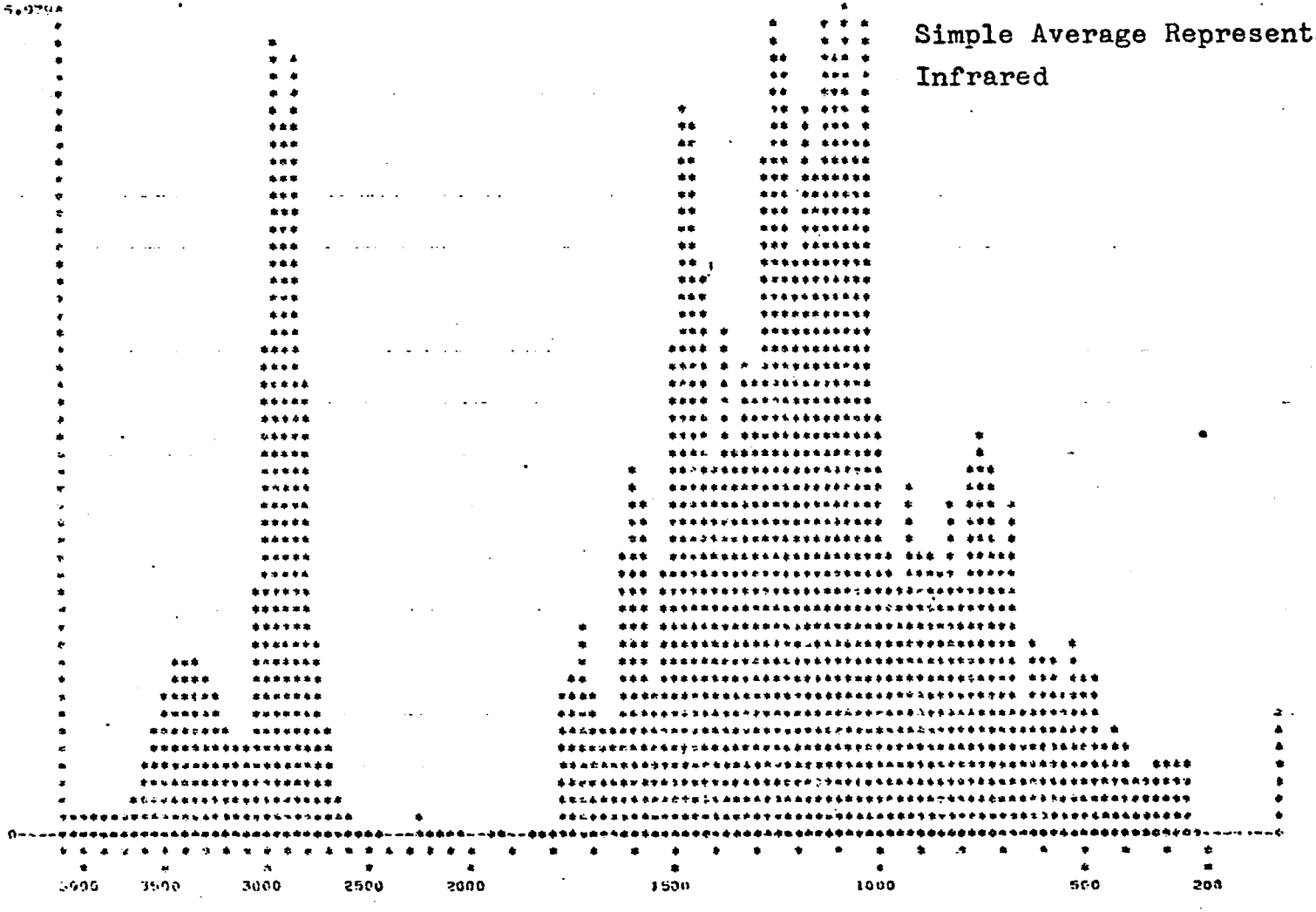
DOUBLE BONDS
Muted Average Representation
Raman Perpendicularly Polarized



ETHERS

Simple Average Representation

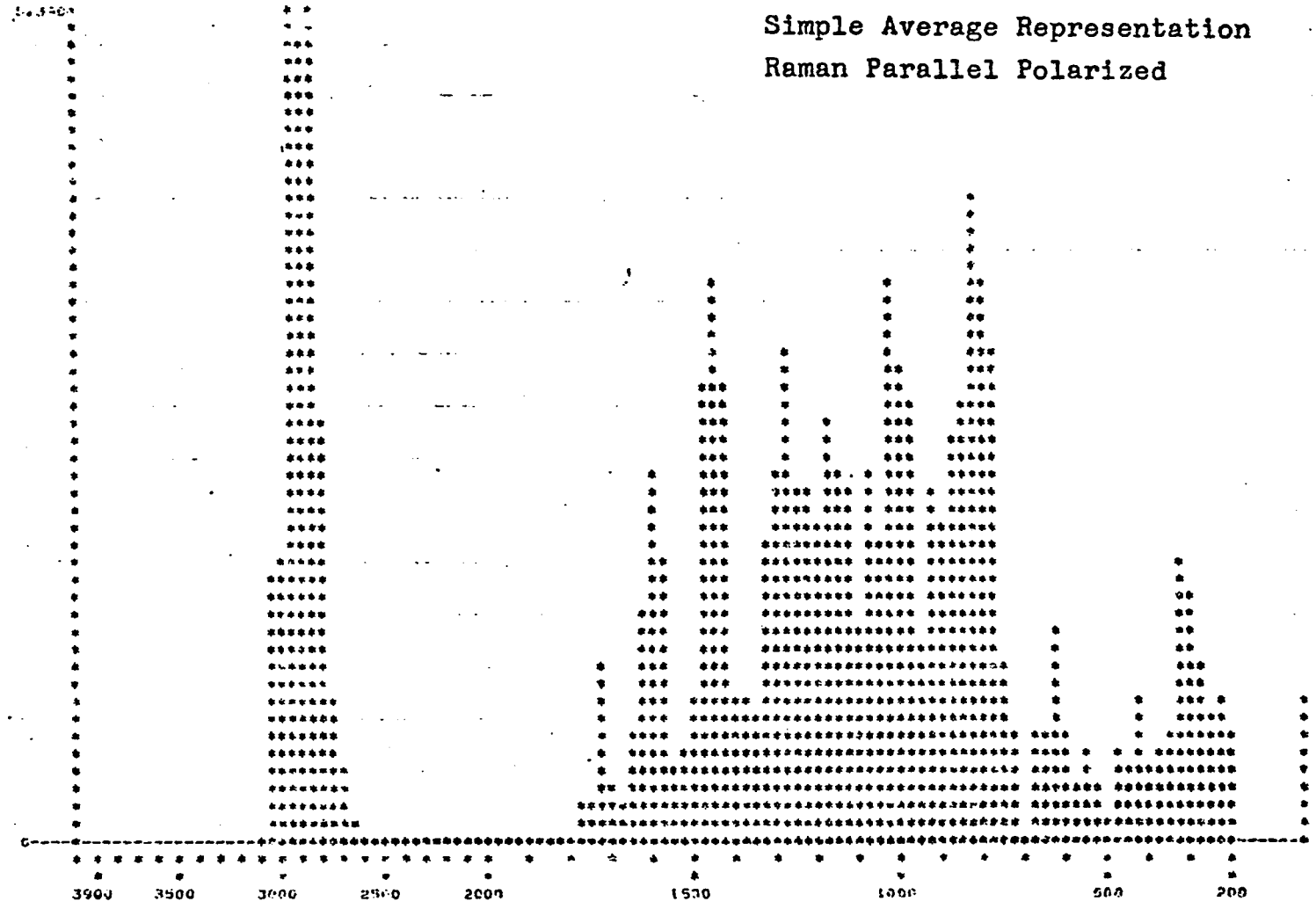
Infrared



ETHERS

Simple Average Representation

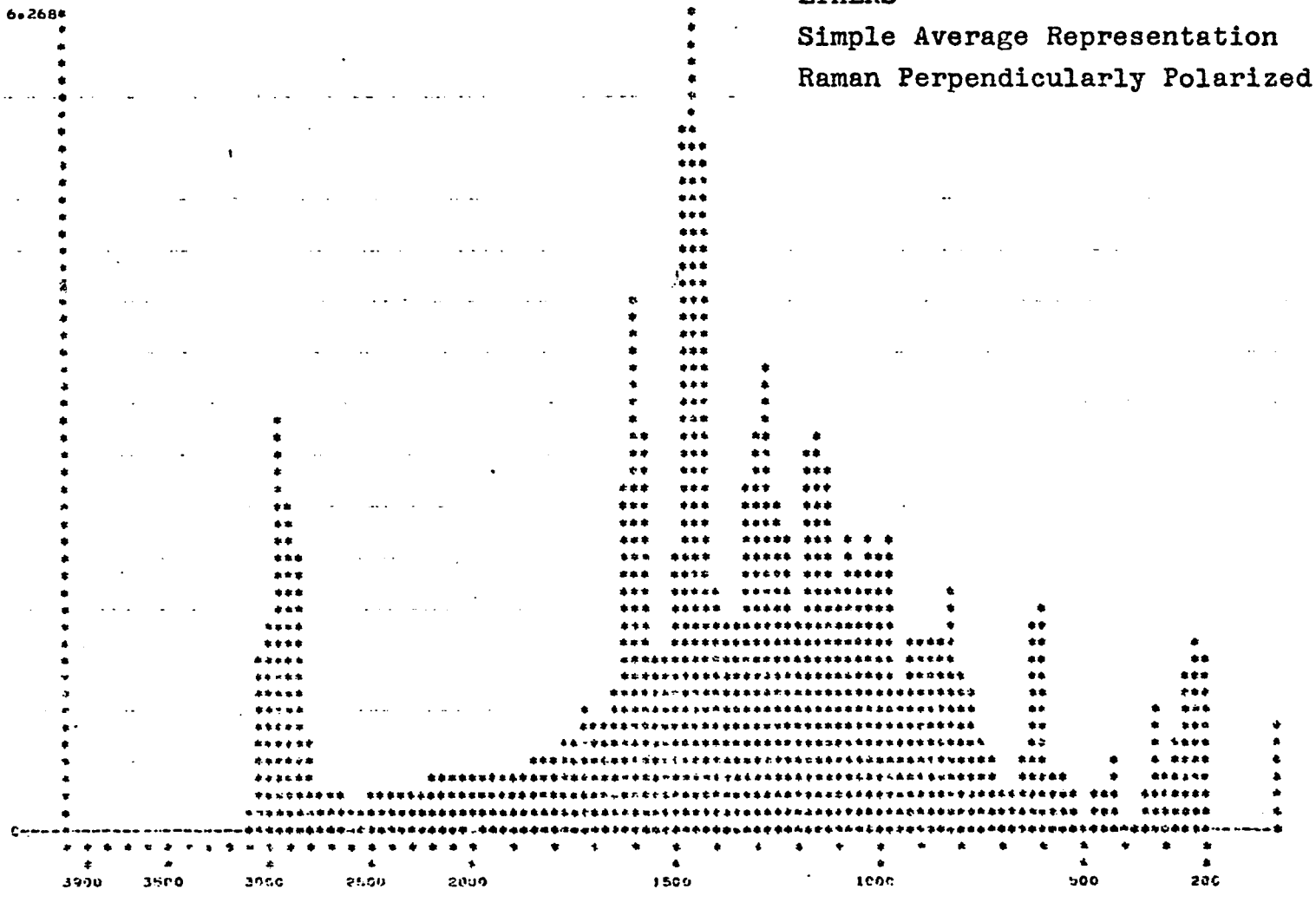
Raman Parallel Polarized



PLEASE NOTE: p.213

This page not included in material received from the Graduate School. Filmed as received.

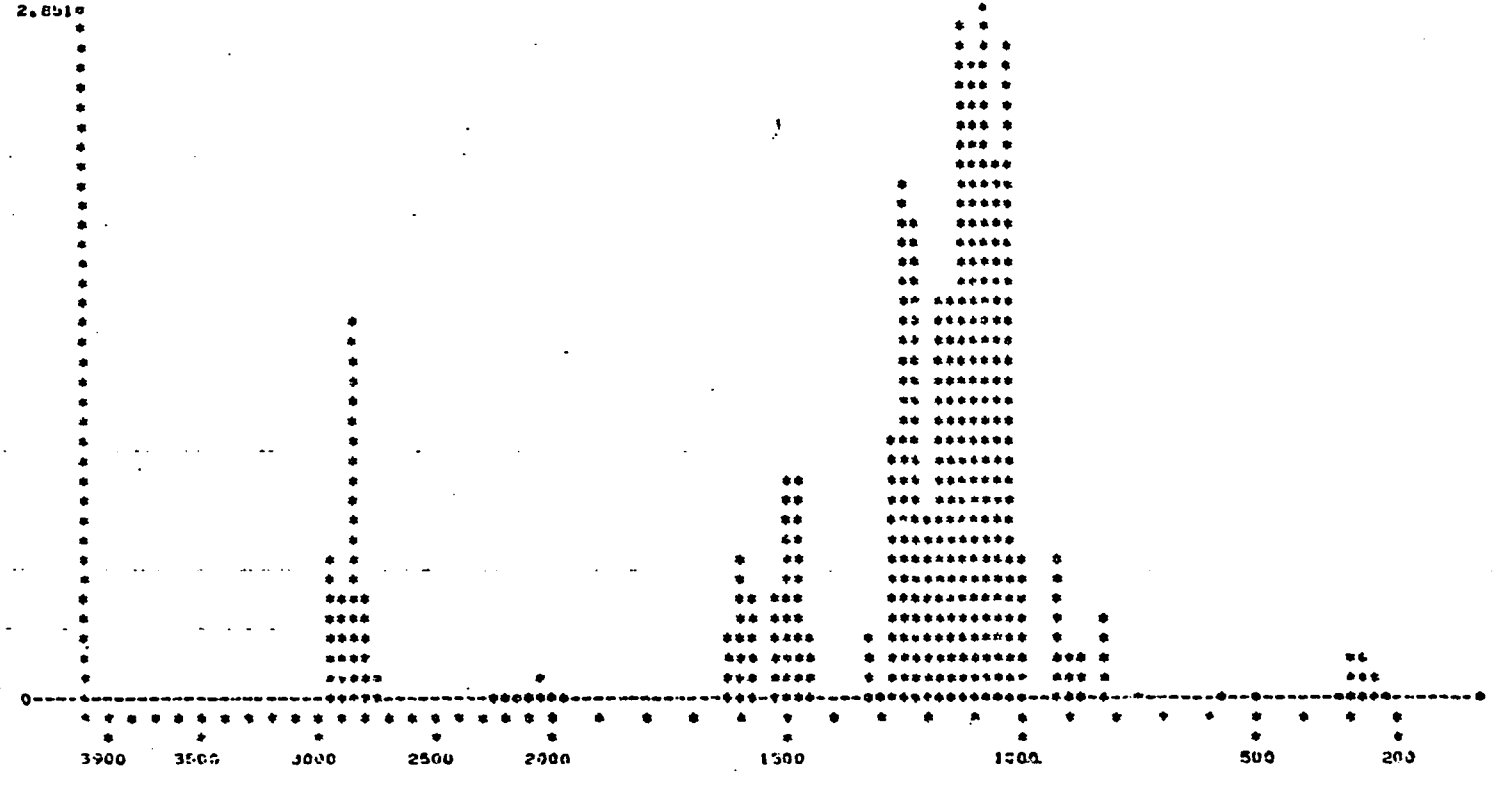
UNIVERSITY MICROFILMS


ETHERS
Simple Average Representation
Raman Perpendicularly Polarized
6.268
0
3000
3500
3000
2500
2000
1500
1000
500
200

ETHERS

Muted Average Representation

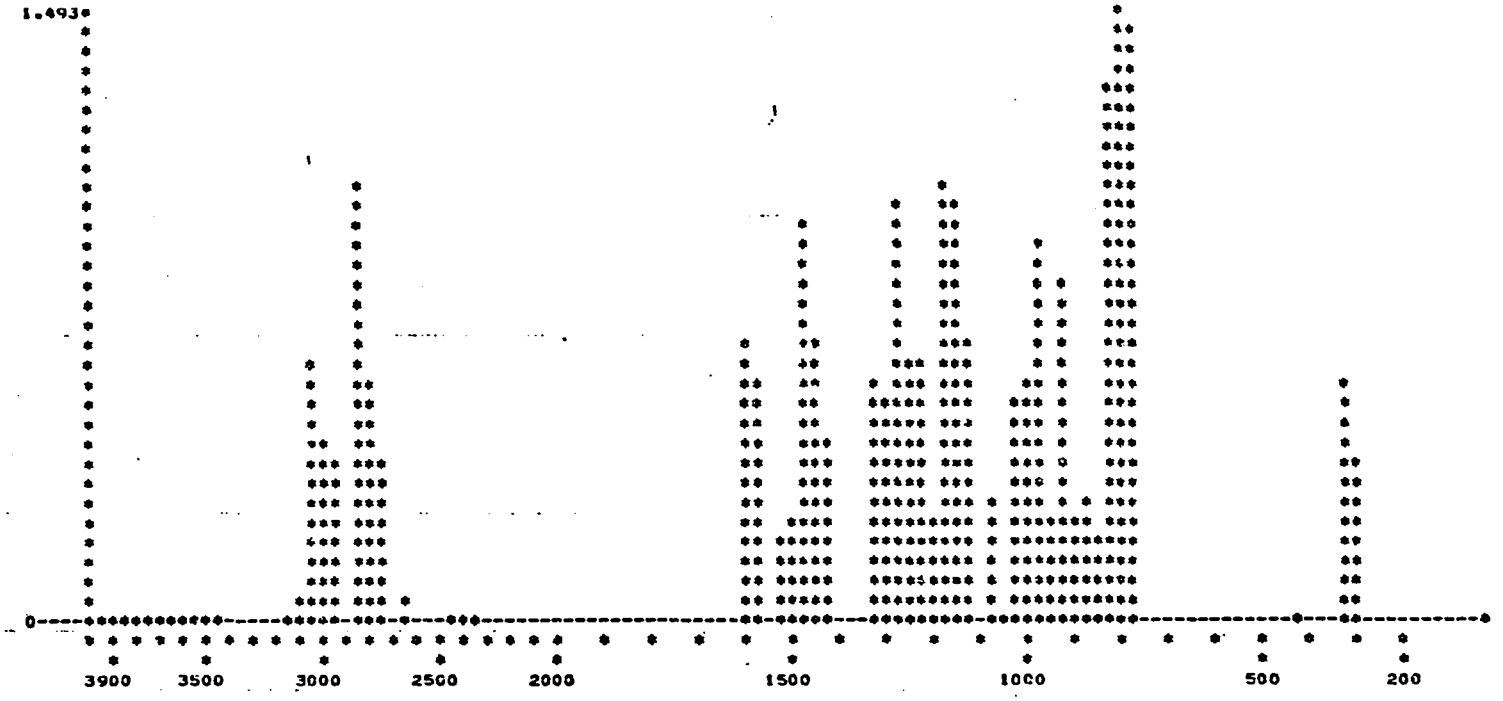
Infrared



ETHERS

Muted Average Representation

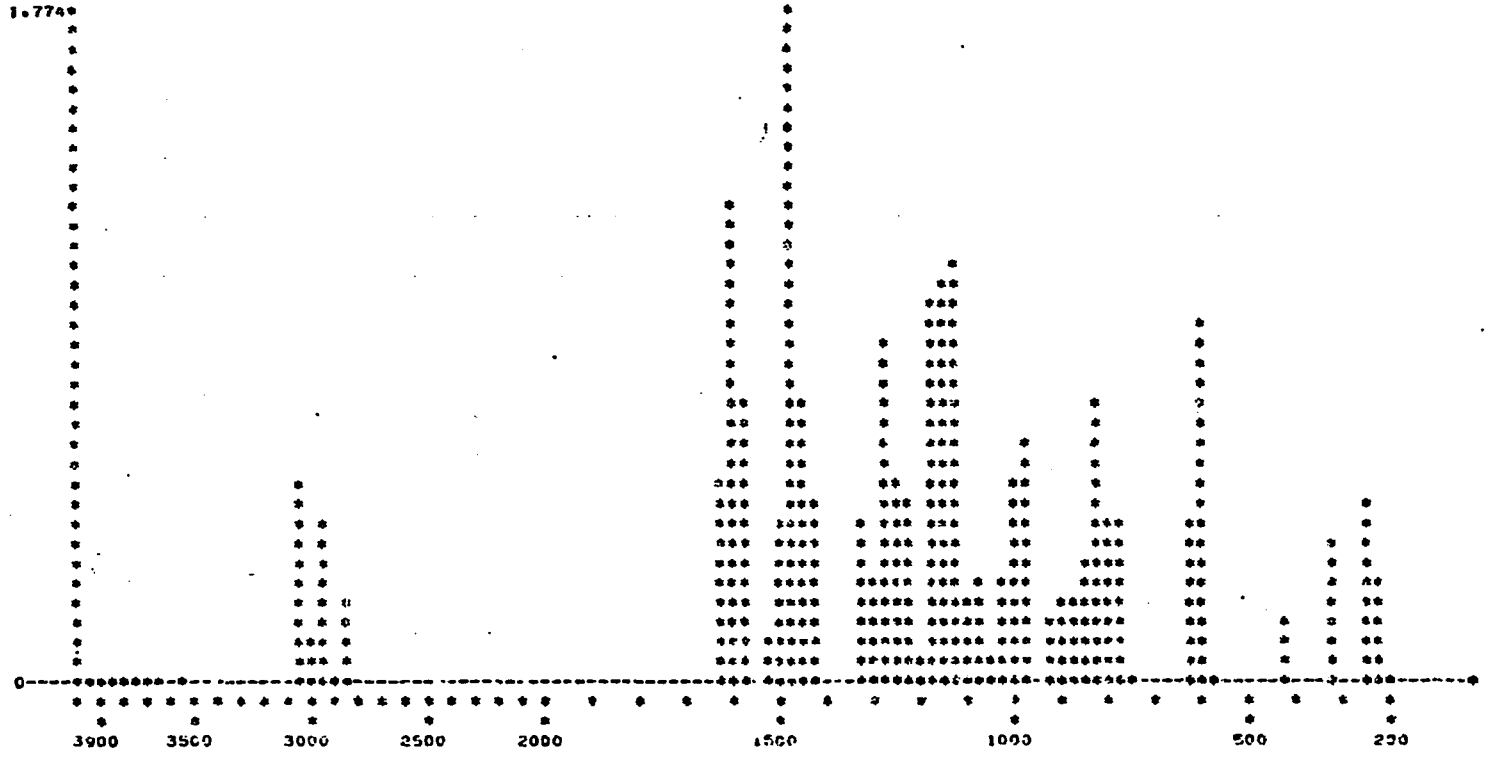
Raman Parallel Polarized

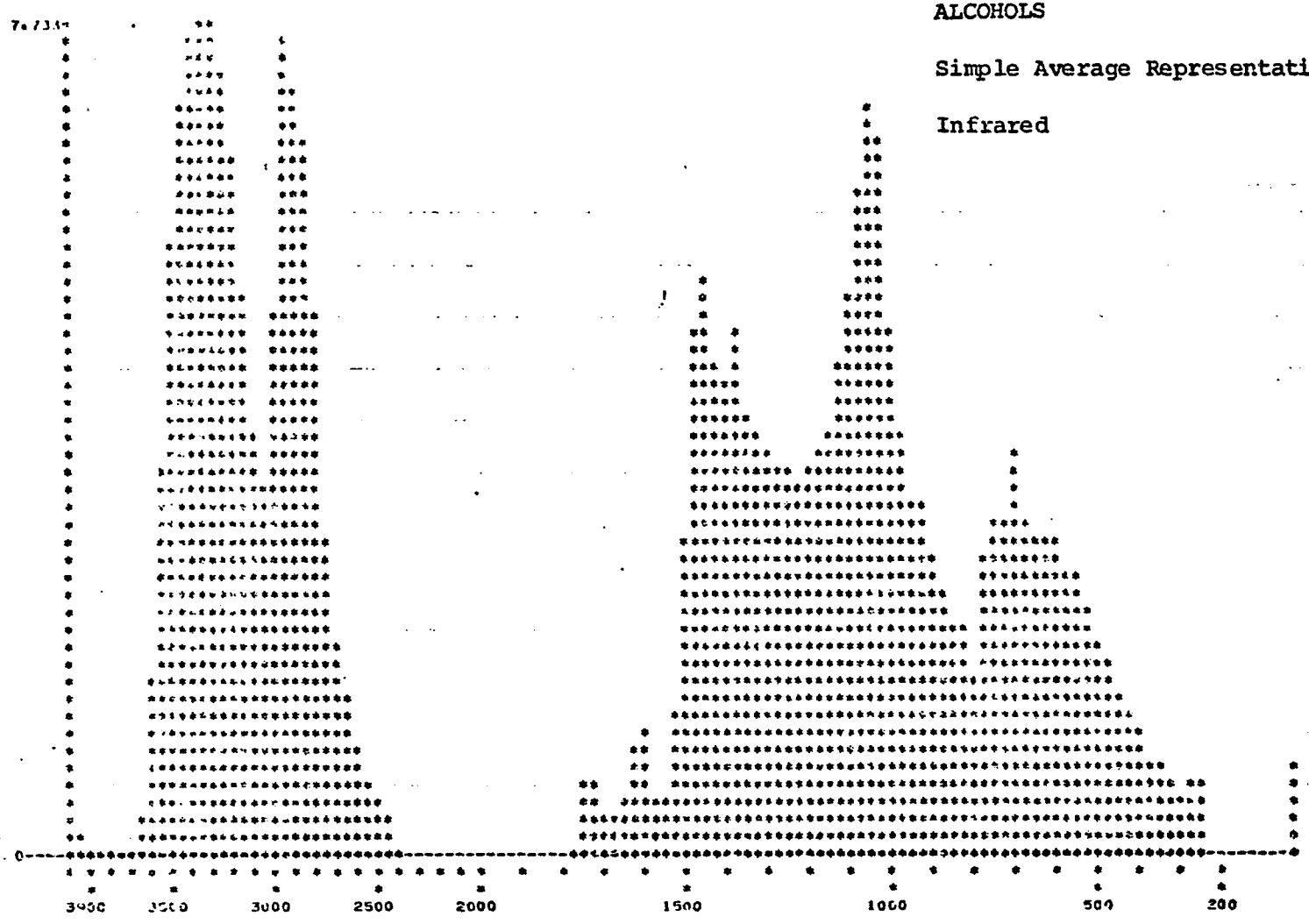


ETHERS

Muted Average Representation

Raman Perpendicularly Polarized






ALCOHOLS

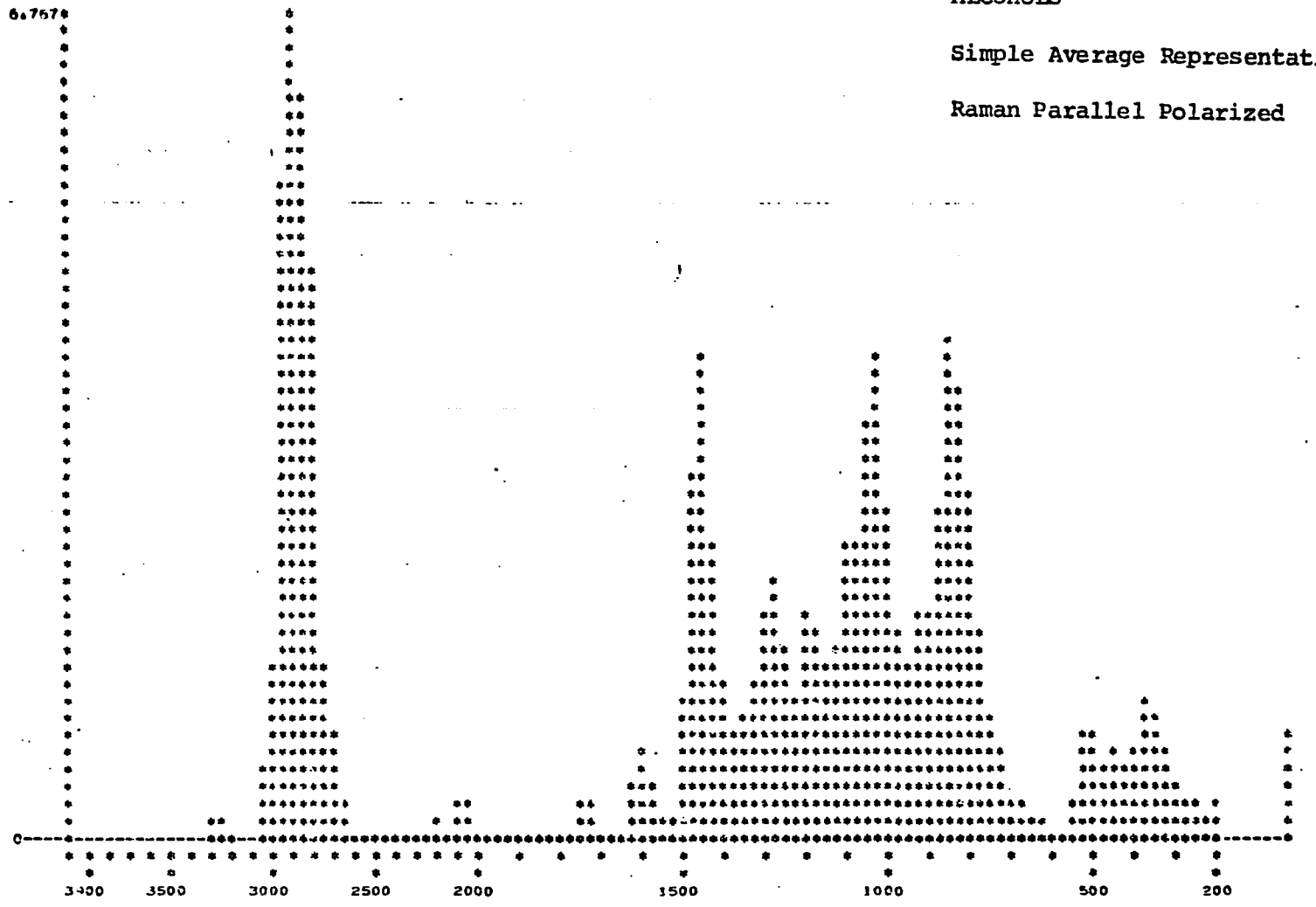
Simple Average Representation

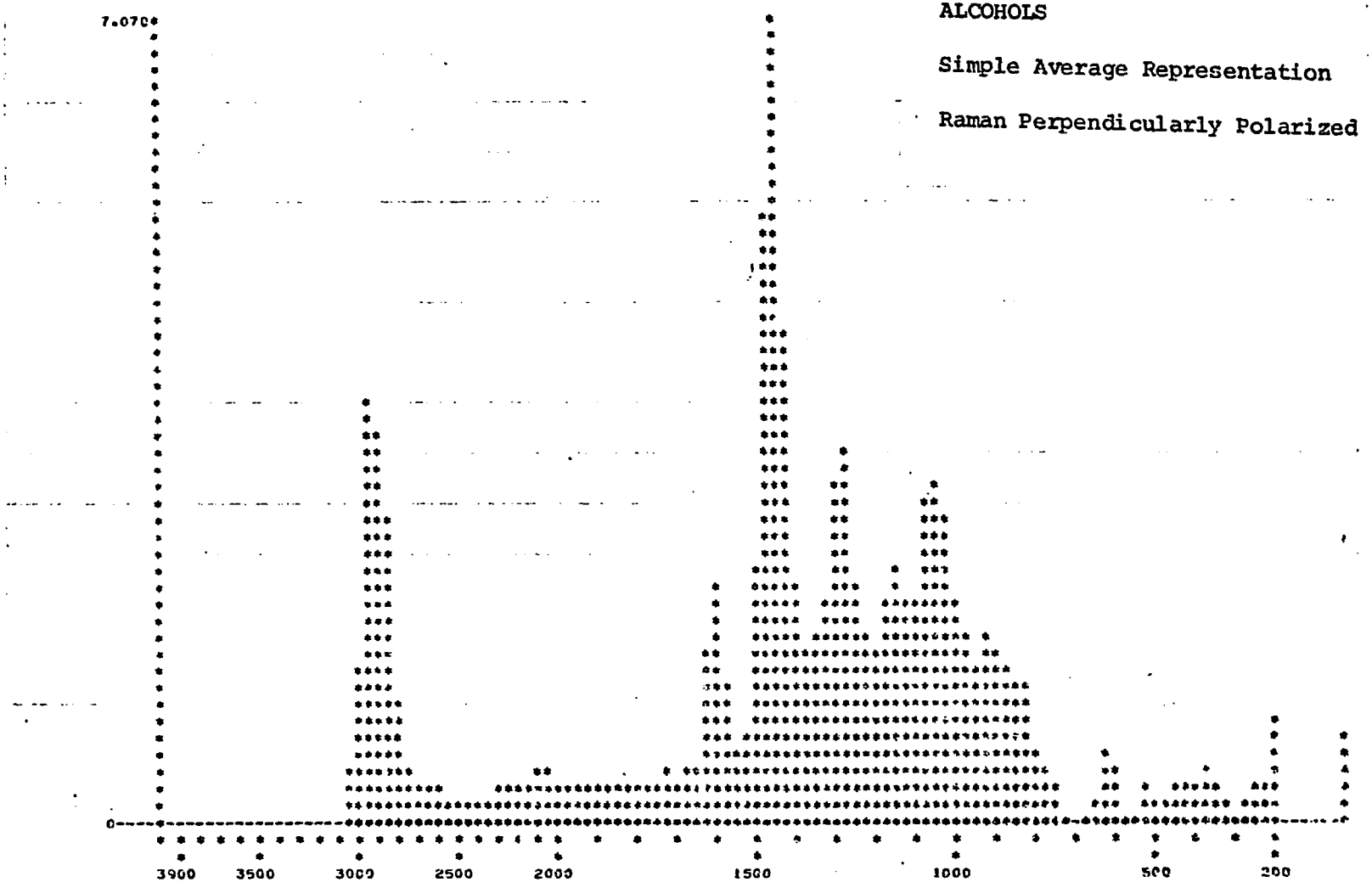
Infrared

ALCOHOLS

Simple Average Representation

Raman Parallel Polarized






ALCOHOLS

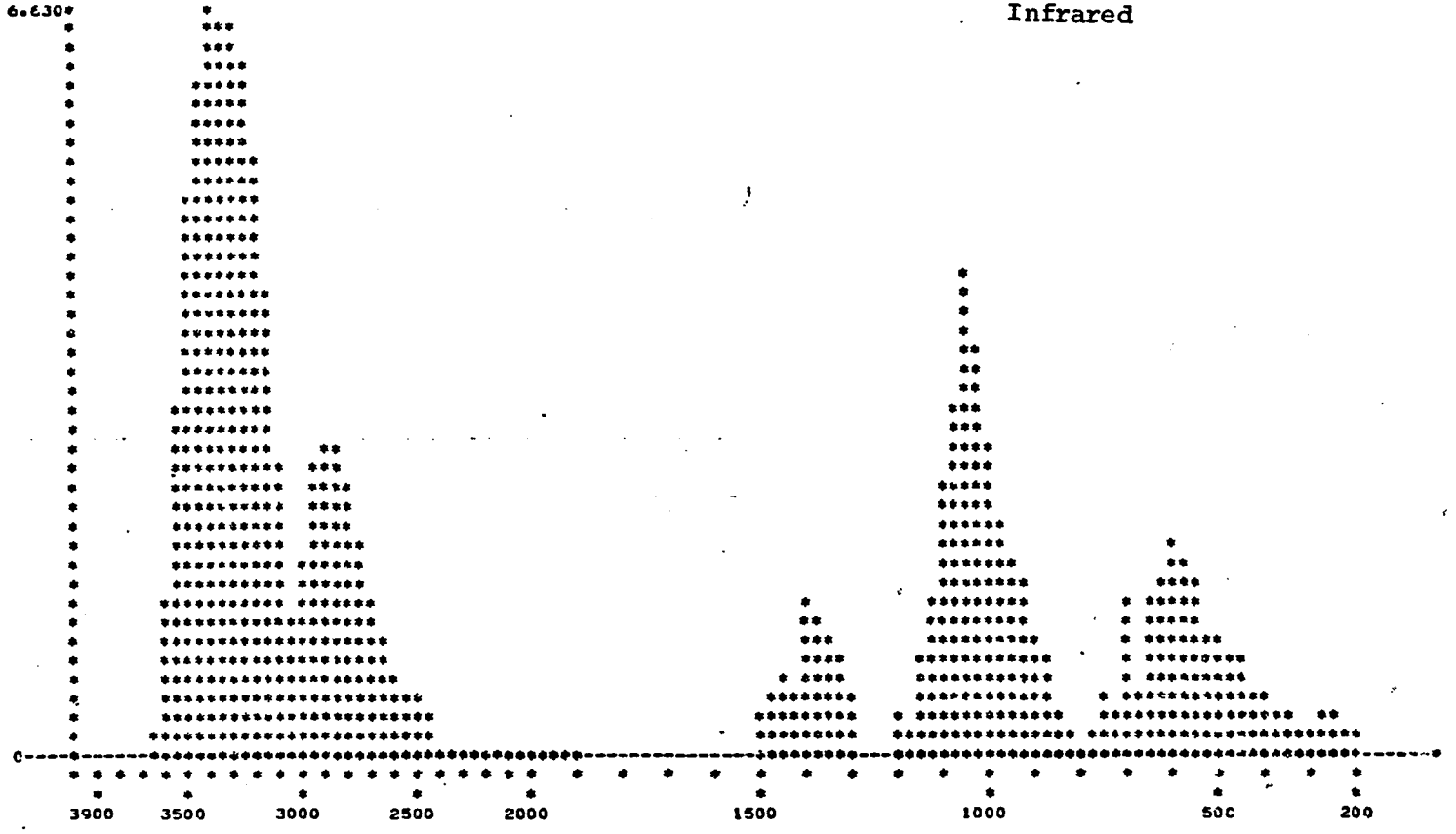
Simple Average Representation

Raman Perpendicularly Polarized

ALCOHOLS

Muted Average Representation

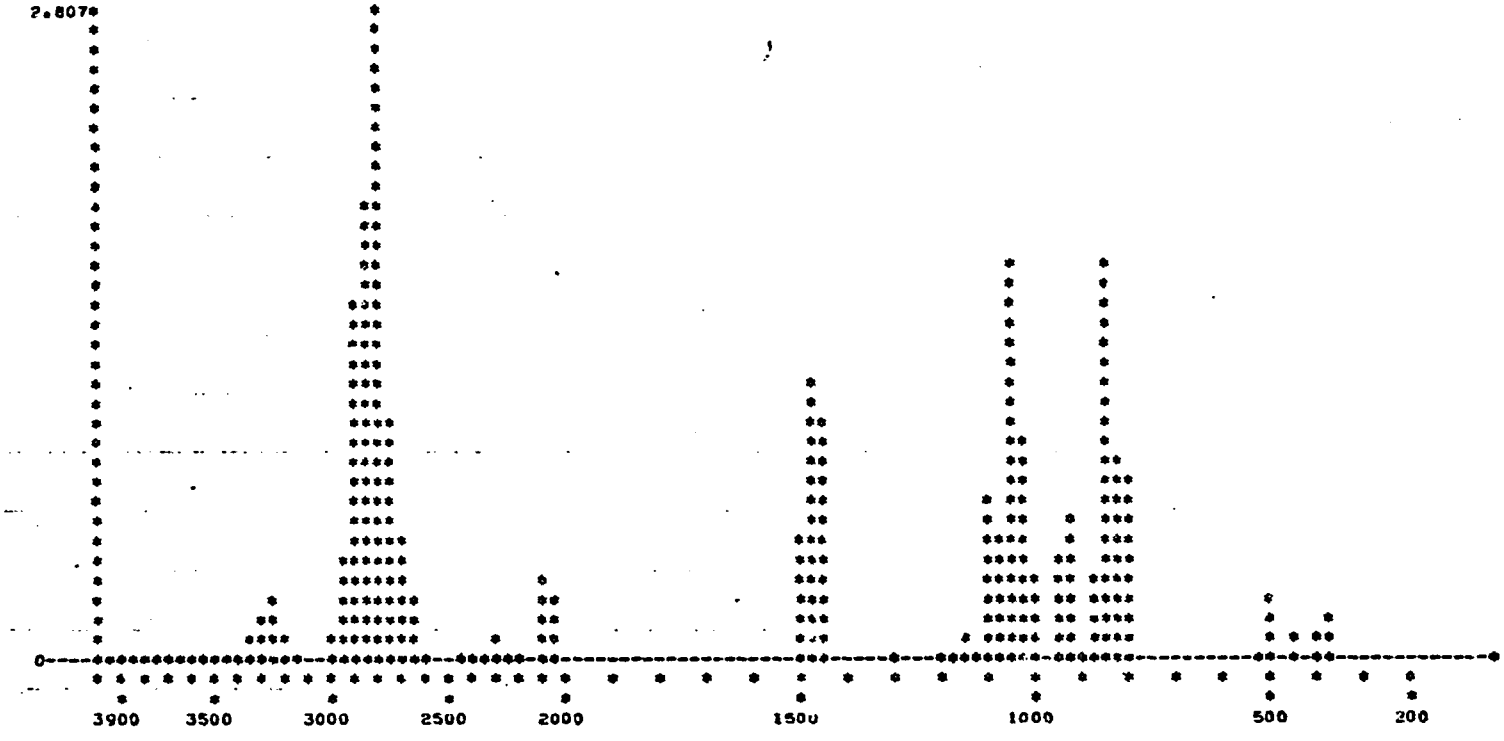
Infrared



ALCOHOLS

Muted Average Representation

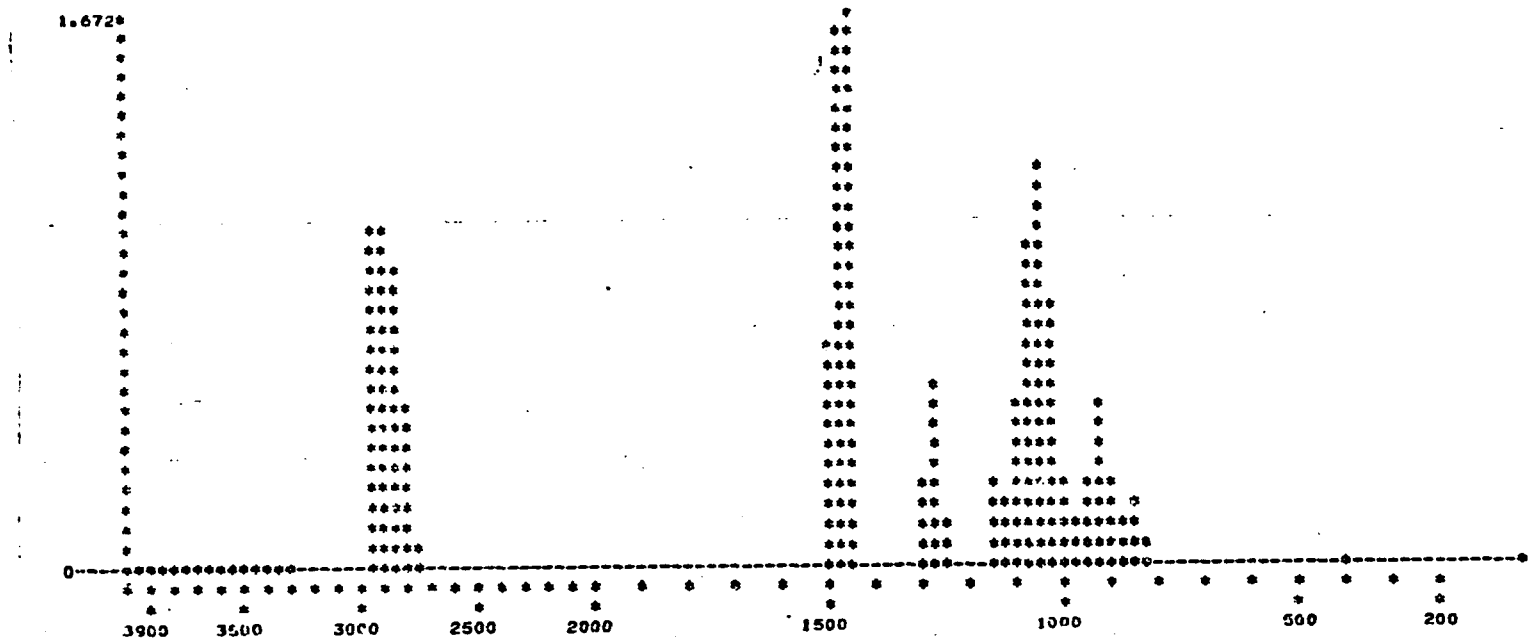
Raman Parallel Polarized



ALCOHOLS

Muted Average Representation

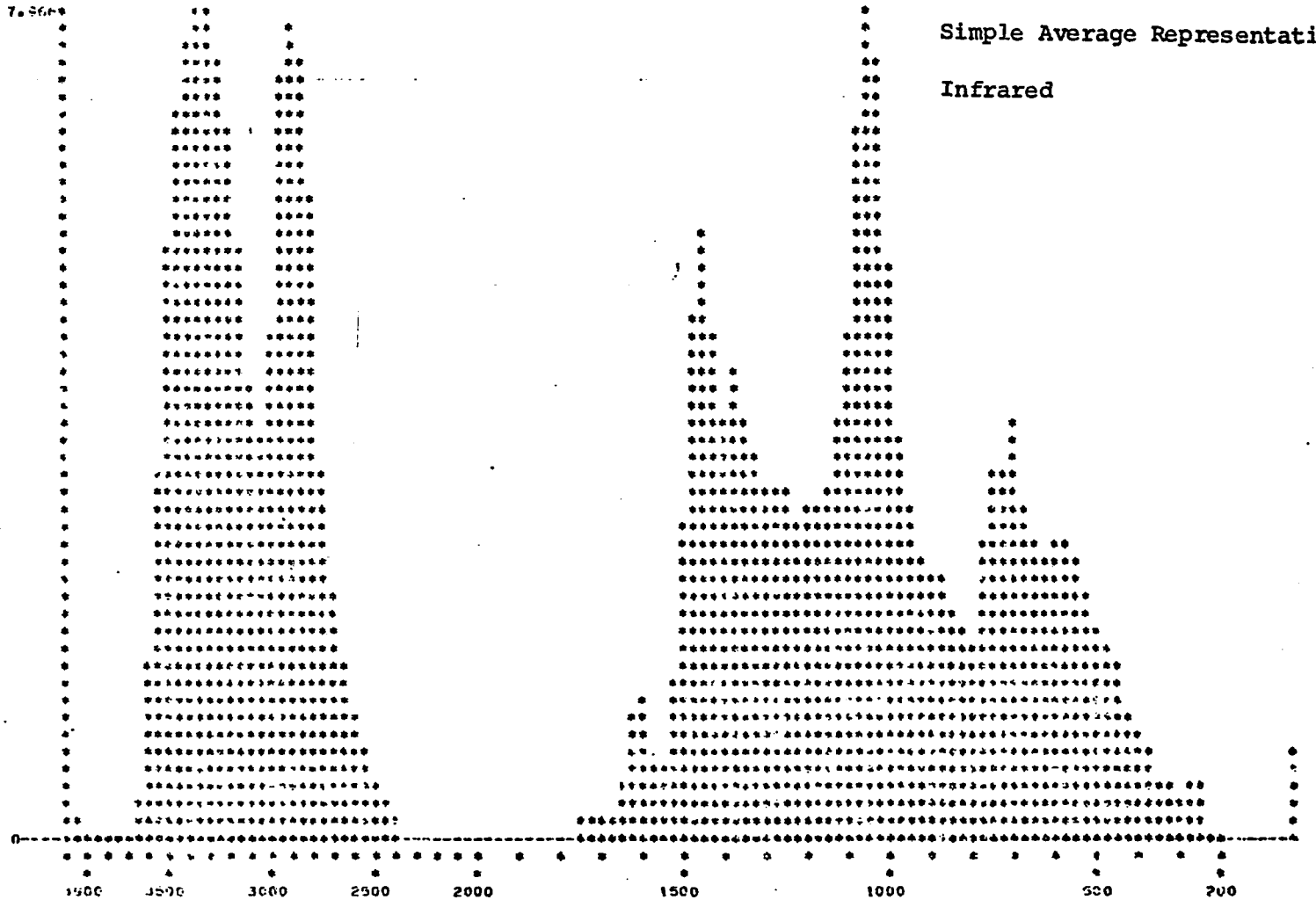
Raman Perpendicularly Polarized



PRIMARY ALCOHOLS

Simple Average Representation

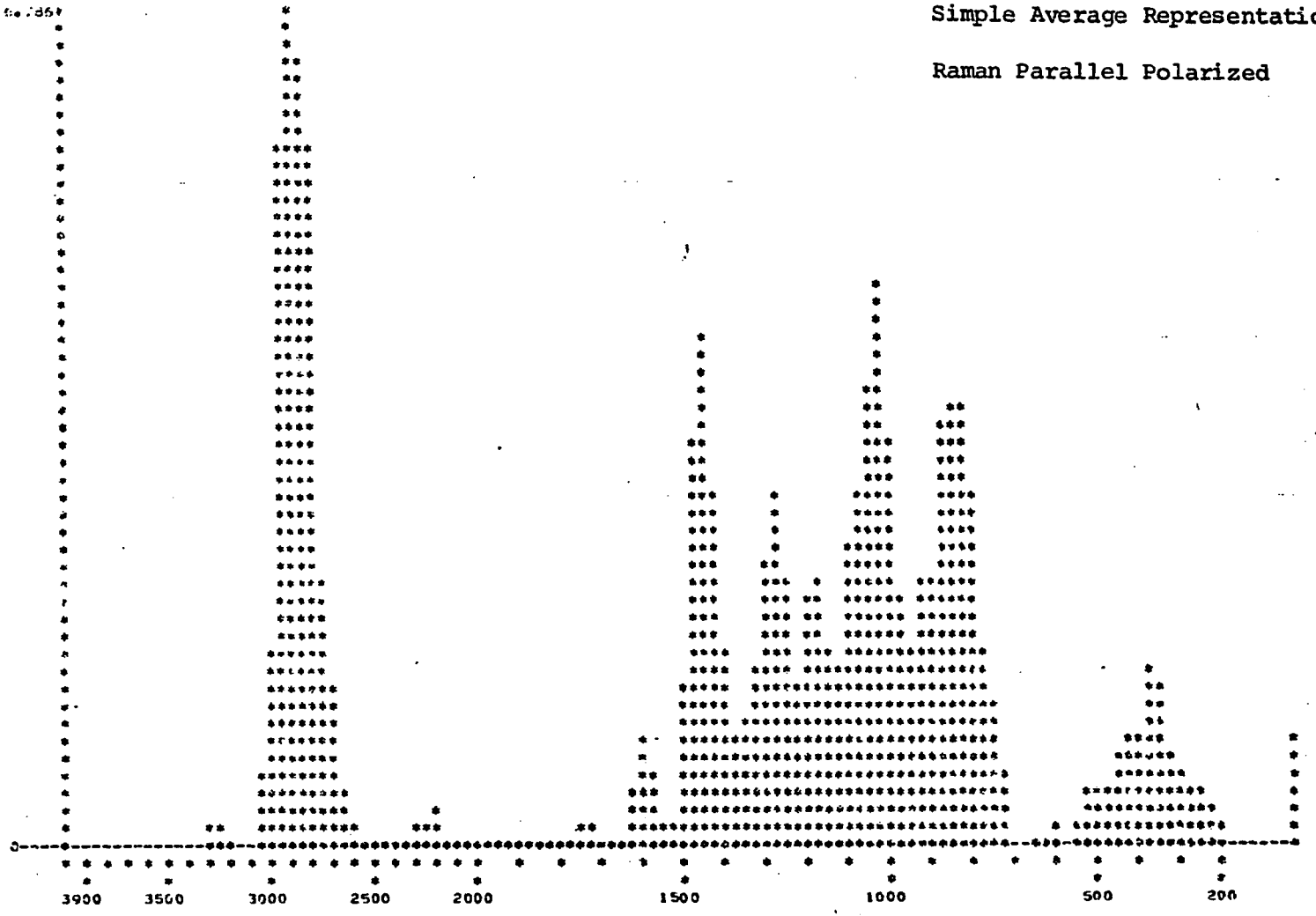
Infrared



PRIMARY ALCOHOLS

Simple Average Representation

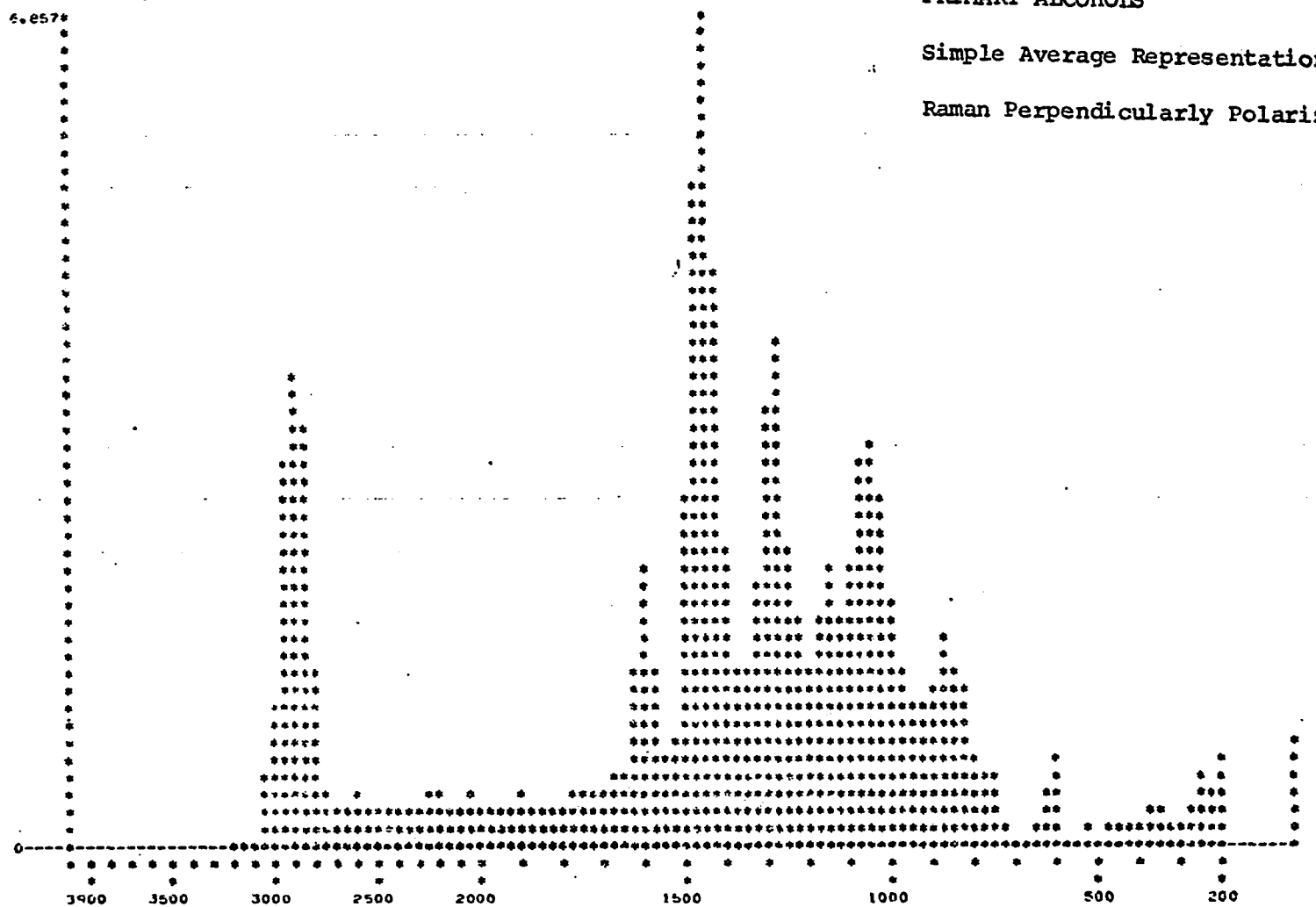
Raman Parallel Polarized



PRIMARY ALCOHOLS

Simple Average Representation

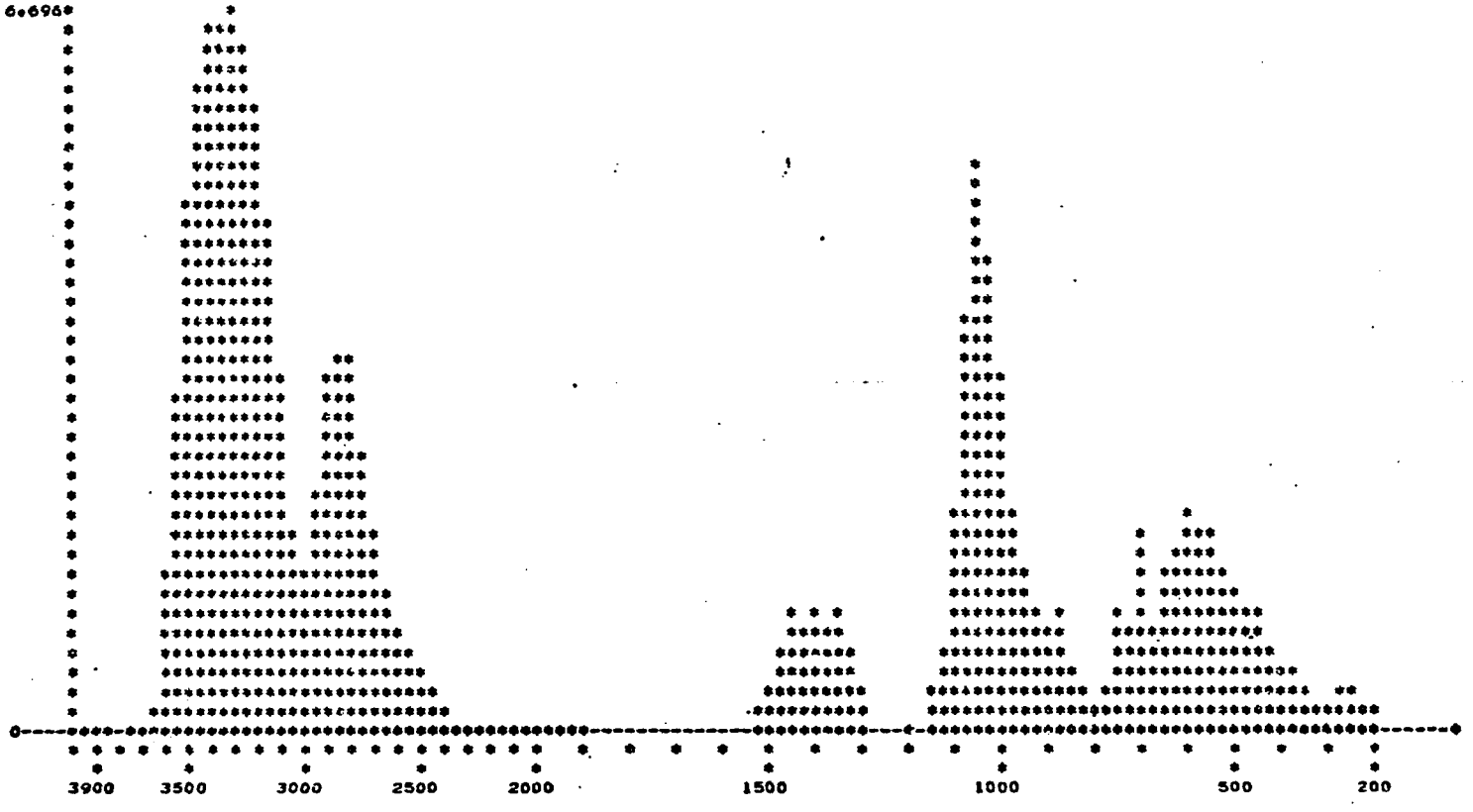
Raman Perpendicularly Polarized



PRIMARY ALCOHOLS

Muted Average Representation

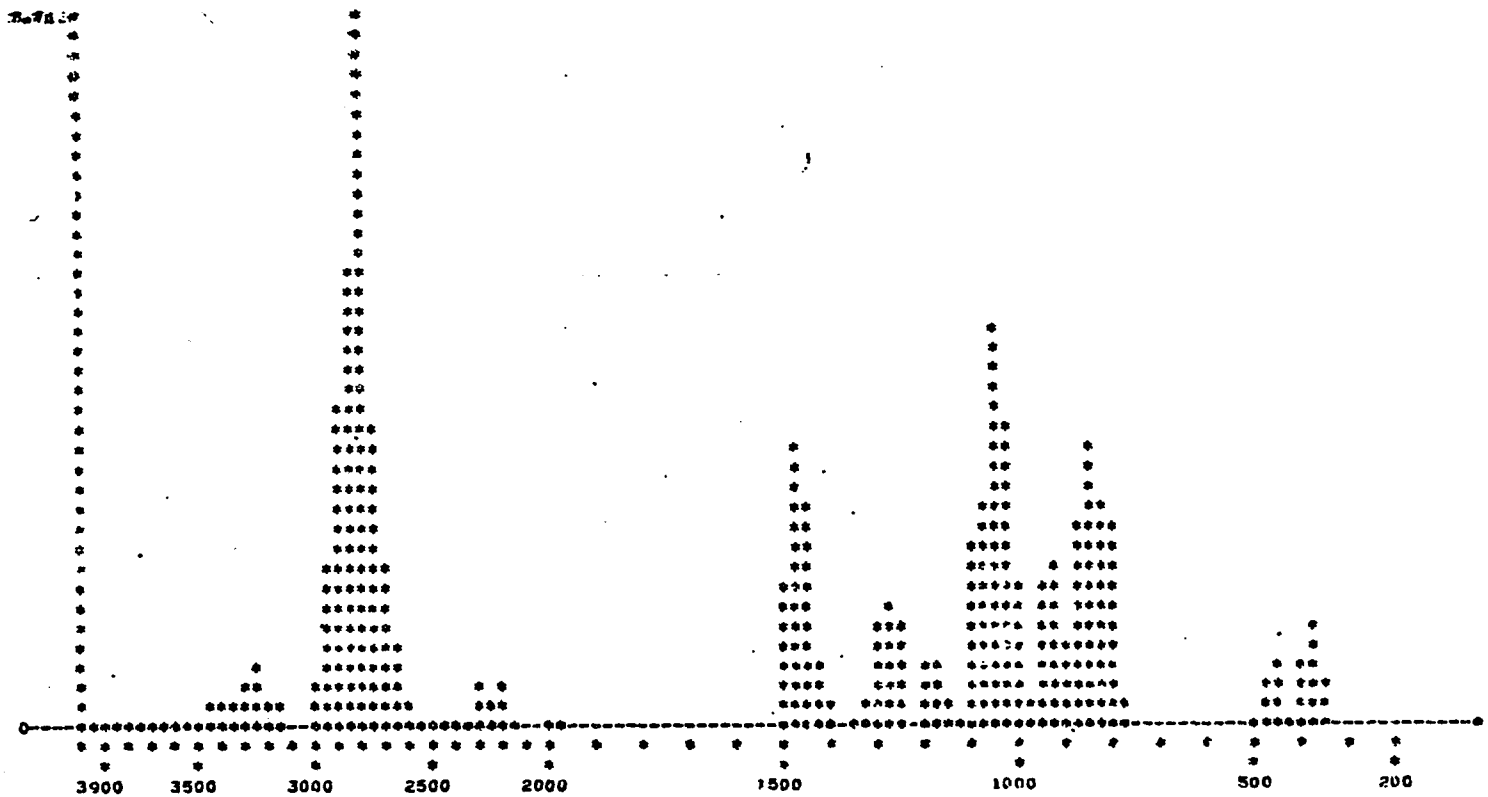
Infrared



PRIMARY ALCOHOLS

Muted Average Representation

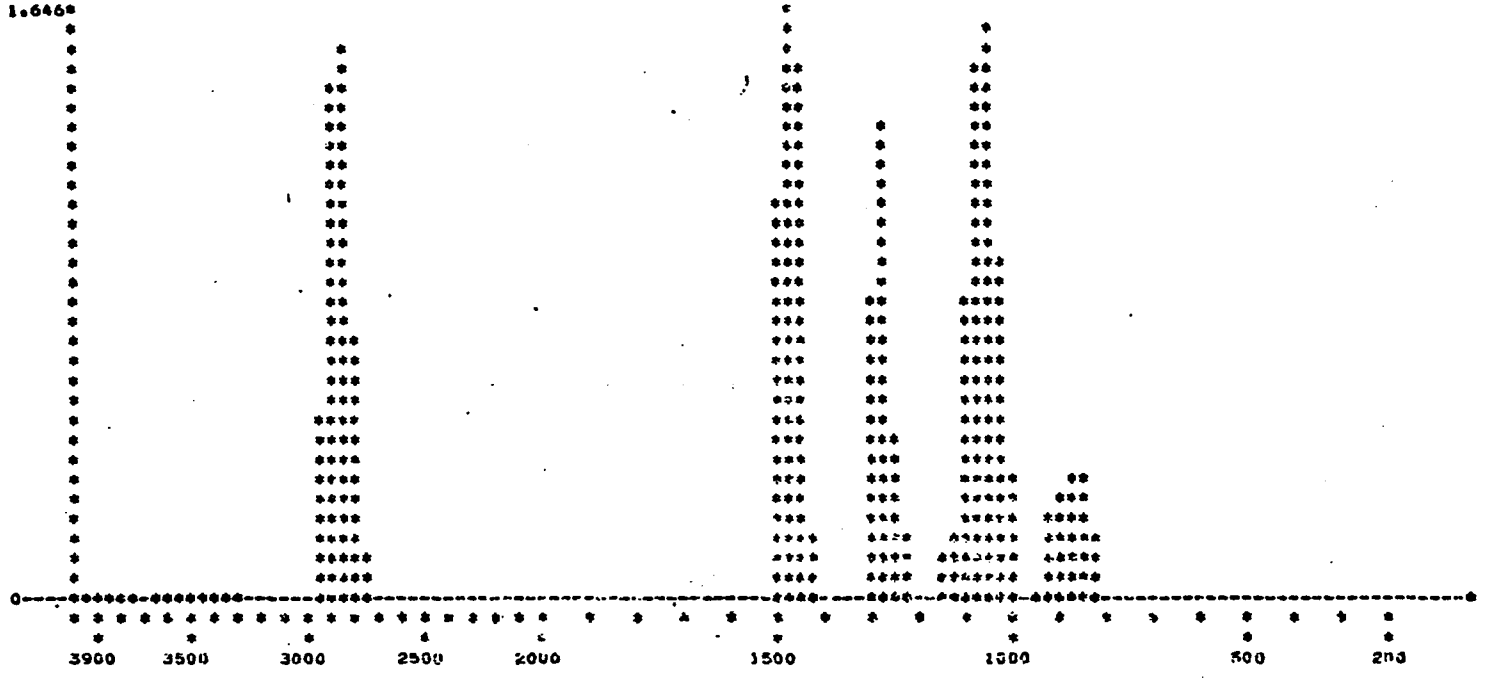
Raman Parallel Polarized



PRIMARY ALCOHOLS

Muted Average Representation

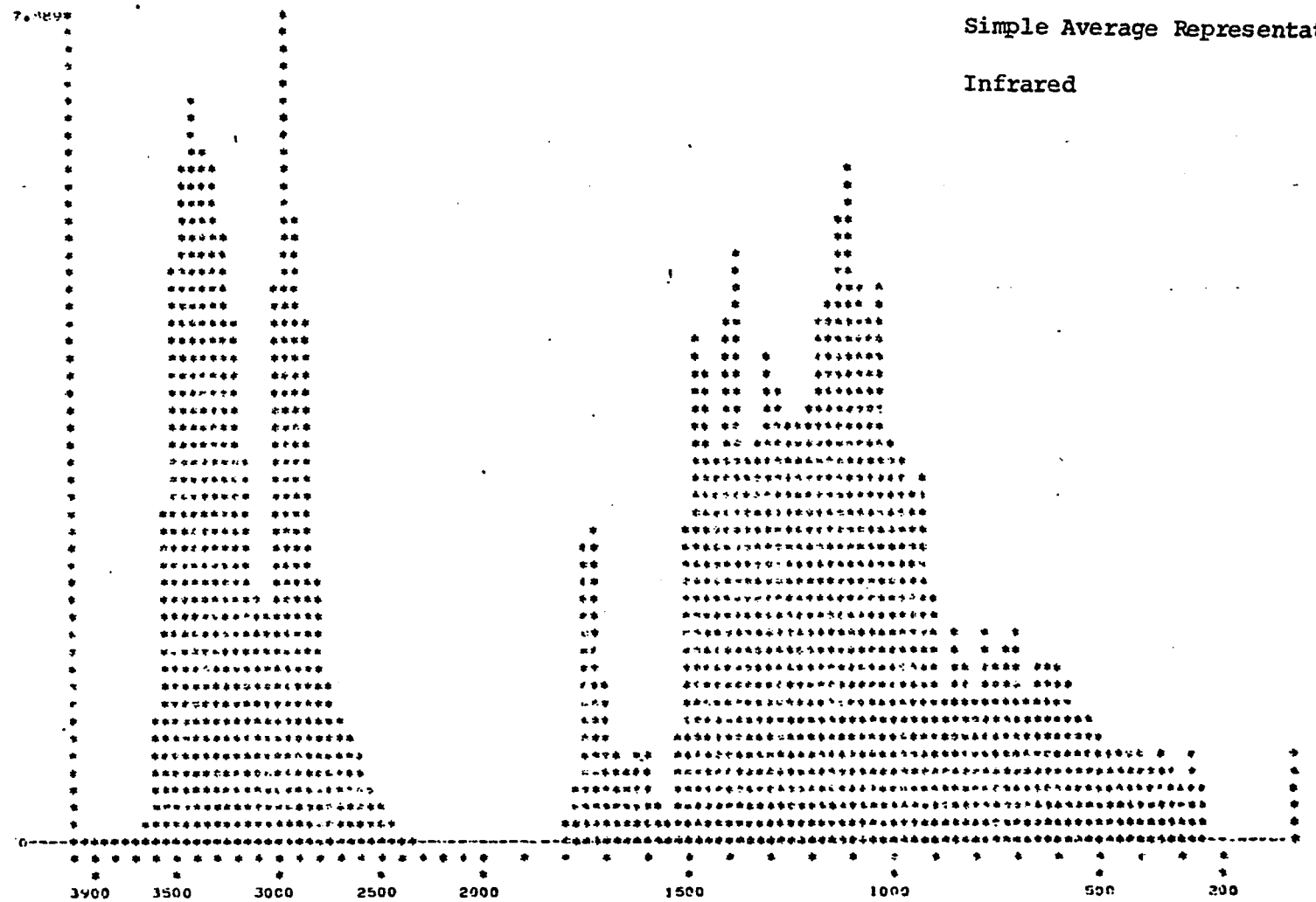
Raman Perpendicularly Polarized



SECONDARY ALCOHOLS

Simple Average Representation

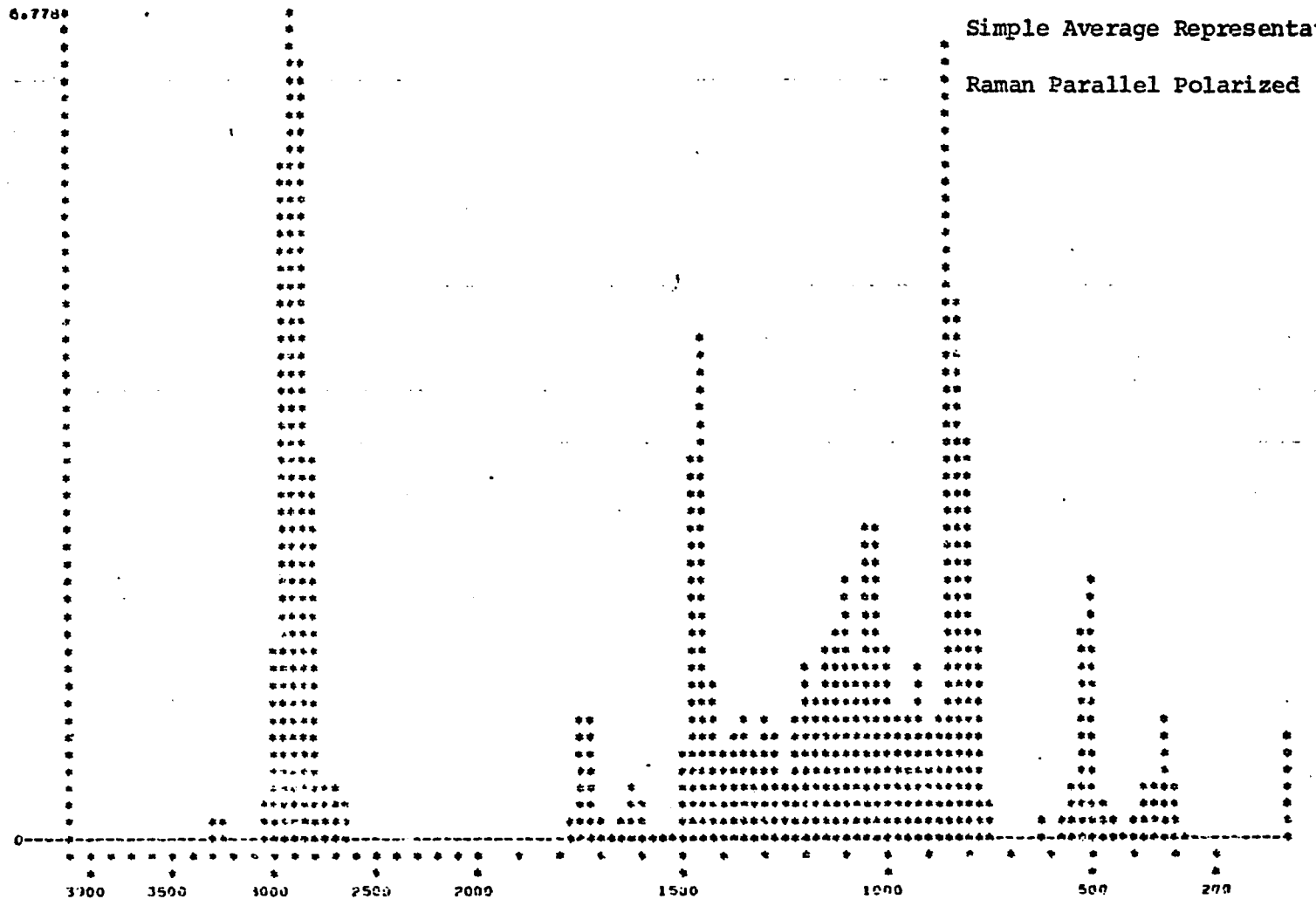
Infrared



SECONDARY ALCOHOLS

Simple Average Representation

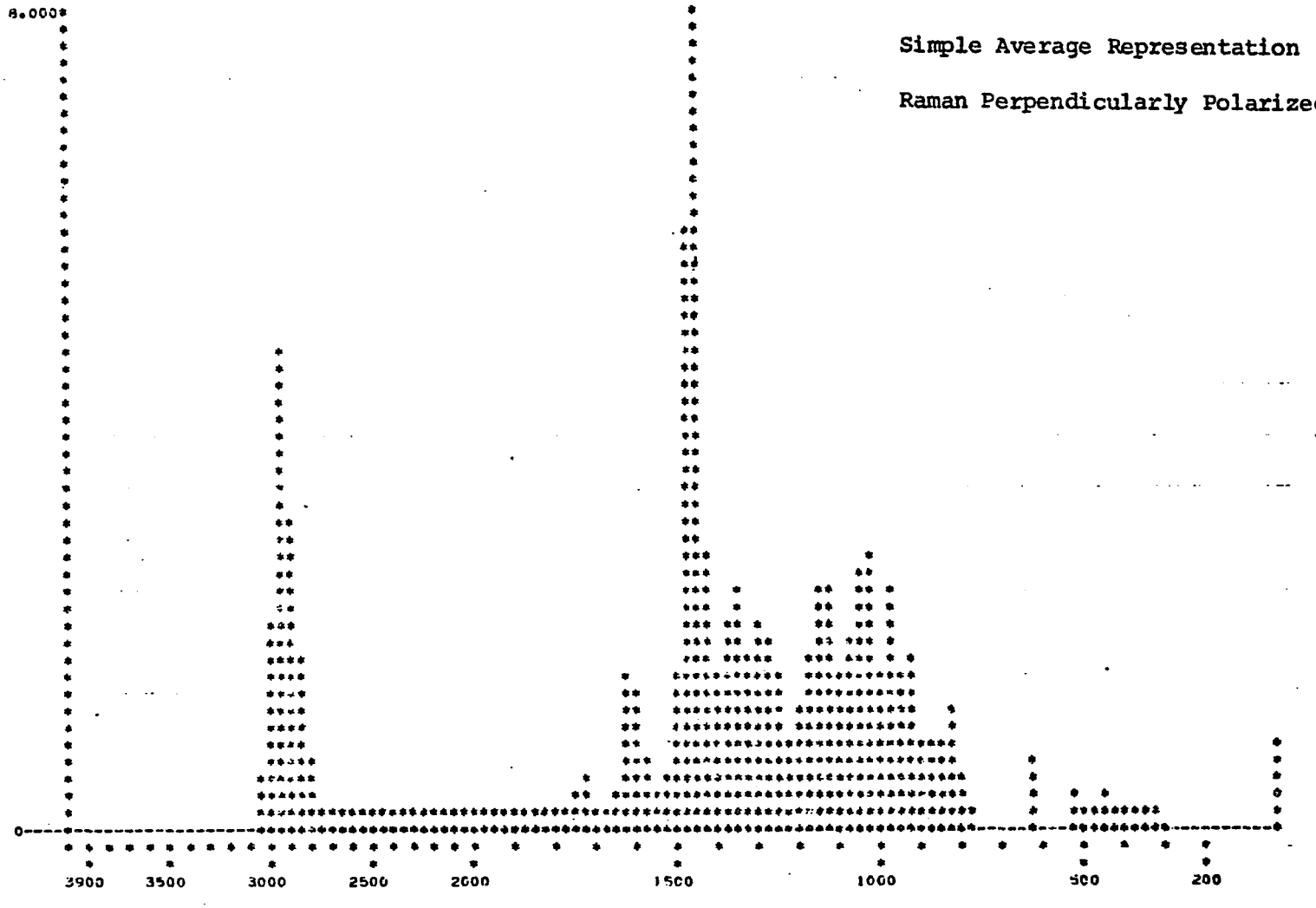
Raman Parallel Polarized



SECONDARY ALCOHOLS

Simple Average Representation

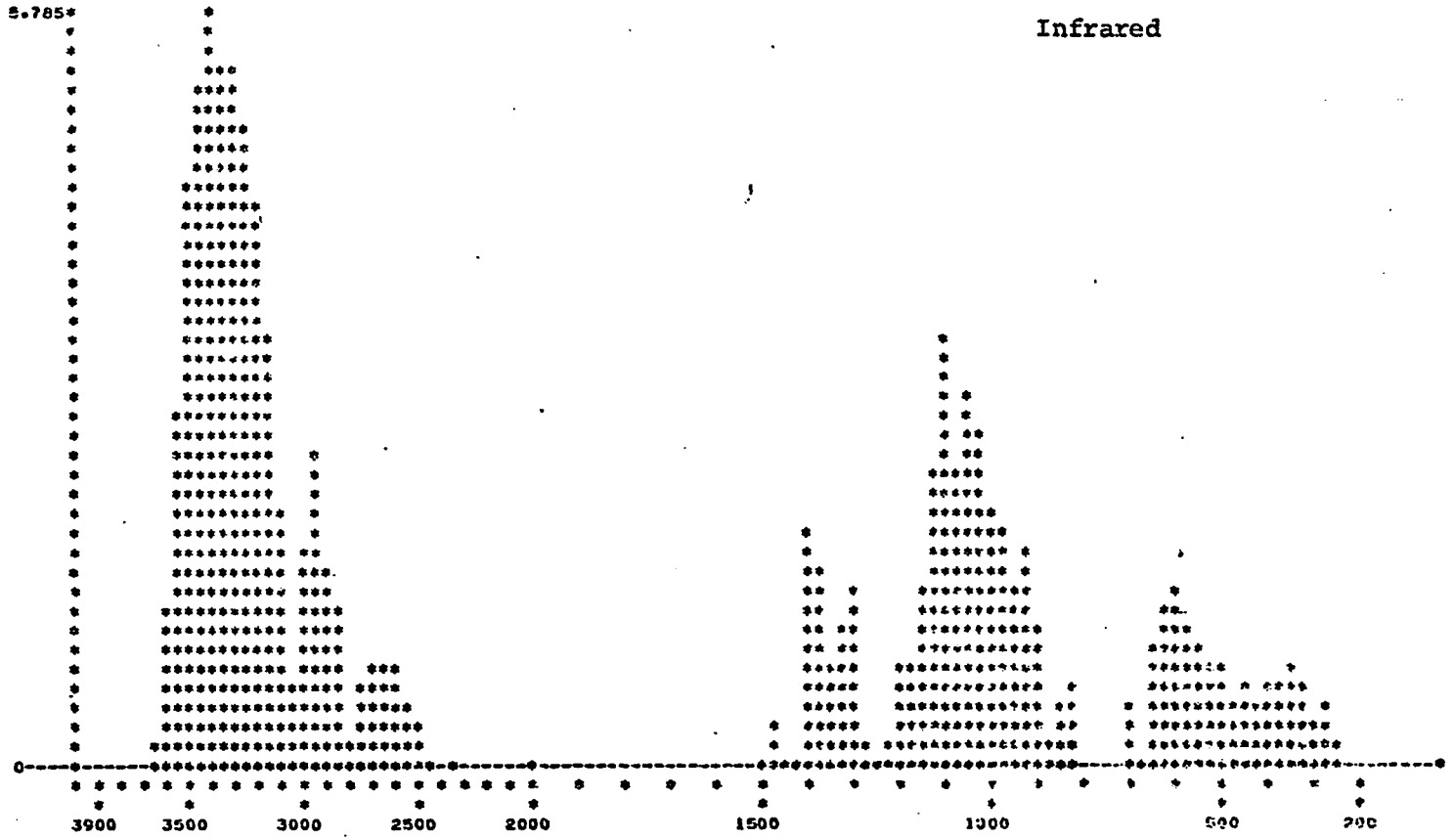
Raman Perpendicularly Polarized



SECONDARY ALCOHOLS

Muted Average Representation

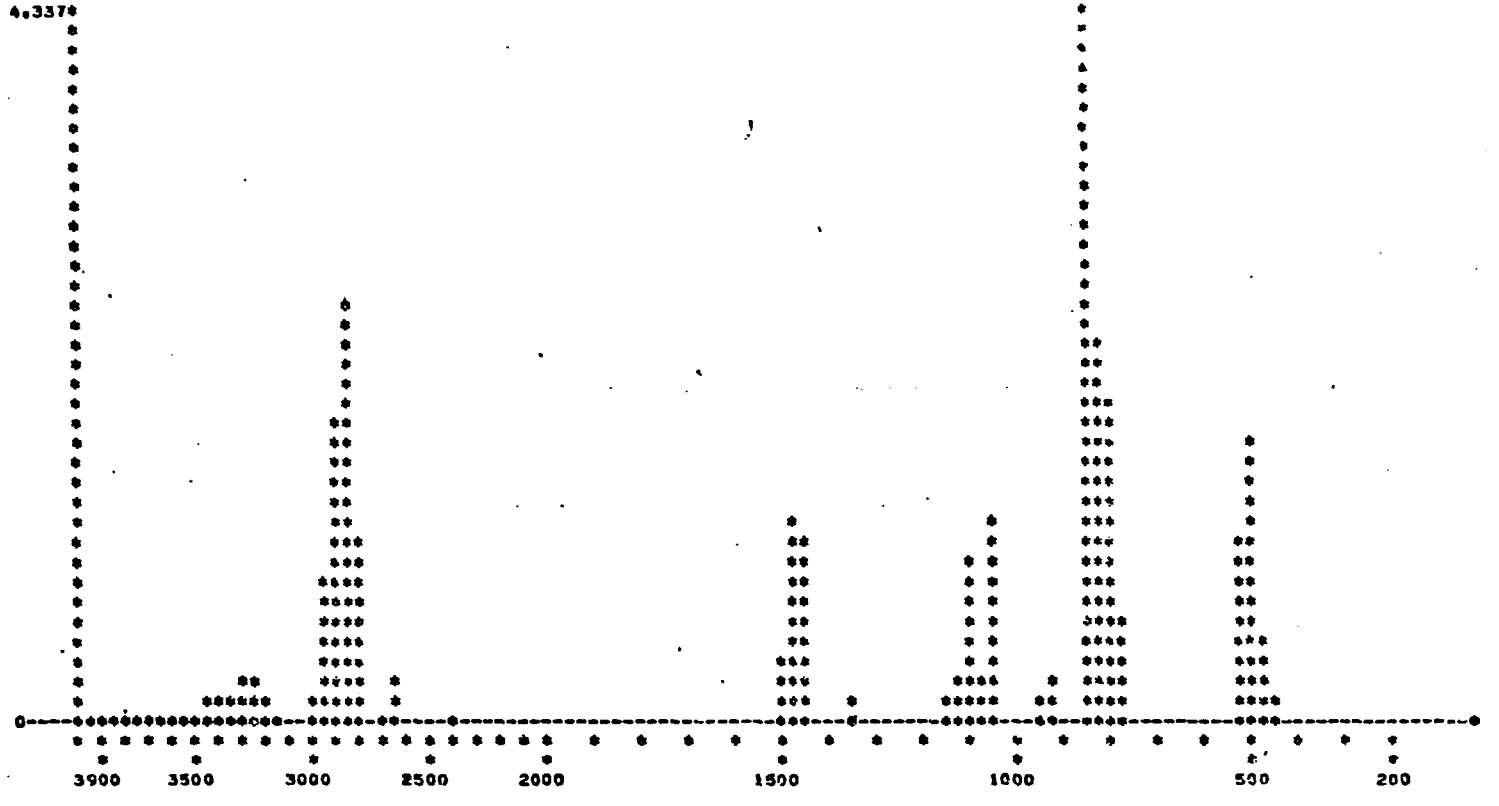
Infrared



SECONDARY ALCOHOLS

Muted Average Representation

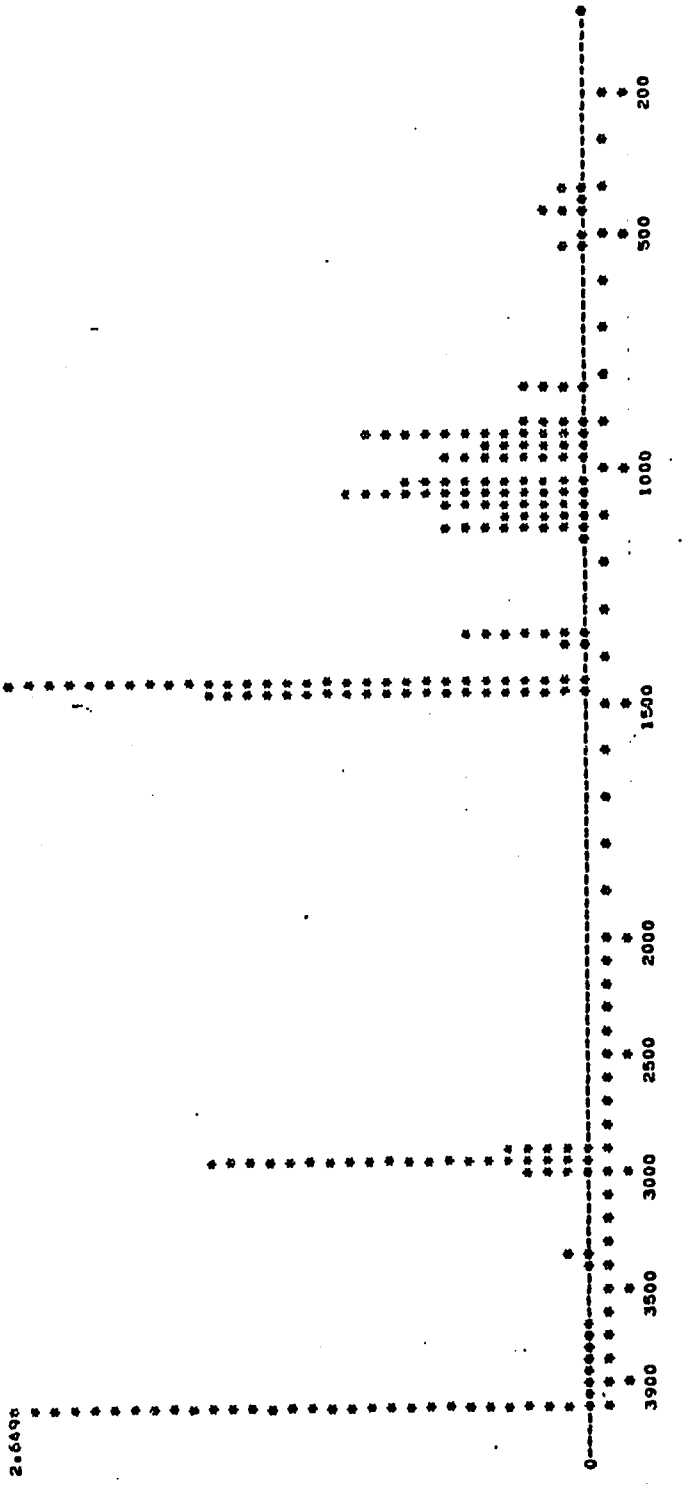
Raman Parallel Polarized



SECONDARY ALCOHOLS

Muted Average Representation

Raman Perpendicularly Polarized



# VITA

The author was born in , . He graduated from Northeast Catholic High School, and received the Bachelor of Science degree in Chemistry from Niagara University, graduating Summa Cum Laude.

In 1967 he married Marie A. Grove of Trenton, New Jersey and they now have four children.

In 1970 he received the Master of Science degree from Newark College of Engineering studying under the guidance of Dr. Howard Kimmel and Dr. William H. Snyder. His doctoral dissertation was also done under Dr. Kimmel, while Dr. Snyder served as one of the members of his advisory committee.

Since 1966 the author has been employed by Personal Products Company, one of the Johnson and Johnson Family of Companies, and it was they who funded the entire graduate program and provided the sabbatical year to fulfill the residence requirement.