JSC-07[illegible]07

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

JSC INTERNAL NOTE NO. 73-FM-53

May 4, 1973

(NASA-TM-X-72177) THE APPLICABILITY AND EFFECTIVENESS OF CLUSTER ANALYSIS (NASA) 32 p HC $3.75 CSCL 09B

N75-15332

Unclas
G3/61 06631

THE APPLICABILITY AND EFFECTIVENESS OF CLUSTER ANALYSIS

Mathematical Physics Branch

MISSION PLANNING AND ANALYSIS DIVISION

LYNDON B. JOHNSON SPACE CENTER
HOUSTON, TEXAS

JSC-07907

JSC INTERNAL NOTE NO. 73-FM-53

THE APPLICABILITY AND EFFECTIVENESS OF CLUSTER ANALYSIS

By D. S. Ingram, IBM, and
A. L. Actkinson, Mathematical Physics Branch

May 4, 1973

MISSION PLANNING AND ANALYSIS DIVISION
NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
JOHNSON SPACE CENTER
HOUSTON, TEXAS

Approved: E R Schiesser
Emil R. Schiesser, Acting Chief
Mathematical Physics Branch

Approved: James C McPherson for
John P. Mayer, Chief
Mission Planning and Analysis Division

CONTENTS

FIGURES

THE APPLICABILITY AND EFFECTIVENESS OF CLUSTER ANALYSIS

By D. S. Ingram, IBM, and
A. L. Actkinson, Mathematical Physics Branch

1.0 SUMMARY

The objective of this internal note is to provide insight into the characteristics which determine the performance of a clustering algorithm. It demonstrates that, in order for the techniques which are examined to accurately cluster data, two conditions must be simultaneously satisfied. The first condition is that the data must have a particular structure, and the second is that the parameters chosen for the clustering algorithm must be correct. By examining the structure of the data from the C1 flight line, it is clear that there is no single set of parameters that can be used to accurately cluster all the different crops. The effectiveness of either a noniterative or iterative clustering algorithm to accurately cluster data representative of the C1 flight line is questionable. This means that, in order to use cluster analysis in its present form for applications like assisting in the definition of field boundaries and evaluating the homogeneity of a field, one must have extensive a priori knowledge. Modifications to existing techniques, or entirely new techniques, are necessary for clustering to be a reliable tool for representative data sets.

A modification to existing clustering methods is proposed. This involves the use of goodness of fit tests to determine, in a quantitative manner, a measure of the unimodality of a cluster. This also has applications to quantitatively evaluating the homogeneity of test and training fields.

2.0 INTRODUCTION

Cluster analysis is a decision-making process in which similar measurements are grouped together. The primary advantage of cluster analysis is that it is not necessary to assume a statistical model for the data. Typical applications which have been identified are evaluating field homogeneity, boundary definition, selecting homogeneous data from nonhomogeneous data, and use as an unsupervised classifier. An objective of this internal note is to determine the factors which affect the ability of a clustering algorithm to perform these functions. These factors are examined in view of the data analysis requirements associated with processing multispectral scanner data for agricultural crops from the C1 flight line.

For the clustering algorithms (ref. 1) which are examined to accurately cluster data, two conditions must be simultaneously satisfied. First, the data must have a particular structure, and, second, the correct parameters must be used in the clustering algorithm. To demonstrate these conditions some experiments using two sets of simulated data are described. The structure of the first set of data is such that the clustering algorithm will accurately cluster the data. The statistics of the second set of data are determined from the Cl flight line. These experiments indicate that the results obtained by using existing cluster analysis techniques to evaluate field homogeneity, boundary definition, selecting homogeneous data from nonhomogeneous data, and as an unsupervised classifier are likely to have little meaning unless one essentially knows the answer before the data are processed. This is particularly significant because it means that it would be necessary to determine the parameters to be used in the algorithm for each flight line and for each different set of crops.

The key question which must be answered to make clustering a scientific technique rather than an art is whether a set of data is unimodal or multimodal. It is proposed that two goodness of fit tests be investigated in order to quantify the concepts of unimodality and homogeneity. This would be very valuable in determining the appropriateness of the assumptions of the probability density function of the multivariate data. The assumption of the multivariate normal distribution is used extensively in feature selection and pattern classification.

3.0 ANALYSIS

In this section clustering is initially described from an intuitive point of view. The relationship between the form of the data and the result produced by a clustering algorithm is investigated for some limiting cases. The results obtained by processing data that corresponds to an agricultural image are presented for two sets of simulated observations. The first set is chosen such that the algorithm can produce accurate results. The statistics of the second set of data were determined from the Cl flight line. Both a noniterative and an iterative algorithm are used to process the data, which are representative of the Cl flight line.

As is demonstrated, the structure of the data corresponding to the agricultural crops on the Cl flight line is such that no single set of parameters can be used to accurately cluster the data. The cluster results is dependent on the parameters used in the algorithm. Hence, any measure of field homogeneity is input-parameter-dependent. The fundamental problem, then, is to determine whether or not a set of data is unimodal.

3.1 Cluster Analysis

Cluster analysis is a decision-making process in which similar measurements are grouped together. The performance of an algorithm to group data together depends on the structure of the data. To illustrate this condition, consider two sets of two dimensional data. In figure 1(a) the data are uniformly spaced in the x_1, x_2 coordinates and in figure 1(b) the data are neatly grouped into three distinct subsets.

In order to apply a cluster algorithm, a function must be used which determines whether two observations are similar. For the sake of illustration let the similarity function be a distance measure. If a measurement is within a specified radius of a cluster mean, then that measurement is an element of the specified cluster. The specified radius is a parameter of the algorithm. Consider the relationship between the number of clusters and the radius, R, for the data in figure 1(a). If R is very small, then the number of clusters, N, equals the number of data points, and if R is very large, then there is one cluster. As R changes from very small to very large the structure of the N versus R cu[illegible] would be similar to the graph shown in figure 2(a) for the data of figure 1(a). The same procedure can be carried out for the data of figure 1(b). The limiting cases are the same; however, the number of clusters is constant over a range of R. Since the form of the data is known, it is clear that the correct number of clusters is three and that a value of R such that $R_1 < R < R_2$ is acceptable. This is equivalent to finding the set of parameters that produce the correct answer or to training the algorithm.

This example clearly shows that, in order for a clustering algorithm to effectively cluster data, two conditions must be simultaneously satisfied. The first condition is that the structure of the data must be such that the data can be clustered. If this condition is satisfied then it is necessary to choose the correct parameter in the algorithm. A value of R outside the range of $R_1 < R < R_2$ would not cluster the data correctly.

3.2 Simulated Data

The concept of clustering data from an image is further developed by considering two sets of simulated data. The clustering algorithms used include both a one-pass and an iterative technique. The spatial configuration of the classes in the image is similar to an agricultural scene. The first set of data is such that the clustering algorithm will effectively cluster the data. The second set of data is representative of the C1 flight line. Neither the noniterative nor the iterative algorithm is effective on the simulated C1 data.

3.2.1 Ideal case.- The simulated data generated for this case are described in reference 2. The noniterative clustering algorithm used is the CLUST1 option in ASTEP (ref. 1). Figure 3 illustrates the way the image is subdivided. For this example only two channels of data, 11 and 12 of reference 2, are processed. The mean ±1 standard deviation for each class are plotted in figure 4. Each element of the field is generated from a normal distribution with a mean of μ_i and standard deviation of σ_i for $i = 1,5$ for each channel. The data are uncorrelated from channel to channel.

The data were clustered for several values of R with the condition that C = 2R. Although it is not obvious that the condition C = 2R will yield "best" results, this condition does appear to be a reasonable way of relating C and R. In each case the initial value for the maximum number of clusters was 20 and the initial values for the means of those clusters was 0. The results of the plot of the number of clusters versus R is shown in figure 5. The number of clusters is constant for values of $10 < R < 30$ and at each value of R the clusters are the same. The image map displayed by ASTEP is shown in figure 6 for R = 20. Each observation is classified correctly. This is exactly the result one would expect for the structure of the data in figure 4.

The question which one must ask is how to know that each of the five clusters is unimodal. For values of $32 < R < 50$ there are three clusters. The clusters are not the same three clusters for all values of R. However, for R = 32, 34, and 36 the three clusters are the same and one might suspect that there are three clusters instead of five.

3.2.2 Simulated C1 data.- The statistics of the crops along the C1 flight line are listed in table 1. The mean ±1 standard deviation for each class in channels 6, 10 and 12 is plotted in figures 7 and 8. The statistics listed in table I were used to generate a data tape which represents the fields in figure 3. The ellipses which are drawn in figures 7 and 8 have their principal axes parallel to the measurement coordinates. This is not the case for C1 data, as the principal axes of each of the ellipses would be inclined to the measurement coordinates. Examining figures 7 and 8 it is clear that the structure of the data is such it would be difficult to find a set of parameters that could cluster all the crops. The most obvious reason for this is the size of the standard deviation of Wheat2 as compared to the difference between the means of the other crops, such as Alfalfa.

The noniterative algorithm described in reference 1 was used to process these data with the same set of initial conditiors described in section 3.2.1. The results of N versus R is shown in figure 9. There is clearly no well-defined interval for which there are eight clusters. For this set of conditions the best results appear to occur for $R = 5$ as shown in figure 10. It is possible to decide what is best only because we know the answer.

Soybeans and Bare Soil are accurately classified. Corn1 and Oats are classified with fair accuracy. It is not possible to distinguish among Red Clover, Red Clover2, Alfalfa, and Wheat 2. For this case the Red Clover field appears to be nonhomogeneous while the Corn1 field appears to be homogeneous. As another illustration of the concept of homogeneity consider figure 11. In this case $R = 16$ and the image divides nicely into two categories, B and C. The field labeled C appears to be homogeneous and indeed is Wheat2. The field labeled B appears to be homogeneous and indeed consists of seve: different crops.

The same data were processed with ISODATA (ref. 4) using the same parameters suggested in reference 5, namely DLMIN and STDMAX equal to 3.2 and 4.5, respectively. For the best case (fig. 12), Red Clover and Alfalfa were indistinguishable and the accuracy of the classification of Oats and Corn is fair. This case took 20 iterations and used in NMIN value of 30. The term NMIN is the minimum number of points allowed in a cluster. Changing the NMIN value to 15 resulted in Wheat, Red Clover, Red Clover2, and Alfalfa being poorly classified while the accuracies of Corn and Oat classification improved somewhat (fig. 13). The cases illustrate that the choice of NMIN affects the accuracy of clustering. Changing NMIN may cause some accuracies to improve while others deteriorate.

Chaining was applied in each of the above cases. The results without chaining were much worse (figs. 14 and 15 for NMIN equal to 30 and 15, respectively).

Using a different channel set, channels 1, 6, 9, and 12, the ISODATA classification maps were figures 16, 17, and 18, after 18, 19, and 20 iterations, respectively. The value used for NMIN was 30, and no chaining was used. The 20 iterations case (fig. 18) was less accurate than the corresponding three channel case given in figure 14. Also, the results for iteration 18 were much better than those for iteration 19.

These results demonstrate that the ISODATA classification is dependent on the number of iterations, the number and choice of channels, and on the choice of NMIN. No criteria currently exist for selecting these values without extensive a priori knowledge. Even for the best choice, the accuracies were very poor for some crops. The effectiveness of the iterative algorithm to cluster data representative of the C1 flight line is questionable.

4.0 RECOMMENDATION

In order to determine whether a set of data consists of one or more clusters it is necessary to determine whether the data set is unimodal or multimodal. This can be determined in a quantitative manner by using goodness of fit tests. The two tests considered here are the classical chi-squared test and the Kolmogorov-Smirnov test. These could be applied to each potential cluster and the degree to which the cluster is unimodal could be determined.

A chi-squared random variable is defined as the sum of squares of independent standard normal variables (ref. 6). Let X be normally distributed with zero mean and unit variance; then the chi-squared random variable is

$$Q = X_1^2 + X_2^2 + \dots + X_k^2 \tag{1}$$

where there are k independent values of X. The parameter k represents the number of degrees of freedom of the system. The chi-squared random variable can be related to the multivariate normal distribution by noticing that the quadratic form in the multivariate normal p.d.f. is a chi-squared random variable, that is,

$$p(x) = \frac{1}{(2\pi)^{n/2}\left|\sum\right|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \sum{}^{-1} (x-\mu)\right\} \tag{2}$$

and

$$Q = (x-\mu)^T \sum{}^{-1} (x-\mu) \tag{3}$$

where x is the $n \times 1$ random variable, μ is the mean vector and Σ is the $n \times n$ covariance matrix of the multivariate normal distribution. The chi-squared variable in equation (3) has n degrees of freedom.

The probability density function of a chi-squared random variable Q is

$$p(Q) = \frac{1}{2^{n/2}\Gamma(n/2)} Q^{(n-2)/2} e^{-Q/2} \qquad Q > 0 \tag{4}$$

where n is the number of degrees of freedom and Γ is the gamma function. The cumulative distribution function (c.d.f.) of Q is

$$P(Q) = \int_0^Q p(Q)\, dQ \tag{5}$$

Equation (5) can be evaluated in closed form when n is even. The results for $n = 2$, 4, and 6 are

$$\underline{n = 2} \qquad P(Q) = 1 - e^{-Q/2} \tag{6}$$

$$\underline{n = 4} \qquad P(Q) = 1 - e^{-Q/2}(1 + Q/2) \tag{7}$$

$$\underline{n = 6} \qquad P(Q) = 1 - e^{-Q/2}\left[1 + \frac{Q}{2} + \frac{Q^2}{8}\right] \tag{8}$$

The number of degrees of freedom is the same as the number of independent channels of a multispectral scanner.

If a data set has a multivariate normal distribution, then the numerically constructed p.d.f. and c.d.f. should match the functions generated by equation (4) and equation (5), respectively. The question of how well one function matches another introduces the concept of goodness of fit. Two goodness of fit techniques are developed, one related to the p.d.f. and the other related to the c.d.f. The advantages of the goodness of fit techniques is that it is possible to establish the percentile level of the fit.

In 1900 Pearson introduced the following measure (ref. 6), which is large when the differences $(f_{oi} - f_{ci})$ are large,

$$\chi^2 = \sum_{i=1}^{K} \frac{(f_{oi} - f_{ci})^2}{f_{ci}} \tag{9}$$

where f_{oi} is the ith observed frequency of occurrence, f_{ci} is the ith expected or computed frequency, K is the number of measurements. It has been shown that χ^2 is a chi-squared variable with $k - 1$ degrees of freedom. Hence, one could compute the frequency distribution of Q and evaluate χ^2 for K intervals along the distribution and determine the percentile level from a table of percentiles of the chi-square distribution. In the case of an application to multispectral scanner data, the number of independent channels would determine the number of degrees of freedom to generate the p.d.f. given by equation (4), which is related to the computed frequency, f_c. Given the measurements the observed frequency could be constructed. Then K values along the χ^2 axis could be chosen and equation (9) evaluated. The use of a table of percentiles of the chi-square distribution would determine the accuracy to which the data base follows the chi-squared assumption.

A method of determining the goodness of fit based on the distribution function uses the Kolmogorov-Smirnov statistic. If $P(x)$ is the theoretically constructed cumulative distribution function and $P_0(x)$ is the numerically constructed cumulative distribution function then the Kolmogorov-Smirnov statistic is

$$D = \max_{\text{all } x} |P(x) - P_0(x)| \tag{10}$$

This situation is illustrated in figure 19. The value of D and the number of samples determined the accuracy to which $P_c(x)$ approximates $P(x)$. In the case of multispectral scanner data the number of independent channels determines the value of n to be used in equation (5). $P_c(x)$ would be generated from the experimental data. The value of equation (10) and the number of samples would be used as inputs to a table of acceptance limits for the Kolmogorov-Smirnov test of goodness of fit.

The effectiveness of the chi-squared statistic and the goodness of fit tests should be evaluated by using synthetic data. This is an effective procedure for checking out the implementation and power of the algorithms. Synthetic data which are representative of aircraft and spacecraft data should be generated and analyzed. This will provide insight into the applicability of the statistical tests for different data bases.

Actual remotely sensed data from aircraft and spacecraft should be processed to obtain better insight into the characteristics of the data base. The topics of field homogeneity, feature selection, pattern classification, and error analysis should be investigated in terms of the characteristics of the data base.

5.0 CONCLUSION

This internal note has demonstrated that, in order for the techniques which are examined to accurately cluster data, two conditions must be simultaneously satisfied. The first condition is that the data must have a particular structure, and the second is that the parameters chosen for the clustering algorithm must be correct. By examining the structure of the data from the C1 flight line, it is clear that there is no single set of parameters that can be used to accurately cluster all the different crops. The effectiveness of either a one-pass or iterative clustering algorithm to accurately cluster data representative of the C1 flight line is questionable. This means that, in order to use cluster analysis in its present form for applications like assisting in the definition of field boundaries and evaluating the homogeneity of a field, one must have extensive a priori knowledge.

Modifications to existing techniques, or entirely new techniques, are necessary for clustering to be a reliable tool for representative data sets. This involves the use of goodness of fit tests to determine, in a quantitative manner, a measure of the unimodality of a cluster. This also has applications to quantitatively evaluating the homogeneity of test and training fields.

TABLE I.- MEAN/STANDARD DEVIATION OF C1 AGRICULTURAL DATA

Class	Channel Number 1	2	3	4	5	6	7	8	9	10	11	12
SOY	84.46	79.76	61.06	62.32	85.78	87.92	63.90	84.19	70.55	81.87	91.57	73.28
	2.40	2.58	1.82	1.81	3.55	3.37	2.17	3.78	3.20	3.47	5.83	3.60
CORN 1	83.89	77.89	59.34	59.95	82.33	86.75	61.93	76.82	61.93	72.61	105.94	81.56
	3.02	3.00	1.78	1.83	3.10	2.49	1.82	3.18	2.97	3.06	7.37	4.58
CORN 2	75.90	71.53	56.07	56.48	75.55	81.21	59.51	72.36	59.10	71.66	114.78	88.73
	1.59	1.77	1.30	1.22	1.96	1.72	1.46	2.25	2.12	2.30	6.31	3.80
OATS	75.78	73.25	57.15	58.29	79.60	84.63	63.07	85.15	71.33	87.73	106.97	85.69
	3.00	2.88	1.93	1.94	4.01	3.85	2.39	4.35	4.34	4.21	7.00	4.97
WHEAT 1	72.16	71.40	57.47	59.64	80.47	81.15	62.82	92.74	83.75	95.68	81.18	69.35
	2.82	3.00	1.98	2.03	3.91	3.85	2.68	5.23	5.26	7.33	7.27	5.55
WHEAT 2	80.42	80.86	64.42	67.35	99.32	102.06	76.86	116.76	100.79	116.27	98.31	77.83
	2.69	4.50	3.27	3.85	8.51	9.07	6.46	11.79	9.64	12.76	10.97	7.36
RED CLOV	72.34	69.38	54.00	54.18	72.19	79.96	57.60	68.92	55.35	76.58	140.02	107.96
	2.02	2.06	1.31	1.44	2.77	3.18	1.90	2.79	2.40	3.37	14.54	8.87
RED 2	71.53	68.36	53.24	53.71	71.28	78.54	56.92	67.68	54.17	71.15	121.77	94.59
	1.57	1.66	1.31	1.21	1.80	1.99	1.22	1.88	1.86	2.77	14.29	8.79
ALFALFA	76.84	71.68	55.27	55.47	75.13	84.59	60.08	69.64	54.42	77.97	154.21	114.00
	1.88	2.37	1.71	1.87	3.16	2.17	1.81	4.14	3.80	2.41	14.25	8.28
RYE	80.12	79.79	63.68	65.36	93.92	96.71	72.57	102.91	86.13	99.84	96.05	75.86
	2.19	2.56	1.83	1.78	2.95	2.37	1.93	3.80	3.90	3.86	4.18	2.48
SOIL	90.05	85.31	65.42	66.84	92.52	89.47	67.16	95.12	83.46	90.24	71.81	60.38
	1.80	1.82	1.35	1.17	1.78	1.50	1.24	1.82	1.47	1.90	2.79	1.76

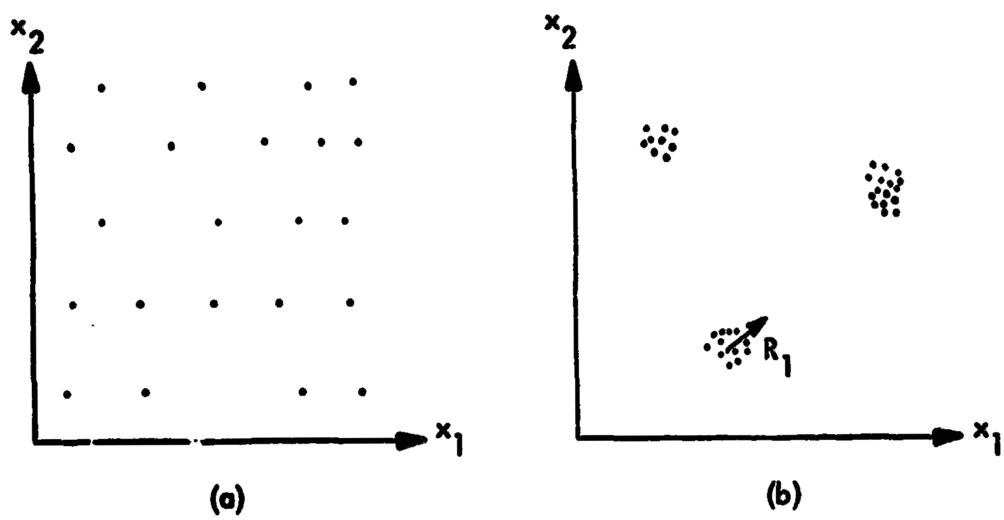


Figure 1.- Uniform and grouped data.

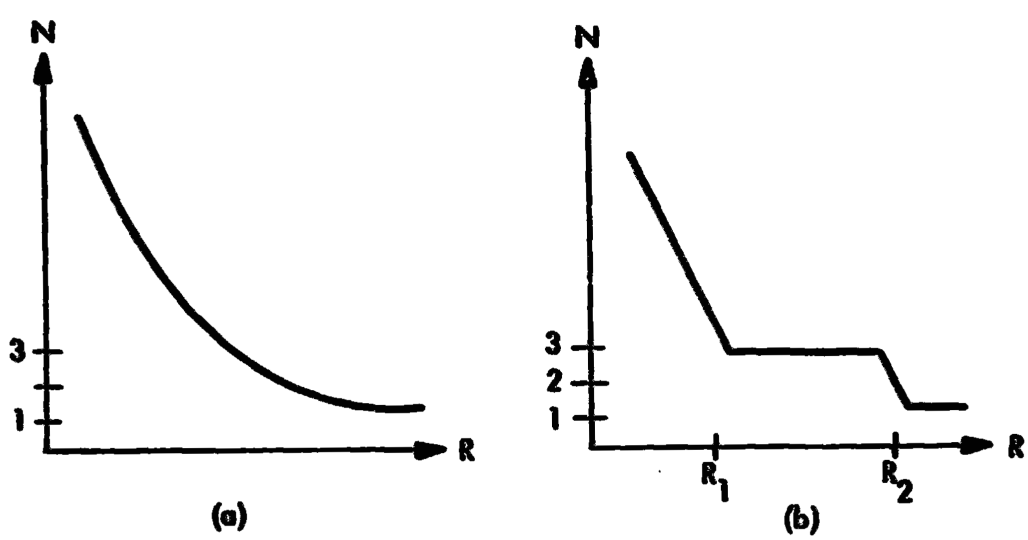
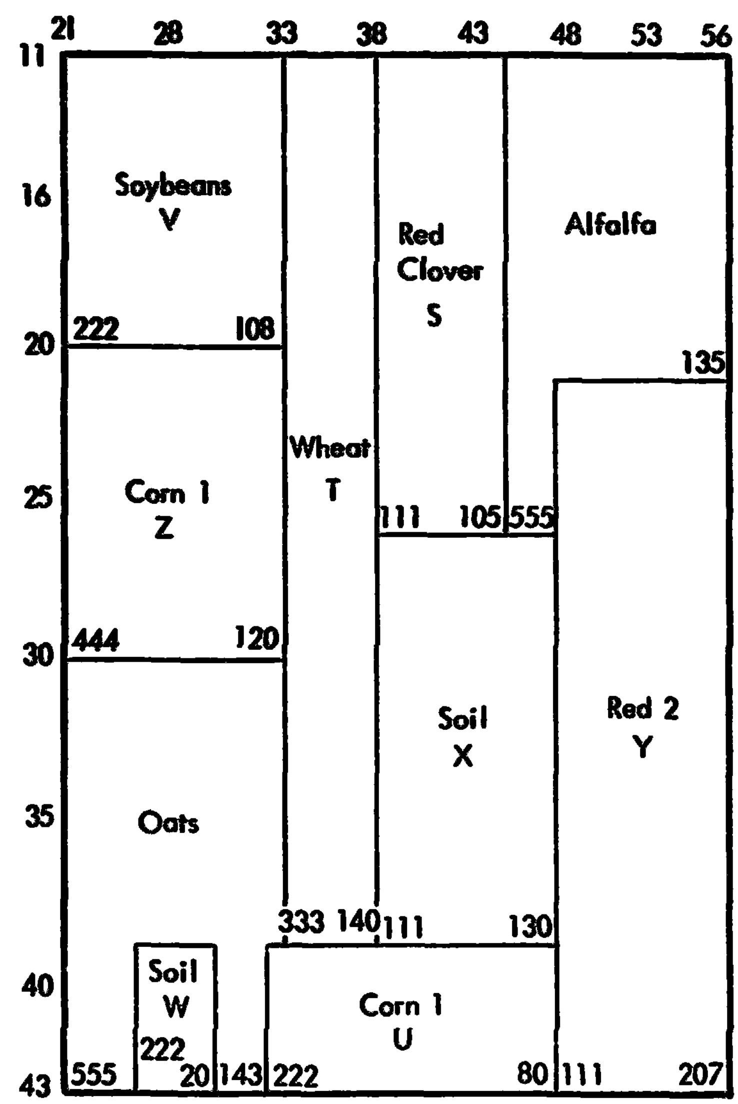


Figure 2.- N versus R for the data of figure 1.



For each field:

Letter in middle corresponds to FIELD ID
Number in lower left corresponds to Class ID
Number in lower right corresponds to number of pixels
The agricultural crop is used in simulating C 1 data.

Figure 3.- Data image (fig. 1 of ref. 2).

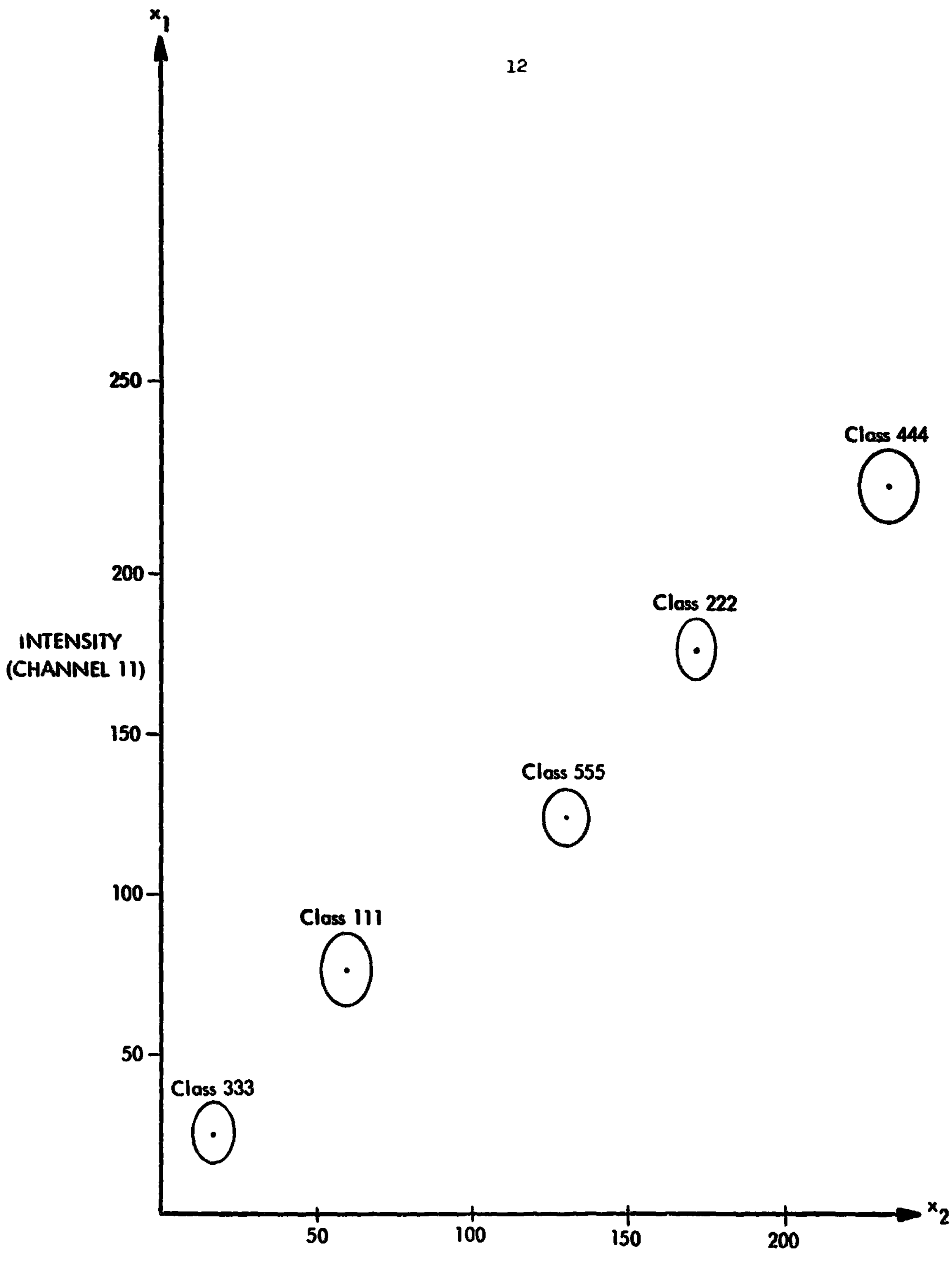


Figure 4.- Simulated data from the SERID program.

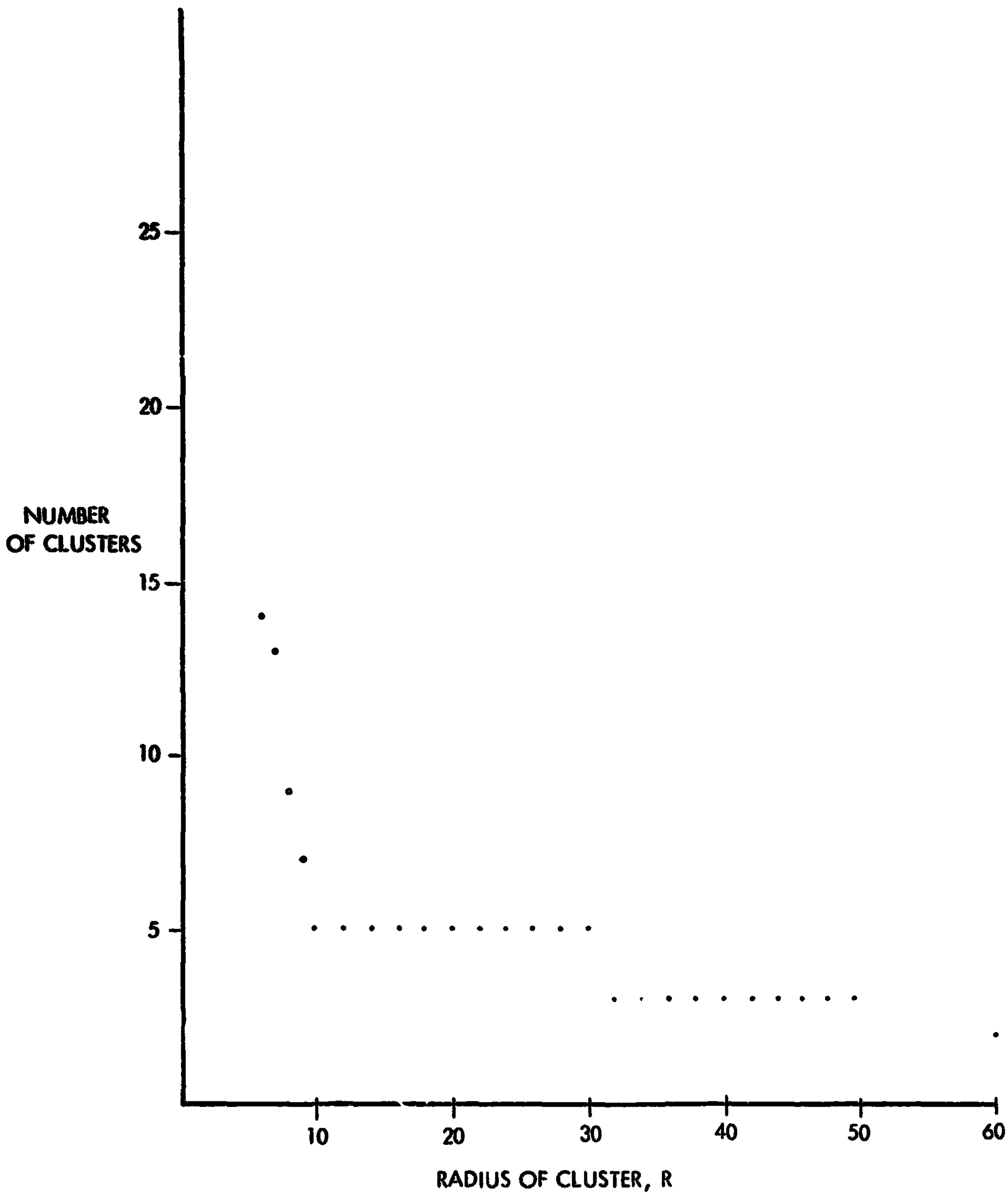


Figure 5.- Number of clusters versus R for the ideal case.

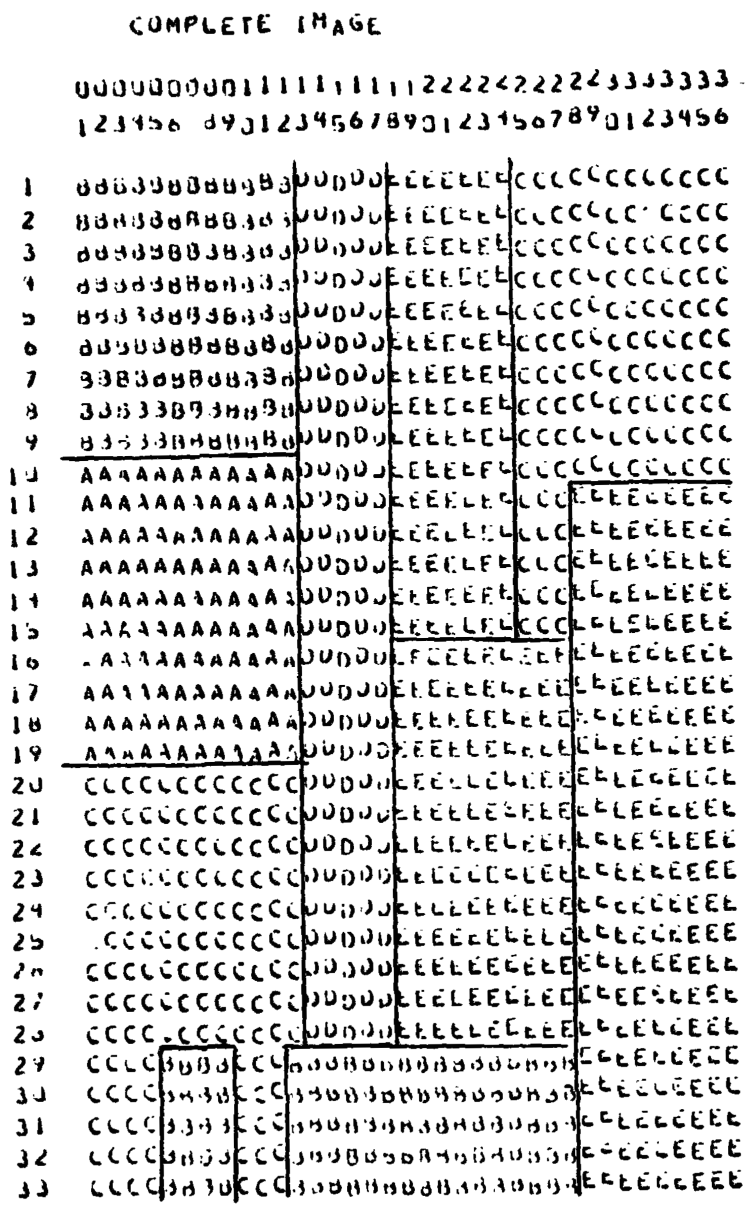


Figure 6.- Image map from the noniterative clustering technique for the ideal case, R = 20.

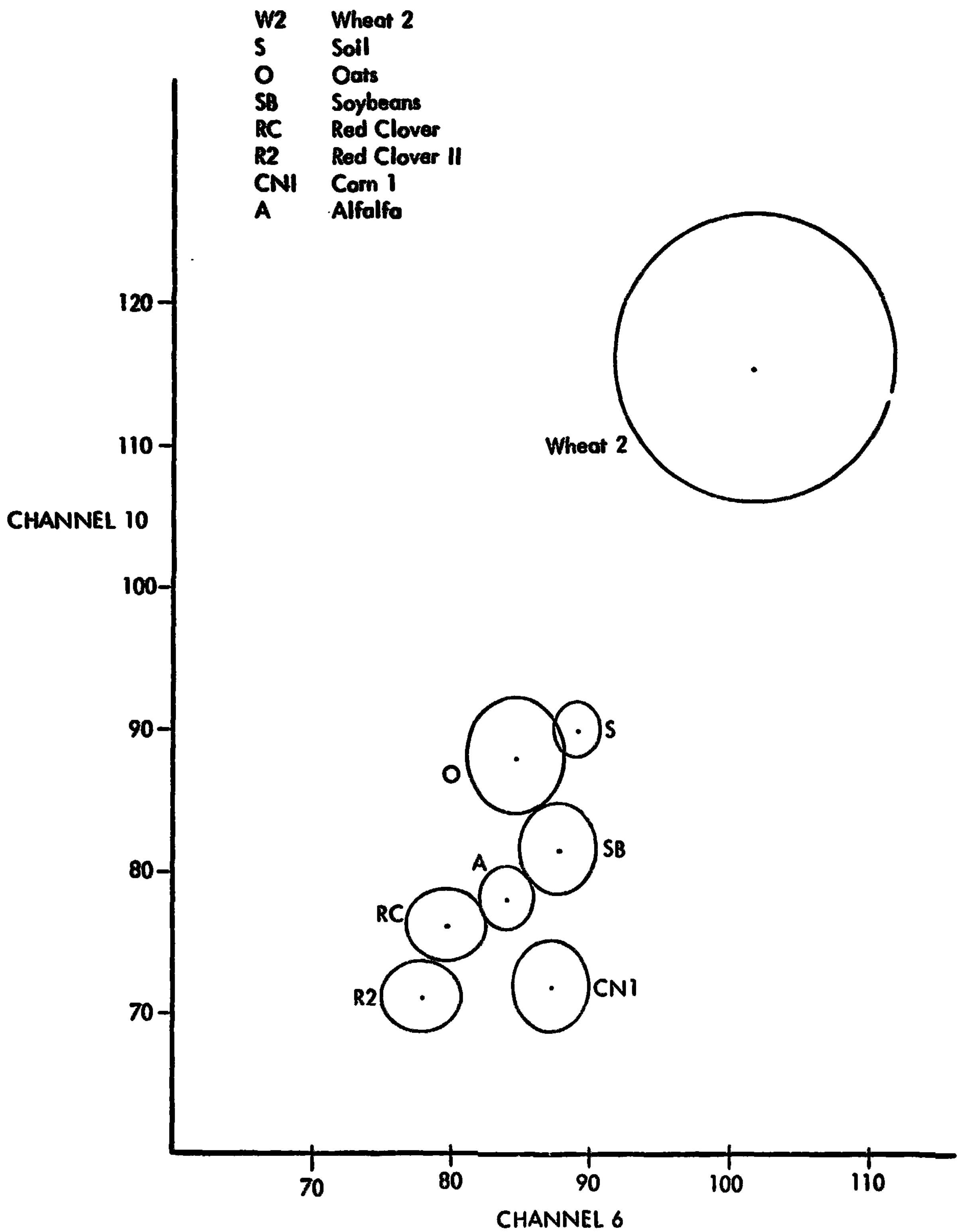


Figure 7.- Intensity of channels 6 and 10 for C1 data.

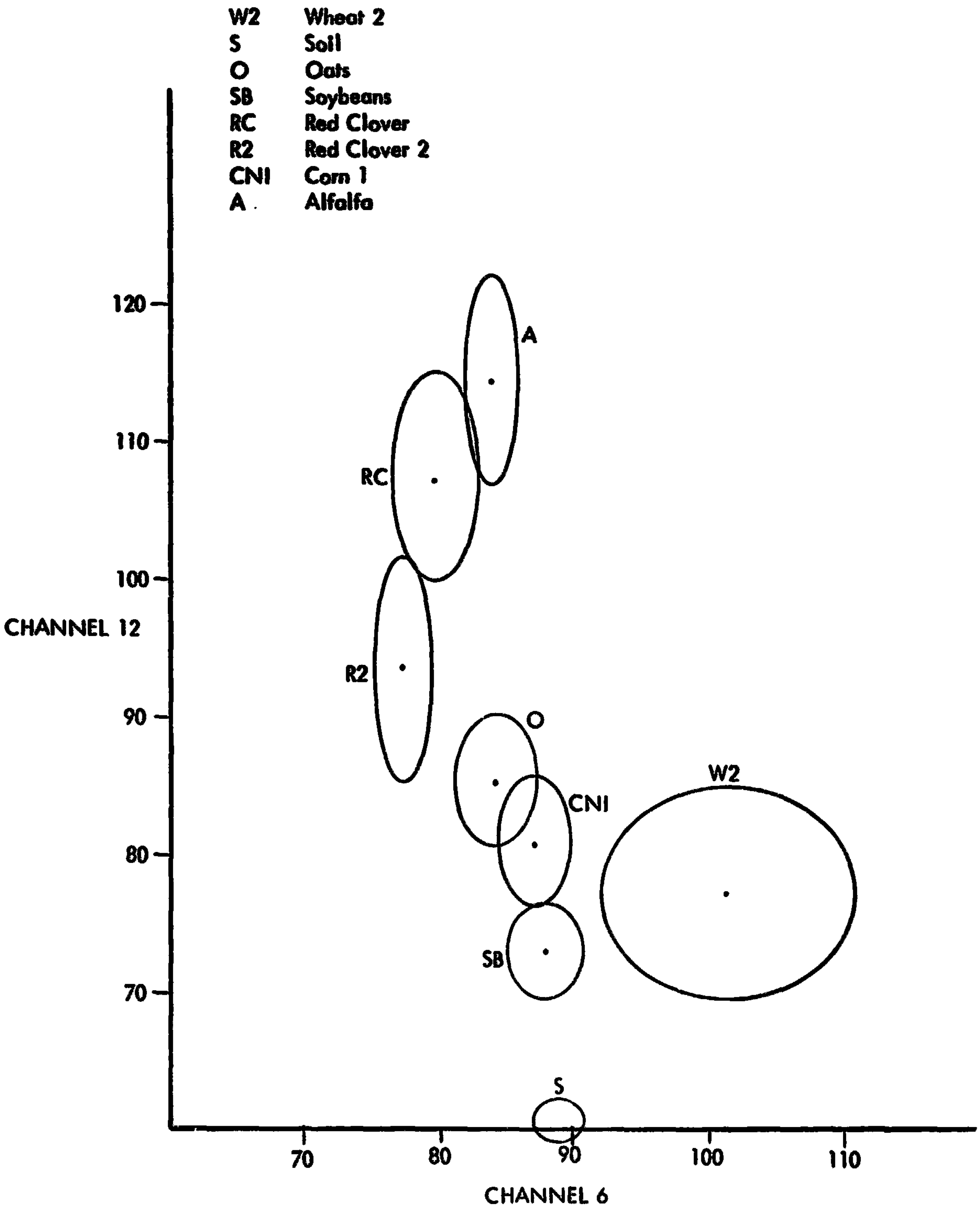


Figure 8.- Intensity - channels 6 and 12 for C1 data.

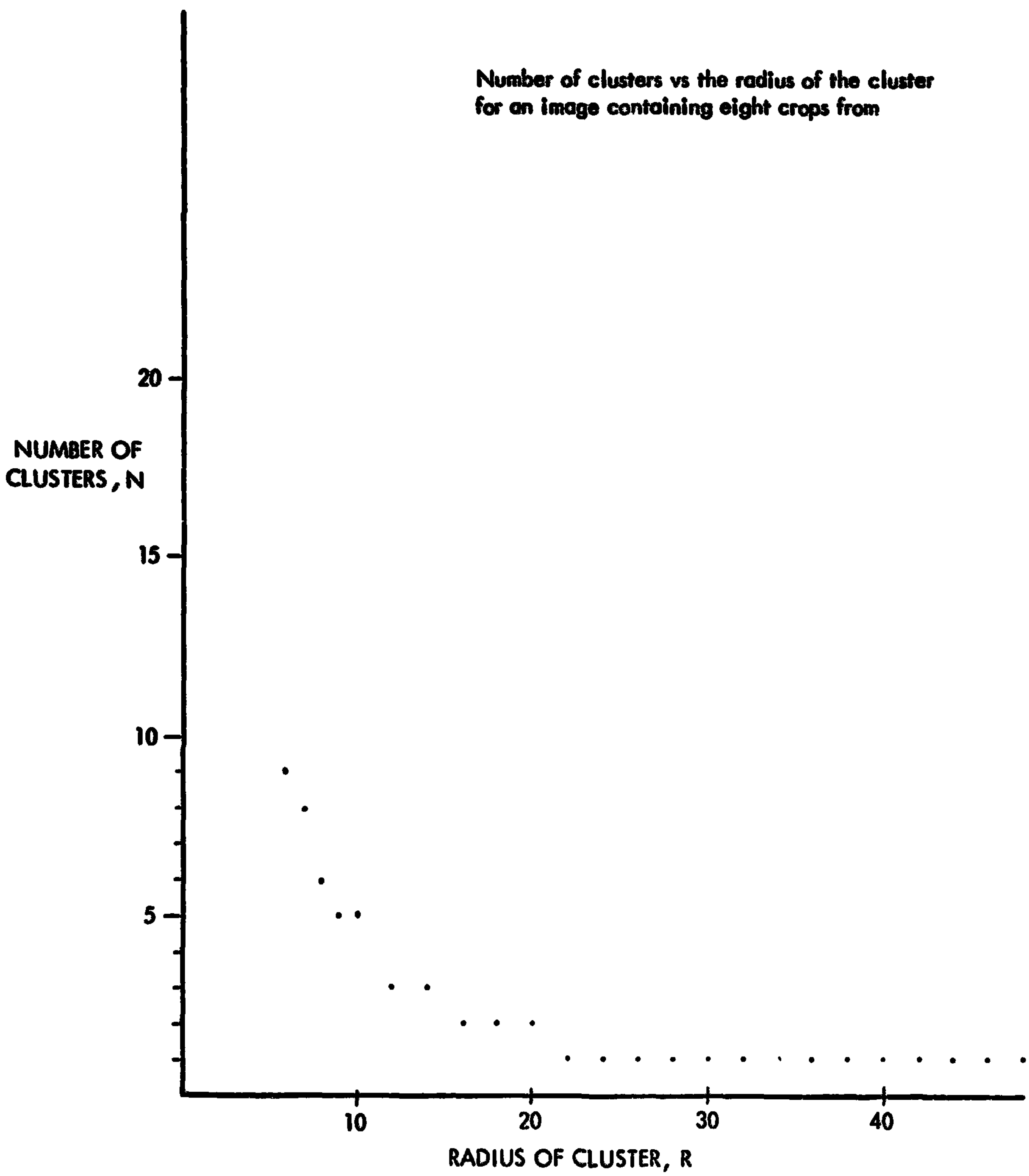


Figure 9.- N versus R for simulated C1 data.

```
            COMPLETE IMAGE

            000000000111111111122222222223333333
            123456789012345678901234567890123456

         1  BBBBBBBBBCBBBMGAAAAIAIIANNNANINANNNNN
         2  BCBBBBBBBCBBBAMAAMAIAIIINIIINAINNIANA
         3  CBBBBBBBBBBBBHLAAJAIAIIIAANINAININNNN
         4  BBBCBBBBBBBBBKAAAAAIAAIAANNNNANAIANNI
         5  BBBBBCBBBBBBBLAMAAAIAAAAAIINAAIIINANN
         6  BBBBBBBBBBBBBMAQAMAIIAIAANNAIAAINNAAI
         7  BBBEBBBBBBEBBMAOAAAIAAIAIINANAINAIINN
         8  CBBEBBBBBBBBBJAMAAIIAAIAIIINAAINNNNNN
         9  BBBBBHBBBBEBBHAAAAIAAAAANIIIAANNINNAA
        10  CCCCCCCCCCCCCAWAAAAIAAAAIHAAAAAAIIANI
        11  CCCCCCCCCCCCCNAAAAAIAIAANAISASSSNASI
        12  CCCCCCCCCCCCCAAAKUAAIIIAAAINSASSAAIAS
        13  CCCCCCCCCCCCCFAAALAIIAAAINIAAACASISAS
        14  CCCCCCCCCCCCCFAAAJIAAAAAIIAIAASASSIAS
        15  CCCCCCCCCCCDCMAAAAIIIIIAAIAISAIISASSA
        16  CCCCCCCCCCDCCMAAJAEEEEEEEEEESASISIASS
        17  CCCCCCBBCCCCCAAJMAEEEEEEEEEEAAACAIAIS
        18  CCCBCCCCCCCCCAAAAKEEEEEEEEEEAAASSSASS
        19  CCCCCCBCCCCCCAAAAHEEEEEEEEEEAASSSSSAA
        20  DDDDBDDDDDDDDAAAPAEEEEEEEEEEAASASAASA
        21  DDDDDDDDDDDDDFAAAAEEEEEEEEEESAAASASSA
        22  DDDDDDDDCFDDDMAAAAEEEEEEEEEESSASAACSI
        23  DDDDDBDDDDGDDAAAAAEEEEEEEEEEASASSSAAS
        24  DDDDDDDDDDDDDMAFAAEEEEEEEEEEAAAASIAIC
        25  CDDDDDDDDDDDCAAJMAEEEEEEEEEEASISSSCSS
        26  DDDDBBDDDDDDDAAALAEEEEEEEEEECSSSSSSAC
        27  DDDDDDDDDDDBDPAKAAEEEEEEEEEECAASSSSSC
        28  DDBDDDDDDDDDDJAKAAEEEEEEEEEEAASSSSSAS
        29  DDBBEEEEDDDCCCCACCCCCACCCCCCAASAAASAS
        30  DDDDEEEEDDDCCCAACCCACCCACCCCAIIAAASAA
        31  DDDDEEEEDDDCCAACCCCCCCCCCLACAISSSAASS
        32  DDDDEEEEDDDCCCACAACCCCCCCAAAASACSSSSS
        33  DDDDEEEEDDDCCACCACAACCCCCSACSSASSISST
```

Figure 10.- Image map from the noniterative clustering technique for R = 5.

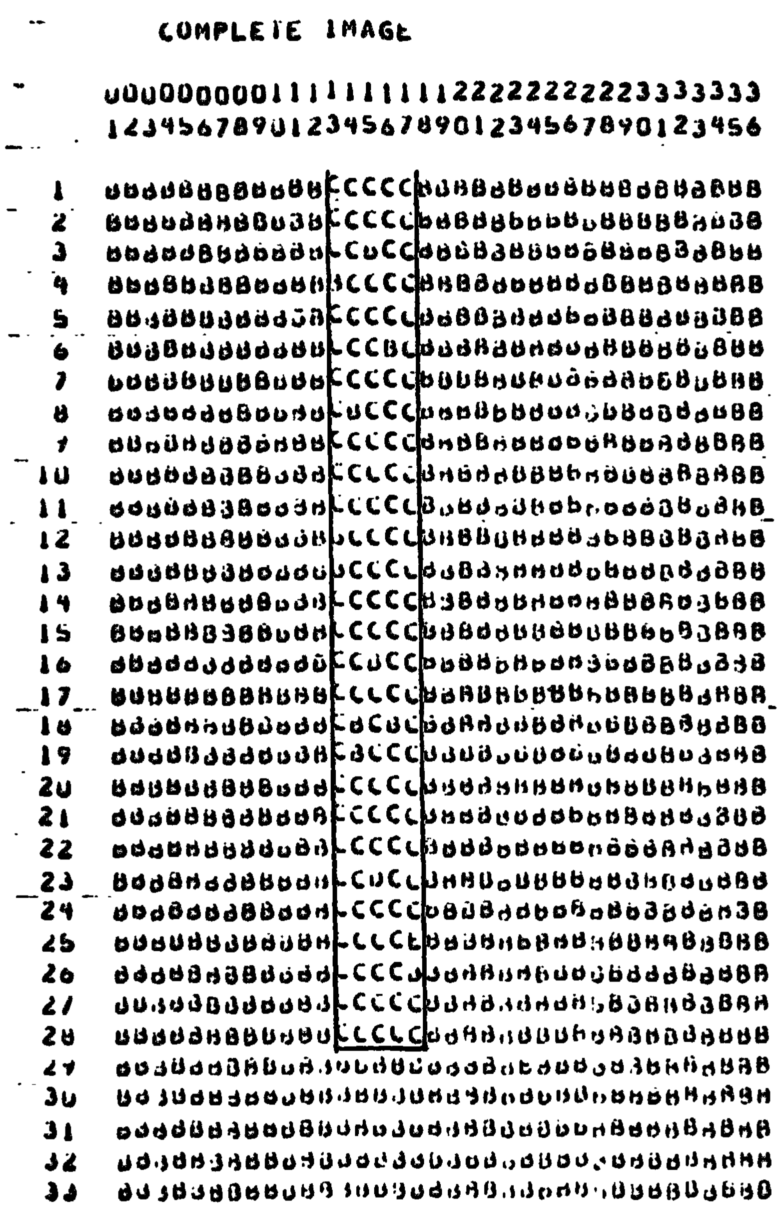


Figure 11.- Image map from the noniterative clustering technique for R = 18.

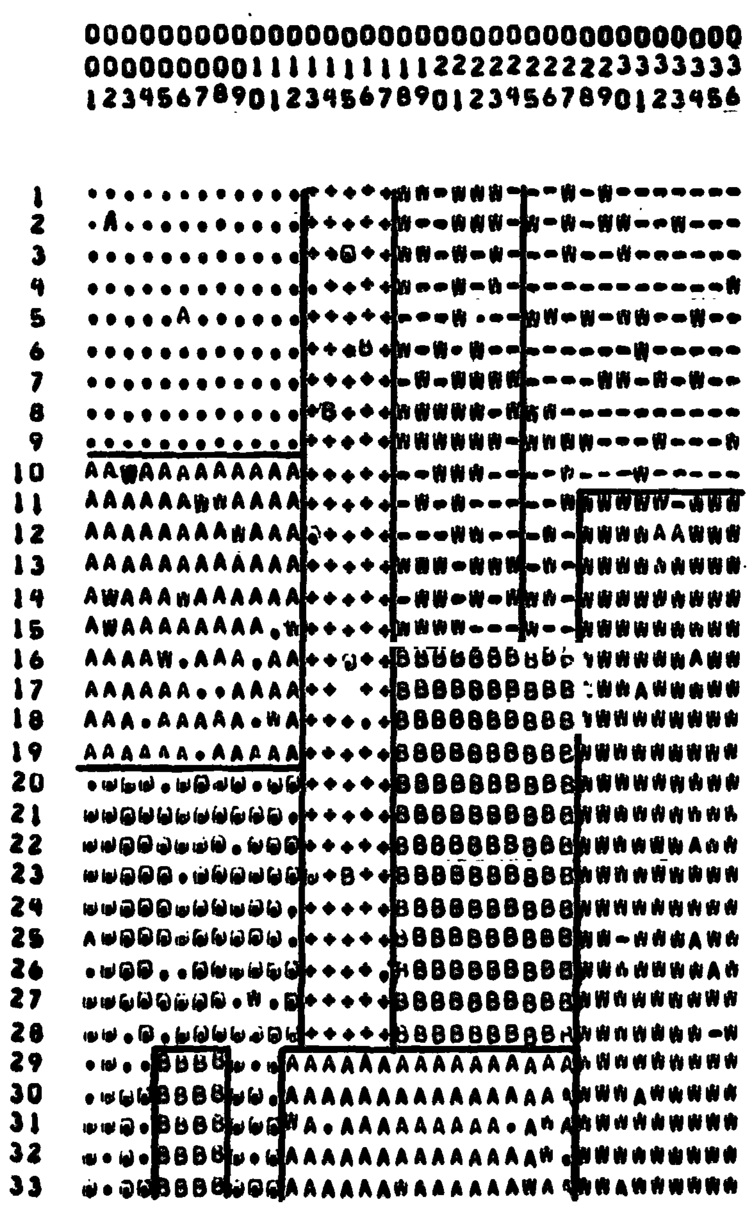


Figure 12.- ISODATA results (channels 6, 10, 12; NMIN = 30; no. of iterations = 20; with chaining.

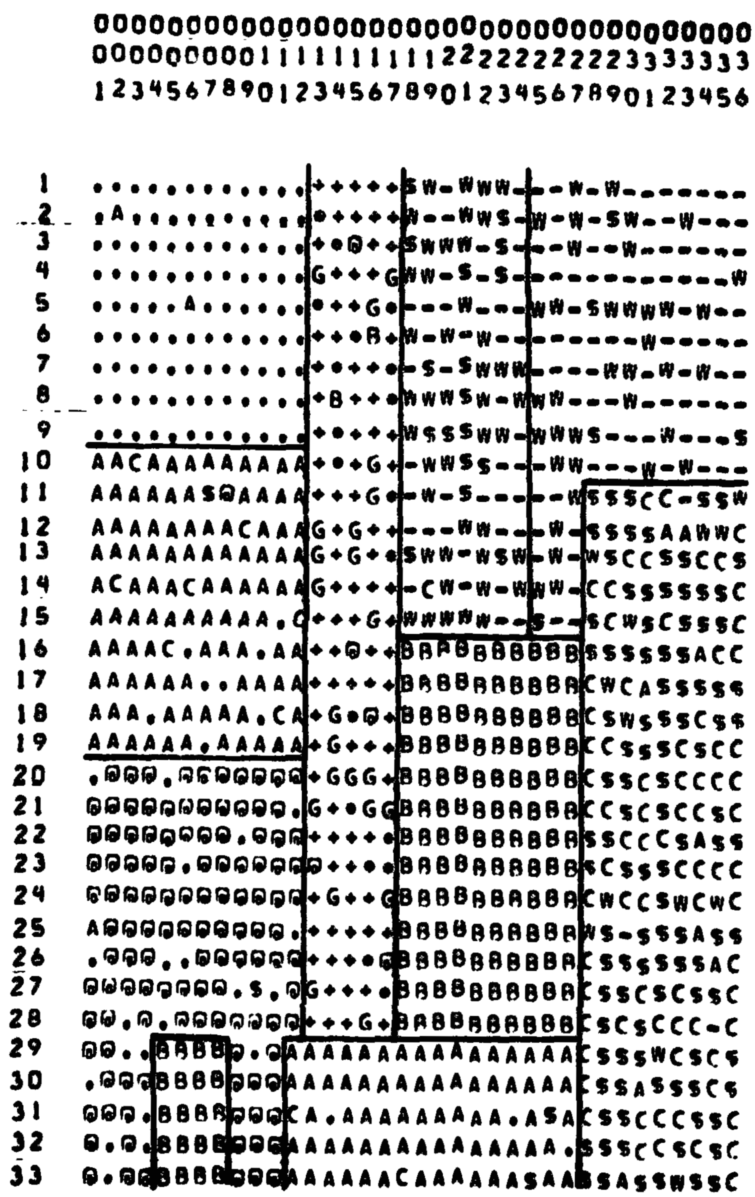


Figure 13.- ISODATA results (channels 6, 10, 12; NMIN = 15; no. of iterations = 20; with chaining

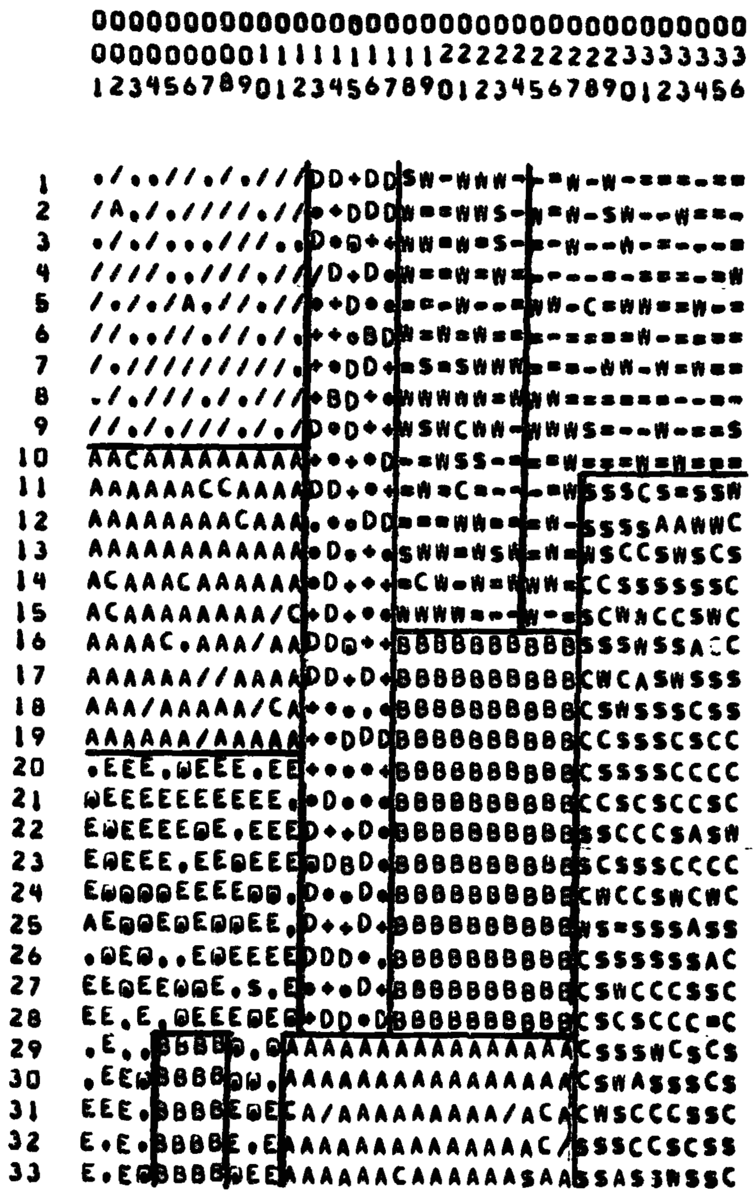


Figure 14.- ISODATA results (channels 6, 10, 12; NMIN = 30; no. of iterations = 20; no chaining

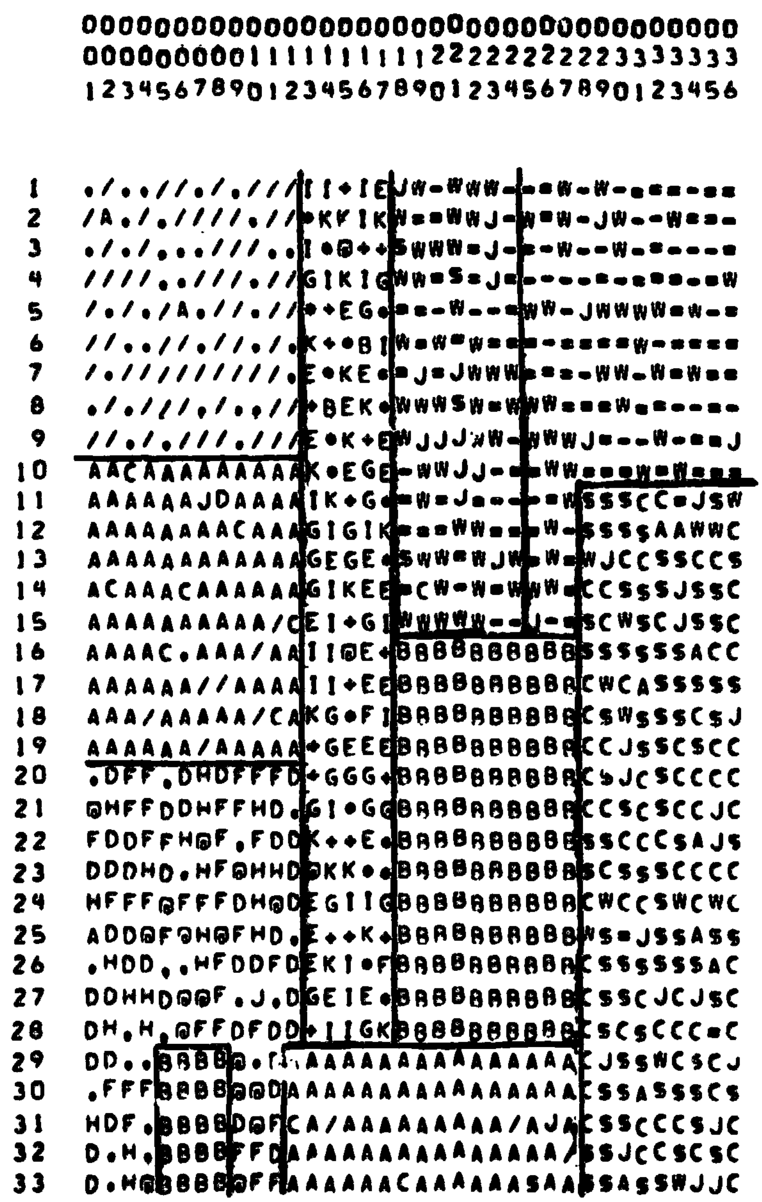


Figure 15.- ISODATA results (channels 6, 10, 12; NMIN = 15; no. of iterations = 20; no chaining

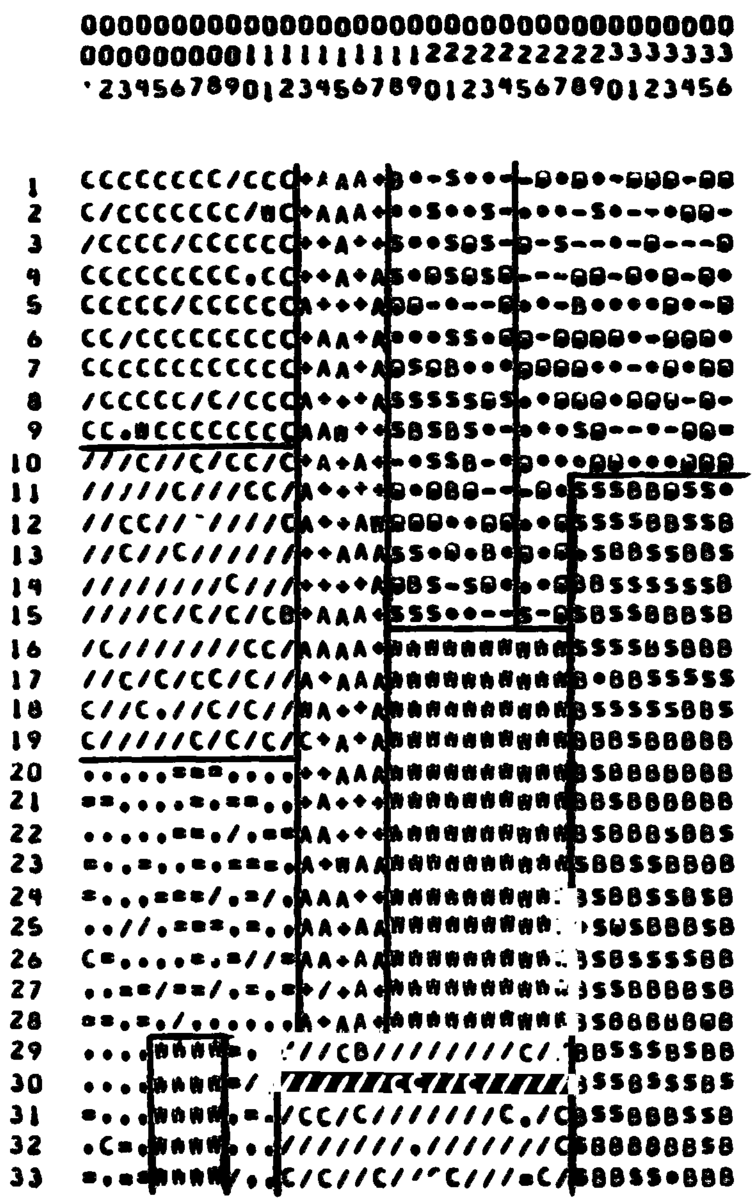


Figure 16.- ISODATA results char[illegible] , 12;
NMIN = 30; no. of iterations = [illegible] ing

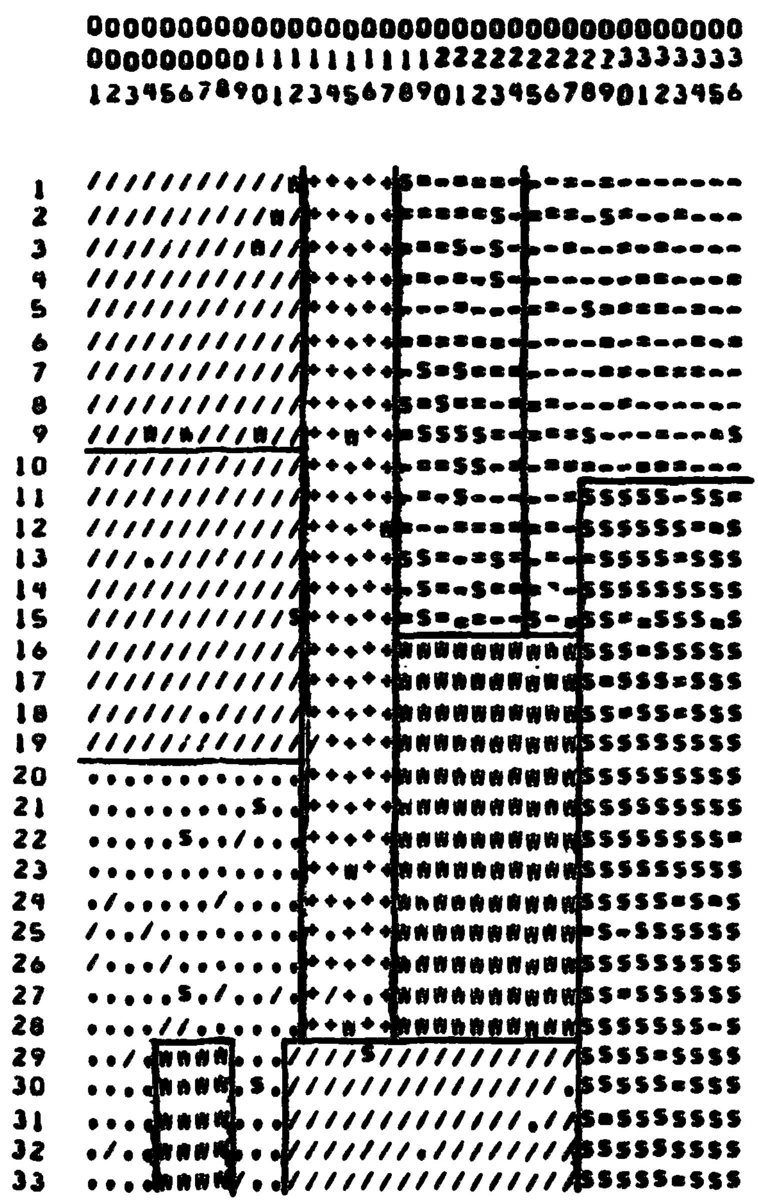


Figure 17.- ISODATA results (channels 1, 6, 9, 12; NMIN = 30; no. of iterations = 19; no chaining

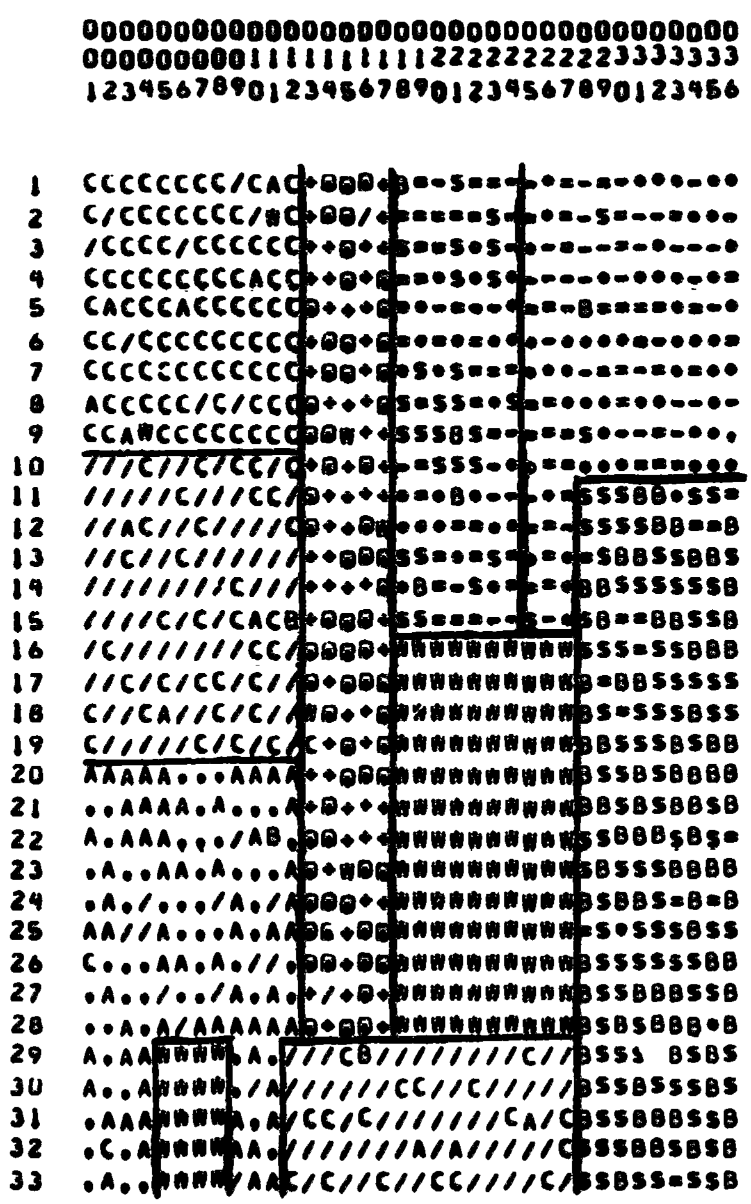


Figure 18.- ISODATA results (channels 1, 6, 9, 12; NMIN = 30; no. of iterations = 20; no chaining

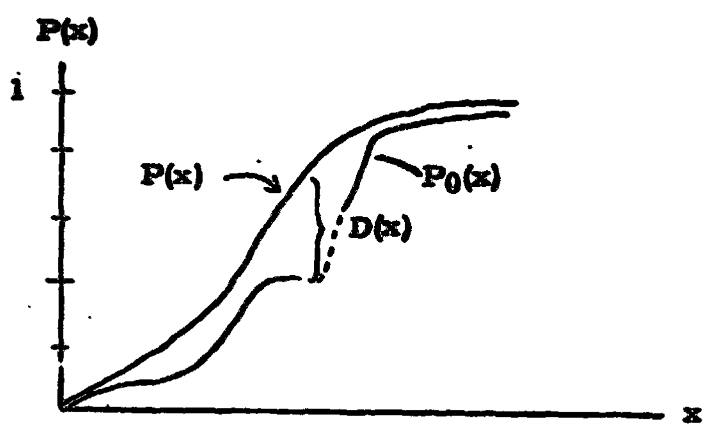


Figure 19.- Kolmogorov-Smirnov statistic.

REFERENCES

1. Detchmendy, D.; and Wylie, A.: RTCC Requirements for Implementation of a Clustering Program in the ERIPS. MSC IN 72-FM-268, Jan. 18, 1973.

2. Wylie, A. D.: Simulated Test Case for Verifying the SERID Program. MSC Memo FM8(72-154), June 21, 1972.

3. Detchmendy, D., et al: User's Guide and Software Documentation for the Algorithm Simulation Test and Evaluation Program (ASTEP). MSC IN 72-FM-198, Aug. 4, 1972

4. Kan, E. P. F.; and Holley, W. A.: More on Clustering Techniques with Final Recommendations on ISODATA. Lockheed Electronics Co., HASD, Houston, Texas, Technical Report 640-TR-112, May 1972.

5. Kan, E. P. F.: The Latest Version of ISODAT(A)/ISOC[illegible]. Lockheed Electronics Co., HASD, Houston, Texas, Technical Memo 63-0257-2115, Sept. 1972.

6. Lindgreen, B.: Statistical Theory. The MacMillan Co., New York, 1962.