A COMPARISON OF SAMPLING METHODS
FOR PERCENT COVER IN TUNDRA VEGETATION
and
A CLUSTER ANALYSIS OF SATELLITE (ERTS) DATA

A
THESIS

Presented to the Faculty of the
University of Alaska in Partial Fulfillment
of the Requirements
for the Degree of
MASTER OF SCIENCE

By
Richard Edward Furnas, B.A.
Fairbanks, Alaska
May 1975

A COMPARISON OF SAMPLING METHODS

FOR PERCENT COVER IN TUNDRA VEGETATION

and

A CLUSTER ANALYSIS OF SATELLITE (ERTS) DATA

RECOMMENDED:

Samuel J Harbo Jr

Barbara M. Lewis

Benita J. Wieland

S. F. MacLean Jr.
Chairman, Advisory Committee

James E. Morrow
Head, Department of Biological Sciences

APPROVED:

George C West
Dean of the College of Biological Sciences and Renewable Resources

5/14/75
Date

Provost

Date

ABSTRACT

I. A Comparison of Sampling Methods for Percent Cover in Tundra Vegetation

Line intercept and several random point techniques are compared. Random point techniques are found preferable to line intercept in providing error estimates and speed of execution. Several lines with random points selected from them is a good compromise between a simple random sample of an area and the speed possible by restricting attention to a line. A technique is given for sequentially drawing a simple random sample from a line.

II. A Cluster Analysis of Satellite (ERTS) Data

Data from the Earth Resources Technology Satellite (ERTS) are examined for a "natural" classification. Classification criterion involves commonness of classified spectral signatures and rarity of intermediates. Such a "natural" classification does seem to exist. The method makes no assumptions about underlying distribution of data. It involves order N calculations where N is the number of recognized different spectral signatures.

ACKNOWLEDGMENTS

Many people contributed materially and spiritually to my thesis. First and foremost I would like to thank Tom Wetmore, programmer at the Geophysical Institute. He wrote the computer programs for the analyses of the ERTS data and without his assistance that portion of my thesis could have been little more than a discussion of techniques and conjectures as to how they might work.

The success of my field work was greatly enhanced by Fairbanks-based moral and physical support from Jeff Kline and field assistance from my brother, Bill Furnas.

For his advisory assistance and as an editorial adversary I thank Steve MacLean.

I gained statistical insight and criticism from discussions with Barb Williams, Sam Harbo, and Pat Fox.

I thank Bill Cibula for his communications regarding related work that he has done with ERTS-type data, and Jim Anderson for his thoughts, the loan of some reconstituted imagery, and a light table for making figures.

My appreciation for problems in vegetation sampling was aided by discussions with Bonita Neiland and Sam Harbo. Chuck Racine helped me identify plant species.

My mathematical understanding and formalization of ideas was assisted by discussions with Barb Lando, Kathy Morgan, and Susan Royer.

Material support in the form of an HP-65 pocket computer-calculator was provided by Molecular Data Corporation.

For their comments on the manuscript I would like to thank the members of my committee: Steve MacLean, Sam Harbo, Barb Lando, and Bonita Neiland; also Barb Williams, Kathy Morgan, and Dan Swift.

Funds were provided by the University of Alaska through the cost-of-education allowance associated with my National Science Foundation graduate fellowship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

GLOSSARY OF SYMBOLS

$x = (x_1, x_2, x_3, x_4)$	name of a point in the 4-space 256 x 256 x 256 x 64
$y = (y_1, y_2, y_3, y_4)$	name of a point in the 4-space 256 x 256 x 256 x 64
$\|x_i - y_i\|$	absolute value of the difference between x_i and y_i
$d(x,y)$	absolute value or taxi-cab distance between x and y
$D(p^{(J)}, p^{(K)})$	Matusita distance between $p^{(K)}$ and $p^{(J)}$, $p_i^{(K)}$ $i = 1$ to u set of probabilities in u categories of probability distribution $p^{(K)}$
$P_0 \ldots P_n$	points in the four-space 10 x 10 x 10 x 10
$\rho(P)$	pixel density at point P
c	cutoff value for cluster assignment
$\bar{\rho}(P)$	smoothed value of $\rho(P)$

INTRODUCTION

Data from satellites present a wealth of information about the earth's surface. To the ecologist it provides an opportunity to consider large scale patterns and processes in ecological communities. In order to usefully interpret these data, however, some system of classification of ground features and satellite data is required and the correspondence between them must be understood.

Philosophically there are two approaches to this problem. One is to use familiar surface features as "training sets," discover how the satellite imagery portrays them, and thereby establish a correspondence. In the context of vegetation analysis, classical phytosociological techniques may be used to identify and characterize the training sets. This may become extremely difficult to quantify, however, since quantitative techniques are generally not geared to cope with the scale of resolution (50 m) of the satellite. As a result, qualitative assessment is often made and a "representative subsample" (e.g., a smaller quadrat) is taken if any quantitative work is done at all. This approach has its greatest usefulness and ease of application in agricultural or low-diversity natural communities.

The second approach is to work from the satellite data themselves. Here some notion of classification is used to subdivide the data, either in a geographic context by looking for patches in a reconstituted image (much like an air photo) or, in a more abstract sense, by identifying those combinations of reflected light (spectral signatures) which occur frequently in an entire image. In working with the satellite data these classifications may then be superimposed on the imagery and the resulting map of an area used to direct subsequent ground truth efforts.

My investigation approaches the problem from this more abstract viewpoint. It searches for a "natural" way of classifying the satellite data themselves. My interests lie in the extent to which such a "natural" classification may exist, and the relationship between the classification units and those of a ground-based observer.

I have divided the thesis into two separate parts. The first part deals with the problem of large scale sampling. I have compared several sampling techniques for estimating percent cover in vegetation. I then evaluated their efficiency in terms of the precision, accuracy, and effort required to perform the sampling. From these observations I developed a sampling technique of reasonable precision and accuracy and requiring little effort to perform on the large scale demanded for satellite ground truth.

The second part confronts the problem of structure of the data from the Earth Resources Technology Satellite (ERTS). Here I examine several properties of the data relating to the frequency-of-occurrence of spectral signatures in a portion of an image. I develop a "natural" way of classifying the satellite data based on this frequency information. Ideally such a technique would be fast and require few assumptions about the underlying distribution of the frequency data to yield a classification. This I have tried to do. Since there are not known to be forces operating to rigidly preserve the identity of such classification units, I have not required unambiguous classification of all observed spectral signatures but rather develop an hierarchical classification of relatedness with some spectral signatures remaining unclassified at more refined levels of classification.

A basic assumption in the use of satellite imagery is that surface vegetation differences will have different spectral signatures. I examine a very simple form of this assumption using ground truth data collected using the sampling methodology developed in part one.

PART I.

A COMPARISON OF SAMPLING METHODS FOR PERCENT COVER IN TUNDRA VEGETATION

Introduction

The problems of sampling methodology are fundamental to ecological inquiry. Many sampling techniques have been developed, often addressing different aspects of the organisms under study, and often new approaches are introduced, which claim to be "better" techniques for measuring the same aspects of the organisms. In vegetation sampling, percent cover is often of interest and a number of techniques for estimating percent cover have been developed. These vary in precision (repeatability of measurements), accuracy (amount of bias in measurements), and the effort required to perform the sampling.

My interest in the problem arises from the use of cover estimates to provide ground truth for satellite imagery. In order to provide nearly uniform coverage of the ground on a scale comparable to the data units from the satellite, I needed an efficient sampling technique -- one which would be relatively accurate, precise, and reasonably quick in execution. In order to select such a sampling technique, I compared several promising ones, and chose the one which was the most efficient.

Site

The study was conducted on the Macomb plateau, a plateau north of the Alaska Range, south of the Alaska Highway between Delta and Tok, lat. 63°35'N, long. 144°45'W (Figure 1).

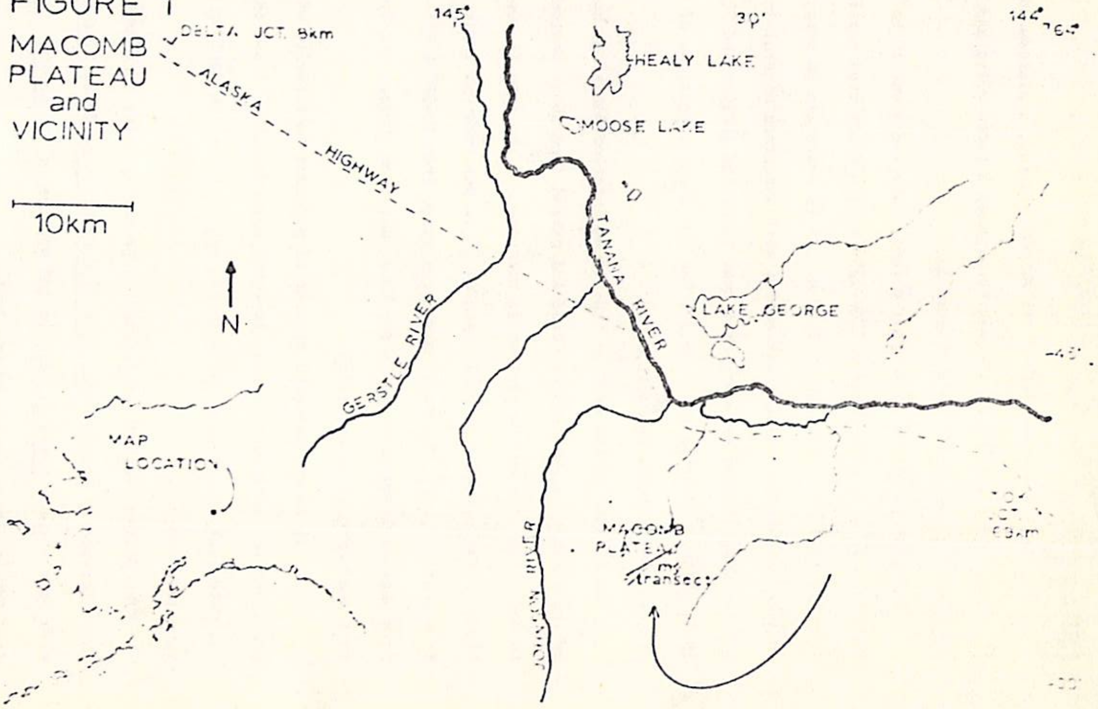
The plateau itself is roughly oblong, covering about 15 km^2. Elevation is primarily between 1200-1300 m with restricted portions dropping to 1100 m or reaching 1500 m. It is underlain by metamorphic rock related to the Birch Creek schist with a surface mineral layer of moraine deposits and deep organics near the lakes (Holmes and Foster 1968). The vegetation mat over all but the highest portions of the plateau is virtually complete.

The climate of the area is typical of interior Alaska as modified by the elevation.. Winters are cold and extend from about September to May. Snowfall on the plateau is moderate and the snow is much windblown. Precipitation probably exceeds 30 cm per year but varies greatly from year to year. Summer temperatures range from 10-30° C with light rain showers frequent and of a few hours duration (Holmes and Foster 1968 and personal observation).

The plateau was selected because of previous familiarity with the area, and the relatively simple logistic arrangements. Also, the plateau is quite flat, minimizing variable sun angle effects in dealing with the satellite imagery.

The plateau vegetation is tundra, primarily moist tussock tundra characterized by Betula nana, Vaccinium uliginosum, Carex bigelowii, with some shrubby Salix planifolia and patches of Eriophorum vaginatum (all species names follow Hulten 1968).

FIGURE 1 MACOMB PLATEAU and VICINITY



Species of wet tundra areas were primarily *Carex aquatilis*, *Eriophorum angustifolium*, and *E. scheuchzeri*. There were also a few areas of *Dryas* coarse sand fell-field vegetation, while creeks and intermittent drainages were bordered by tall (2 m) shrubby willow growth. Along the southern edge of the plateau rise mountains of the Alaska Range with bare rock and lichen-crusted rubble slopes as well as tarn lakes and small glaciers. To the southwest and west are the Johnson Glacier and Johnson River, a glacial meltwater river with markedly braided channel. To the north is the precominantly spruce forest of the Tanana River valley. To the east is more upland tundra vegetation. The plateau itself has several small lakes or ponds on it. Two squares, each 100 ft (30.48 m) on a side, were selected for sampling. They were located in two different vegetation patterns. One, which was sampled most intensively, was located in a *Carex bigelowii* tussock area which was striped by drainages about 10 m apart which contained *E. angustifolium* and a diminutive form of *C. aquatilis*. It was sampled the first part of July. The second square, sampled in mid-August, was in a *Dryas*-lichen-low willow area -- a drier site than the first, much better drained, and sandy.

Methods

Sampling in the two squares was done using several techniques which seemed promising in view of the intent to relate ground information to satellite data. The squares were treated as populations from which various types of samples were drawn. The different sampling techniques were all variations on point sampling. A point was located by some criterion and the species of the uppermost plant was recorded, since it is the uppermost layer which is seen by a satellite. The notion of "species" for these sampling purposes was extended to include other surfaces such as dry, dead (usually light brown) plant material; dark brown to black decaying organic material; water over decaying organics; water over sand; and sand.

The sampling comparisons may be subdivided into four parts in a 2x2 array. The relationships examined were (see also Table 1):

(I) line intercept sampling: methods of sampling which approximate line intercept's precision but which are less time-consuming, especially in finely divided vegetation.

(II) random point sampling from the entire square: methods which may be used to approximate random points' accuracy but which are less time-consuming or more suited to an oriented sampling (as along a transect).

Under each of these basic sampling styles were two considerations in estimation. The basic thought is that sampling may be done for two principal reasons:

A. Estimates of the importance of a single species in comparison with all other species which may be present.

B. Estimates of the spectrum of species which are present, and their relative importances.

These two classes of estimates are not necessarily inconsistent with each other. In terms of the efficiency (which includes accuracy, precision, and effort) of a sampling procedure, there may be differences.

I used two variations on line intercept sampling. My standard of

Table 1. Outline of percent cover estimate comparisons.

I. Line intercept sampling and two alternatives

0) Standard of reference -- line intercept

1) Intensive random points

2) Dominant sp on .1 ft

II. Simple random sample of points from the square and five alternatives

0) Standard of reference -- 1000 random points located using x-y coordinates

1) 1000 random points generated by random walk

2) 100 random points on the center line

3) 100 random points divided onto two lines

4) 100 random points normally distributed around the center line

5) Line intercept techniques from I (in the tussock square)

Under I and II above, I considered:

A. Individual species estimates

1) Overlap of confidence intervals with reference technique

2) sign test for bias

B. Estimates of total species-spectrum Matusita distance from reference technique

These comparisons may then be visualized in a 2x2 array:

$$\begin{bmatrix} \text{IA} & \text{IIA} \\ \text{IB} & \text{IIB} \end{bmatrix}$$

reference was a line intercept sampling with a resolution of intercept measurements of .01 ft (3 mm) along a 20 ft (6.1 m) segment. As one alternative technique, I also sampled 200 randomly located points on the same segment. The points constituted a simple random sample of the line and the precision of locating points was .01 ft (3 mm). The other alternative technique recorded the dominant species on all 200 of the .1 ft (3 cm) intervals along the 20 ft (6.1 m) segment. I sampled five such segments, each 20 ft (6.1 m) long lying in parallel lines with random starting points. This sampling was done in the tussock square. The line intercept references proved so time-consuming that this set of comparisons was not attempted in the *Dryas* square.

As alternatives to truly random sampling from the square, I used four sampling techniques. The standard of reference was 1000 points randomly located by cartesian coordinates drawn from a random number table. On the tussock square I have the line intercept data as another comparison. The first technique (theoretically a good approximation to random point sampling) was a random walk generated by starting at a random location in the square, spinning a soda-straw spinner, and taking five steps in the direction pointed by the spinner. I reflected off the flagged boundary of the square when I encountered it (five steps was about 10 percent of the square's diagonal, making successive points well distributed around the square). A wire with a needle on it was then lowered and the first species touched was recorded.

I also wanted a method which would be readily adapted to the transect style of sampling. With this in mind, I sampled from the squares using 100 randomly located points normally distributed around the center line of the square, i.e., using cartesian coordinates in the square, selecting 100 uniformly distributed random numbers for x-coordinates and 100 normally distributed random numbers for the y-coordinates. The mean was 50 ft (15.24 m) and variance was 16.7 ft (5.09 m) to put all (~99%) of the points in the square. This would be adaptable to a transect sampling scheme and would sample in a way analogous to the reflectance sampling of the satellite.

Since locating points using ordered pairs is time-consuming, I also tried two techniques related to the line intercept on a coarser scale. One was a simple random sample of 100 points along the center line. The other was a simple random sample of 100 points along a pair of lines 16.7 ft (5.09 m) either side of the center line (one standard deviation of the normally distributed random numbers). These 100 random points were randomly partitioned between the two lines.

Confronted with the prospect of much work with random numbers along line segments, I developed a method of sequentially taking a simple random sample of points along a line segment. It involves transforming uniformly distributed random numbers into exponentially distributed random numbers which may be taken as random intervals to be laid end to end along the segment. The exponential distribution of segment lengths insures uniform distribution of the resulting endpoints (see Appendix 1).

A method of data recording that I found particularly well suited to these point-type sampling techniques was to code species names in a two-letter code (e.g., initials of genus and species) and keep records in a cross-section book. In this fashion the squares of the cross-section ruling kept tally of the number of points sampled and permitted fairly rapid tallying of species occurrences onto a master list at a later time.

Interpretation of the data under sampling objective (A) above was done by considering all the techniques to generate binomial variates with the species of interest achieving Np observations where N is the total number of sample points and p is the probability of occurrence of the species of interest (i.e., fraction of total cover). This assumption is valid for considering the entire square as the population in the case of the 1000 randomly located points and for considering the line as the population for the 200 random points on each line-intercept segment. It is only an approximation for the other techniques. The approximation is poor for the .1 ft technique and for considering the line intercept segments as a sample of the entire square. Since most values of p are small, the usual normal approximation for large N is poor and confidence levels based on normal statistics for the observed values of p are misleading. In particular the lower limit often goes negative which is a conceptual impossibility. Binomial confidence limits are tabulated, however, with approximations for large N and interpolation formulae (Diem 1962). I then tallied the number of species for which the 95% confidence limits for the reference and alternative technique failed to overlap. This number was then compared with the number of times such events may be expected to occur by chance (binomial p = $(.05)^2$ = .025) (Table 2). As an indicator of bias, I also treated the cover estimates for each species as separate pairwise observations, comparing the reference technique with an alternative

Table 2. Number of occurrences n of non-overlapping 95% confidence intervals for single species comparisons. n is the number of species for which the 95% binomial confidence intervals for the standard of reference and alternative technique failed to overlap.

I. Line-intercept	Line Number				
Comparisons	(1)	(2)	(3)	(4)	(5)
Intensive random pts	n=0	n=0	n=0	n=2	n=0
Dominant sp on .1 ft	n=0	n=0	n=0	n=0	n=1
2.5% sig. level	n≤2	n≤2	n≤2	n≤2	n≤2

II. Random Points:		
Entire Square	Tussock Square	Dryas Square
Random walk	n=0	n=1
Points on center line	n=2	n=3
Points on two lines	n=0	n=1
Points normally dist.	n=0	n=2
Line intercept	n=10	
Intensive random pts	n=4	
Dominant sp on .1 ft	n=6	
2.5% sig. level	n≤4	n≤5

technique using the sign test (Table 3) (Ostle 1963).

For sampling objective (B), I found no completely satisfactory test. Complications arose due to differing sample sizes and the frequency of null observations in either my reference (1000 random points) or alternative sampling techniques. Also, the fact that my reference was only an estimate and not a true hypothetical distribution made certain tests unacceptable. These difficulties render the usual tests such as chi-square, log-likelihood ratio ("G") and even divergence, a powerful measure related to "G" and used for pattern recognition in communications theory, either undefined or undesirable since they would effectively ignore segments of the data. I did find a test (Walsh 1962) which I used here as an index of the success of the sampling procedures in displaying a spectrum of species. It is called the Matusita distance function defined as:

$$D(p^{(J)},p^{(0)}) = (\sum_{i=1}^{u} (\sqrt{p_i^{(J)}} - \sqrt{p_i^{(0)}})^2)^{1/2}$$

where $p_i^{(J)}$ is the fraction of occurrences devoted to species i using sampling technique J; $p_i^{(0)}$ is the reference sampling technique; and u is the total number of species. It is a distance measure for fractional data. As such, it measures the amount of difference between the species spectra given by the reference sampling technique $p^{(0)}$ and the alternative technique $p^{(J)}$. Note that this is an abstract distance measure in a u-dimensional space, each dimension corresponding to a species. The Matusita distance function may be seen as the Euclidean distance between the unit vectors $(\sqrt{p_1^{(J)}}, \sqrt{p_2^{(J)}}, . . . \sqrt{p_u^{(J)}})$ and $(\sqrt{p_1^{(0)}}, \sqrt{p_2^{(0)}}, . . . \sqrt{p_u^{(0)}})$. This measure has the desirable

Table 3. Tallies for the sign test for bias in species percent cover estimations: + = overestimate with respect to reference technique

- = underestimate

* differences significant at the 5% level are marked *

I. Line Intercept Comparisons

		Line Number				
	(1)	(2)	(3)	(4)	(5)	Total
Intensive random pts +	8	10	11	8	12	49
-	12	13	9	14	10	58
95% range	5-15	6-17	5-15	5-17	5-17	42-65
Dominant sp on .1 ft +	5	9	8	8	9	39 *
-	14	13	13	14	12	66
95% range	4-15	5-17	5-16	5-17	5-16	41-64

II. Random Points: Entire Square

		Tussock square	Dryas Square
Random walk	+	13	23
	-	13	29
95% range		7-19	18-34
Points on center line	+	8 *	15
	-	23	34
95% range		9-22	17-32
Points on two lines	+	8 *	15
	-	23	34
95% range		9-22	17-33

Table 3, continued.

		Tussock Square	Dryas Square
Points normally dist.	+	10	15
	-	21	35
95% range		9-22	17-33
Line intercept	+	15	
	-	21	
95% range		11-25	
Intensive random pts	+	11	
	-	19	
95% range		9-21	
Dominant sp on .1 ft	+	11	
	-	22	
95% range		10-23	

properties that it is well defined for all possible values of $p_i^{(J)}$ and $p_i^{(0)}$ and does not emphasize small values of p_i as much as the other statistics. The hypothesis H_o^J asserts that the reference sampling technique yields the same distribution as the alternative method J, against the alternative hypothesis that the distributions differ by a specifiable amount. The general test states: reject H_o if $D^2(p^{(0)}, p^{(J)}) \geq \frac{(u-1)B}{N}$ with significance level min ($\propto$, 1/B). B may be chosen so as to bound the significance level and $\propto = \left\{u^2 - 1 + \frac{1}{N}(2 - u - u^2 + \sum_{i=1}^{u} 1/p_i^{(0)})\right\} / ((u - 1)B)^2$. This however does not treat β (probability of type II error) which is of concern here since detection of possible differences is the point of interest. Consequently, I use it only as an index and can attach no β significance levels to differences in the values of the measure. Note that the range of possible values for the Matusita distance is zero to two. Note also that $\propto$ involves terms $1/p_i^{(0)}$ which is infinite when the reference technique fails to detect a species. Hence for a 95% significance level B must be 20, making the rejection level about .8 for 41 spp. and 1000 points.

Results and Discussion

Figures 2 and 3 contain the reference data from the 1000 random point sampling of the tussock and *Dryas* squares. As may be seen from the figures, both squares were "dominated" by dry dead vegetation. Dominance in the tussock square was quite pronounced (most of the dry dead vegetation was *Carex bigelowii*) while the *Dryas* square had a more equitable distribution of cover among the species present. Most of the

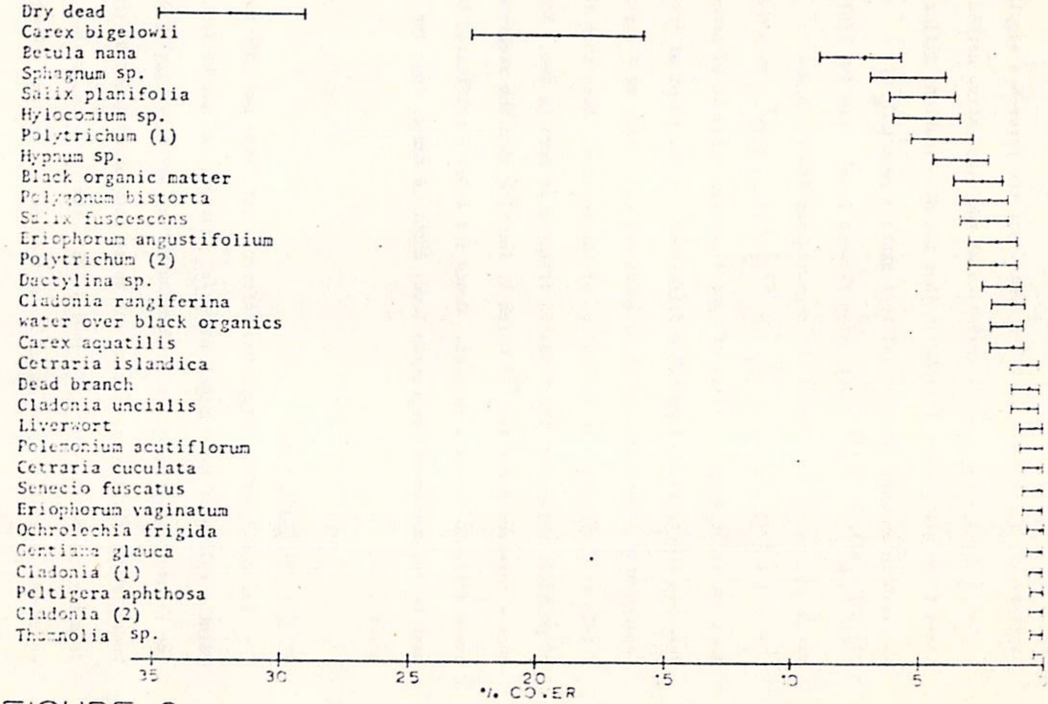


FIGURE 2 Tussock square: percent cover as given by 1,000 randomly located points (95% binomial confidence limits).

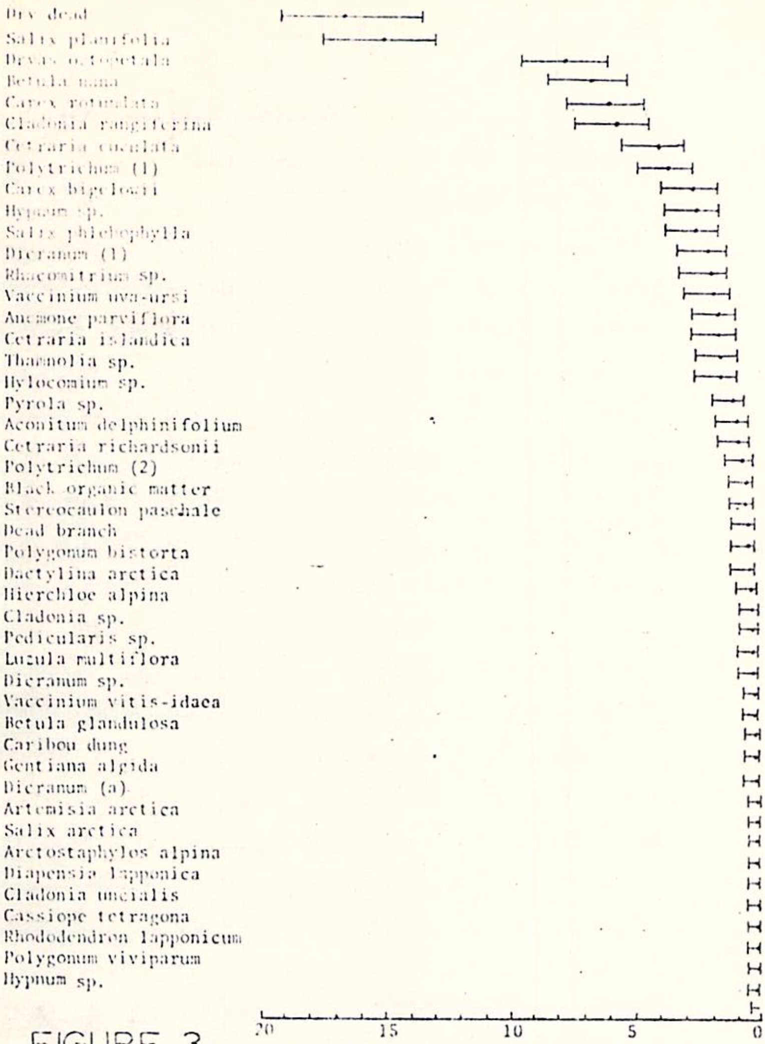


FIGURE 3 Dryas square: percent cover as given by 1,000 randomly located points (95% confidence).

plants in both squares were less than 20 cm high and birch and willows were scraggly. In the tussock square this meant that the line intercept technique was laborious due to the extent to which different plants interdigitated. The Dryas square was similarly interdigitated so the line-intercept comparisons were not done there.

I. Line intercept

A. Data from the single-species comparisons are summarized in Tables 2 and 3. From Table 2 it appears that assigning confidence limits based on the binomial assumption is reasonable when using either of the alternative sampling techniques as an estimate of the line intercept values. Both techniques were far less time-consuming to perform than line intercept itself. The .1 ft technique, however, does consistently under-represent the rarer species, giving rise to a significant bias when considering all five segments together. This makes sense since the rarer species would seldom dominate the .1 ft (3 cm) segment. The intensive random point technique does very well and yields good estimates which are not significantly biased. It does have the disadvantage of being somewhat slower in practice but has the advantages of being insensitive to patchiness and minor disturbances such as losing one's place or bumping the tape measure. Furthermore, the statistics of the sample, considering the line as the population, are precisely binomial.

B. Matusita distances suggest that the random point and .1 ft methods are closely comparable as estimators of species spectrum. The differences are small and the sampling here of five segments does not justify rejecting the hypothesis that the two methods are comparable on the basis of the sign test (Ostle 1963). Values of the Matusita distance function are graphed in Figure 4.

The line intercept method is a highly respected sampling technique (Greig-Smith 1964). Unfortunately, however, in this type of vegetation it is incredibly time-consuming, requiring many hours to complete just one 20 ft (6.1 m) segment. This may be contrasted with 20 to 30 minutes for the intensive random point method and 10 to 15 minutes for the .1 ft (3 cm) method.

II. Random points

Data for the single-species comparisons are summarized in Tables 2 and 3. I have also included the composite data from the line intercept comparisons. I have taken the 1000 randomly located points as the standard of reference for estimating actual cover values and compared others against it. On a species by species basis all techniques do reasonably well. In part this is due to the wider range of values in the tolerance regions for those estimators using fewer points.

The large number of point-equivalents (10000, since resolution was .01 ft (3 mm)) for the line-intercept method restricts its tolerance regions, making its estimates poorer than the others in terms of overlap of confidence intervals. It shows no significant bias, however.

As expected, the random walk technique of approximating uniformly distributed variates worked excellently. Virtually all its estimates, both of individual species and of species spectrum, are very close to those of the points located by cartesian coordinates. The simplicity and speed of the technique strongly recommended it over the use of cartesian coordinate methods which are more involved and slower. In a

Figure 4. Matusita distances between reference sampling techniques and alternatives. The upper two graphs are for the areal comparisons (reference technique: 1000 random points) and the lower one for the line intercept comparisons (reference technique: line intercept).

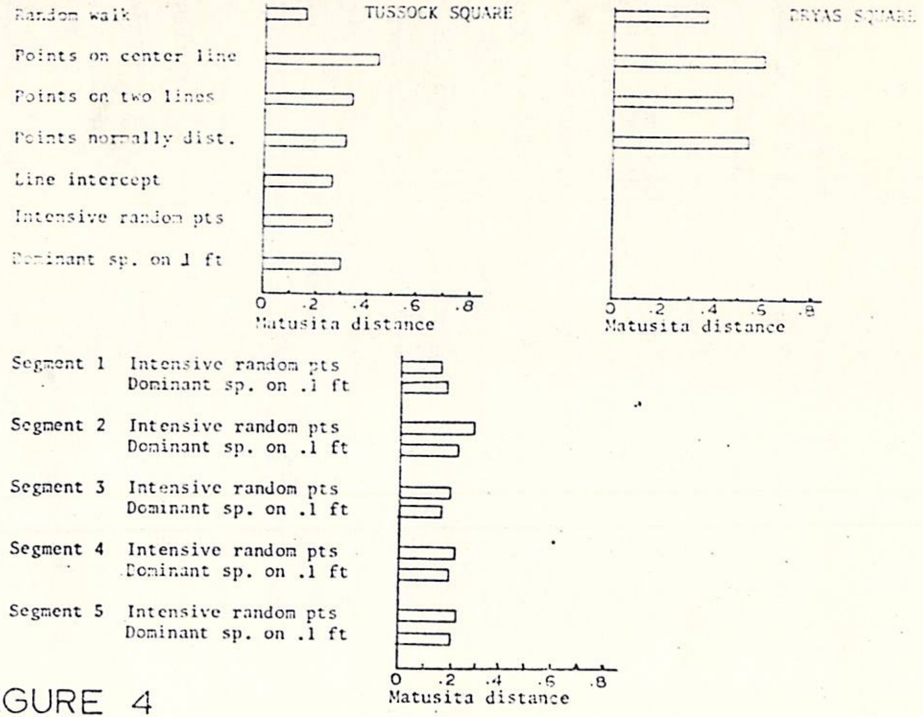

Random walk
Points on center line
Points on two lines
Points normally dist.
Line intercept
Intensive random pts
Dominant sp. on 1 ft
TUSSOCK SQUARE
0 .2 .4 .6 .8
Matusita distance
DRYAS SQUARE
0 .2 .4 .6 .8
Matusita distance
Segment 1 Intensive random pts
Dominant sp. on .1 ft
Segment 2 Intensive random pts
Dominant sp. on .1 ft
Segment 3 Intensive random pts
Dominant sp. on .1 ft
Segment 4 Intensive random pts
Dominant sp. on .1 ft
Segment 5 Intensive random pts
Dominant sp. on .1 ft
0 .2 .4 .6 .8
Matusita distance

FIGURE 4

larger area the number of paces between points would need to be increased. In such situations the savings possible from systematizing in the preparation of a cartesian coordinate method may eventually outweigh the simplicity, convenienve, and relative freedom of the random walk.

Often the interest is not in a particular area but in a region along some line. The range of interest on either side of the line may not be well defined yet an estimate may be possible. My sampling involved just such a problem. I did not know how my transect would cross the digital picture units of the satellite data; I knew the approximate dimensions of the picture units' coverage on the ground; and in order to have reasonably narrow tolerance ranges on individual species cover estimates, I wanted to have about 200 data points in each picture unit. One possibility was to consider points which were scattered around the transect of interest, much as a regression line has observed values scattered around it. This was the rationale for considering points normally distributed around the center line. For estimating cover in the squares, the method worked well, showing good overlap of confidence intervals with the reference of 1000 random points, and showing the species spectrum. It does show significant bias, however, but only due to a record of zero for species of low occurrence. This problem is inherent in lower intensity sampling. Accuracy is good for observed species. Since it is a cartesian coordinate method it was very time-consuming and preparation involved two kinds of random number tables.

At the other extreme is a collapsing of all the points onto the transect line itself. This technique eliminates the second type of random numbers and also the second tape measure used to achieve the second coordinate. To sample the same number of points it takes about one-sixth the time that using an x-y coordinate system does. Sampling along the single line, however, is more susceptible to idiosyncrasies of the vegetation under that particular segment and may give a distorted view of the vegetation. While the method gave reasonable estimates of individual species cover, it generally did not do so well as the other methods and did the poorest job of reflecting the species spectrum. It also, of course, shows the same bias that the normally distributed points did.

As a compromise between the fine performance of the normally distributed scatter of points around the center line and the use of random points on the center line itself, I also tried using two lines, each with half the sampling intensity of the center line technique. The improvement over the center line as the method of estimation was quite substantial and it took only slightly more time to perform, due to setup time for the tape measure. As with the other two techniques involving only 100 points, it does show a significant bias as a result of under-representing the rarer species.

As Grieg-Smith (1964) points out for line-intercept methods, several short lines are preferable to one long one. Similarly, with the point techniques, with sampling intensity on lines replacing line length, the limiting case (essentially one point per line) is that of the randomly located points.

Conclusions

(I) In the finely divided vegetation sampled here, the line intercept technique for cover estimation is not justified. It is too time-consuming and comparable estimates of both individual species cover values and the spectrum of species may be obtained in about 1/10 to 1/3 the time using either intensive random sampling of points from the lines or the dominant species over .1 ft intervals. The random point method is somewhat slower but it apt to be less sensitive to patchiness of cover.

(II) For areal sampling, accuracy of estimates increases with the dispersion of the sample while precision increases with sampling intensity. A random walk was an excellent approximation to random points located by cartesian coordinates. It is much faster and less demanding to execute. For sampling a transect, the trade off between the dispersion possible with a random scatter of points around the line and the reduced time involved in sampling on a single line may be met quite well by dividing the sampling effort into more than one line (here two). A simple random sample of points along a line may be taken sequentially by transforming uniformly distributed random variables into exponentially distributed random variables. A considerable saving in either preparation time or sampling time results.

PART II.

A CLUSTER ANALYSIS OF SATELLITE (ERTS) DATA

Introduction

Satellite imagery provides a superb opportunity to study large-scale pattern in vegetation. One of the biggest problems in such investigations is sampling, and the Earth Resources Technology Satellite (ERTS) does just that. The form of the sample is reflectance in four spectral bands. They are: Band 4 (530-570 nm), Band 5 (570-630 nm), Band 6 (640-680 nm), Band 7 (710-750 nm). From this basic set of data two assumptions are commonly made to relate the satellite information to ground features: (1) different reflectances imply corresponding differences in surface features; and (2) the set of observed reflectances (of bands taken singly or in concert) may be subdivided for purposes of classification.

My interests are: to what extent is assumption (1) true when the surface feature is vegetation? and, is there a pattern to the data which suggests a "natural" classification of vegetation analogous to that which exists with living organisms?

Site

Data from the satellite are subdivided in several ways. One form is the "scene" which corresponds roughly to a 160 km square on the earth's surface. The digital data for a scene is further subdivided into four vertical strips each of which is divided into four blocks. The portion of a scene I have examined in one of these blocks, 40 km on a side. The ground area it covers is an area including the Macomb Plateau. On the plateau I sampled along a transect running between two lakes visible in photographic reconstructions of the digital data (Figure 1). This sampling was done the last week in July and the first week in August. The plateau was chosen as the study area because of its relative accessibility, my previous familiarity with the area, and the minimization of sun angle effects since it is flat. The choice of the entire 40-km square containing the plateau was made to facilitate handling of the digital data. The scene chosen was the only cloud-free summertime image available, taken on August 21, 1972.

The vegetation of the plateau is primarily moist tussock tundra (Carex bigelowii, Betula nana, decumbent willows -- Salix spp. -- and patches of Eriophorum vaginatum), with some areas of wet tundra (Carex aquatilis, Eriophorum angustifolium), and Dryas-dry sand communities. Some of the drainages are overgrown with tall (2 m) shrubby willows.

Along the southern edge of the plateau are mountains of the Alaska Range having bare rock and lichen-crusted talus slopes, as well as small glaciers and tarn lakes. To the southwest and west are the Johnson Glacier and Johnson River, a glacial meltwater river with an extensively braided channel. North of the plateau are the predominantly spruce forests and marshlands of the Tanana River valley. To the east is Berry Creek and more upland tundra vegetation.

Methods

To test the validity of assumption (1), that reflectance difference is a good indicator of vegetation difference, I sampled along a transect 3.93 km long. Sampling was done on two lines 30 m apart. Points were located along the lines so that a simple random sample of points was taken with an expected density of one point every .61 m (see Appendix I). The uppermost plant species under each point was recorded as percent cover. The digital picture units (pixels) of ERTS imagery correspond roughly to rectangular regions on the earth 69m x 50m. The actual ground area from which the associated reflectance values are derived is some 10% larger (72m x 52m) and there is overlap in coverage between adjacent pixels. My transect diagonals across 41 pixels and the entire portion of a scene analyzed includes 473,850 pixels.

Since 30m is beneath the resolution of ERTS, the two transects are effectively one line with a point density of one point each .3 m. The simple random sampling of each line was made sequentially, permitting subdivision into segments each lying within a single digital picture unit (pixel). The transect itself extends between the north ends of two lakes that are visible in the ERTS scene, making location of the sampling area possible.

The four spectral bands that are monitored by the ERTS satellite are in the blue-green, orange, red, and near infra-red portions of the spectrum. ERTS has a sensitivity to 256 discrete reflectance values in 3 of the bands, 64 in the near infra-red band. This provides for over a billion (actually 256^3 x 64 = 2^{30} = 1,073,741,824) different possible combinations of reflectance among the spectral bands.

The four spectral bands may be used as axes in four-dimensional space with the coordinates defined by the range of possible grey scale values. As a measure of the difference between two different combinations of grey scale values (spectral signatures), I use a distance measure in the four-space. Since only discrete values of the grey scales are possible, I chose to measure distance between points using the taxicab or absolute value metric, assigning integer values to the set of possible values:

Distance between (x_1,x_2,x_3,x_4)

and $(y_1,y_2,y_3,y_4) = |x_1-y_1| + |x_2-y_2| + |x_3-y_3| + |x_4-y_4|$

i.e., $d(x,y) = \sum_{i=1}^{4} |x_i-y_i|$

To measure the difference in the species composition of areas on the ground corresponding to two pixels, I have used the Matusita distance function (Walsh 1962). It is a measure of the distance $D(p^{(J)},p^{(K)})$ between two multinomial distributions:

$$p_i^{(J)} \text{ and } p_i^{(K)} \quad , \sum_{i=1}^{u} p_i = 1 \quad p_i = 0$$

$$D(p^{(J)},p^{(K)}) = (\sum_{i=1}^{u} (\sqrt{p_i^{(J)}} - \sqrt{p_i^{(K)}}\ )^2)^{1/2}$$

The Matusita distance function has a number of desirable properties. Its usual application is to multinomial distributions. Percentage cover as used here is such a distribution. Inferential statistics have been developed for it which permit evaluation of the significance of differences between the samples (Walsh 1962). It places a moderate emphasis on categories (species) of low probability, neither ignoring them nor giving them equal weight with others of higher probability.

To make the desired comparison I have used the two distance measures as coordinates and plotted all possible pairwise distances between fifteen pixels along the transect. The extent to which a very simple version of assumption (1) is valid is reflected in the correlation (not necessarily linear) indicated by the graph.

The method I have used for clustering pixels once again uses the four-space as a means of structuring relations between spectral signatures. In order to reduce the problem to one which could be more readily handled, I reduced the billion different possibilities to ten thousand. This was done by restricting attention in each band to the range of values which are commonly observed. For each spectral band I examined the number of pixels having each of the 256 (or 64) possible reflectances. The histograms were readily made using programs already available. The final four-space to be examined was to consist of ten units on a side.

Each spectral band was subdivided into twenty-three categories (see Figure 5). Categories 0 and 22 were the tails of the bell-shaped frequency distribution. In between were 21 equal-sized subdivisions of the grey scale. Depending on the dispersion of the frequency distribution for that spectral band, each subdivision was composed of either one or two of the original grey scale units. Subdivision of the region of interest into 21 pieces allowed the construction of two different four-dimensional spaces, corresponding to grouping together adjacent grey

Figure 5. Subdivision of grey scale for the two histograms. The original grey scale for a single spectral band appears in the leftmost column (truncated at 33 for economy in the figure, actually scale extends to 255 [or 63 in band 7]). The 23 subdivisions which lump the tails of the distribution and provide equal sized subdivisions in the remaining 21 are shown in the second column. The third and fourth columns indicate the final groupings for the ten equal-sized subdivisions for the pair of four-dimensional histograms analyzed. This procedure was done separately for each of the four spectral bands.

```
Grey   .25      hist  hist
scale  subdiv.    I     II                         Frequency

00     00       00    00
01     00       00    00
02     00       00    00    -
03     00       00    00    --
04     01       01    00    -----
05     02       01    01    ---------
06     03       02    01    -----------------
07     04       02    02    ----------------------------
08     05       03    02    -----------------------------------
09     06       03    03    ----------------------------------------
10     07       04    03    -----------------------------------------
11     08       04    04    ----------------------------------------
12     09       05    04    --------------------------------------
13     10       05    05    ------------------------------------
14     11       06    05    --------------------------------
15     12       06    06    ---------------------------
16     13       07    06    ----------------------
17     14       07    07    -----------------
18     15       08    07    -------------
19     16       08    08    ----------
20     17       09    08    --------
21     18       09    09    -------
22     19       10    09.   ------
23     20       10    10    -----
24     21       11    10    ---
25     22       11    11    --
26     22       11    11    -
27     22       11    11    -
28     22       11    11
29     22       11    11
30     22       11    11    -
31     22       11    11
32     22       11    11
33     22       11    11
```

Figure 5. Subdivision of grey scale for the two histograms. (Simulated frequency data)

scale values first as: 1-2, 3-4, 5-6 . . . 19-20, leaving #21 out, then frame-shifted as 2-3, 4-5, 6-7 . . . 20-21, leaving out #1. In this way I would have an indication of the sensitivity of later work to the particular grouping choice.

After I defined the four-space to be used, a four-dimensional histogram was made. Each cell (4-tuple) in the histogram contained a tally of the number of pixels in the 40-km square having the corresponding coordinate set. This histogram served as the basis for the rest of the classification effort. It was this space that was examined for clusters.

To begin looking for clusters I first adopted some conventions:

(1) The only neighbors of a point are its orthogonal ones (i.e., those which differ from the given point by exactly one unit in one and only one coordinate).

(2) A <u>path</u> from a point P_0 to a point P_n is a sequence of points P_0, P_1,P_2 . . . P_n such that P_1 is a neighbor of P_0, P_2 is a neighbor of P_1 . . . P_n is a neighbor of P_{n-1}.

(3) Pixel density at a 4-tuple, ϱ (P) refers to the number of pixels from the 40-km square which are assigned to the associated cell (4-tuple location) in the 4-space.

(4) Two points P_0,P_n are in the same cluster if and only if there is a path P_0,P_1 . . . P_n such that pixel density at each P_i is not less than a specified cutoff value c. (Note: If pixel density at a point is less than c it cannot be in a cluster and hence is unclassified).

I will use a two-dimensional analogy to discuss many of these relationships. The underlying 4-space is represented as the sea-level projection of an area on the earth's surface. Pixel density at the corresponding points may be visualized as a mountain range above the points in the data. Two points are in the same cluster when there is a way to get from one to the other without going below a specified contour line (cutoff value). The definitions of neighbor and path are adaptations of these ideas to the integer-valued coordinate system of the four-space. It amounts to overlaying a contour map with a chessboard. Neighbors are then the single-square moves of a rook and points are in the same cluster if they may be reached by a rook without crossing the specified contour.

With these basic conventions I hoped to formulate an idea of cluster which is close to that used in biological notions of taxonomic units. Since the ERTS sampling is exhaustive in the area of coverage, the data contain "population size" (=frequency) information for each possible spectral signature. As in biological notions of taxonomic unit, my definitions yield a taxon when a particular character set and variations on it are common. A rarity of intermediate forms permits distinction from other taxa. As will be seen later, this notion is a bit too restrictive. It seems that rarity must be relative to yield desirable properties in a final classification.

To identify clusters, the 4-dimensional histogram was first converted into an array of zeros and ones. If ρ (P) $<$ c the point was assigned a "0" and if ρ(P) $\geq$ c the point was assigned a "1". This corresponds to points being unclassifiable (below the contour) and classifiable (above the contour) respectively. I worked out a procedure for identifying clusters from this new array. It was not well suited to computer execution, however, and instead a "tree-search" technique was used to associate points which are in the same cluster. This is a systematic way of starting with an initial point to be classified, searching out a neighbor which may be classified (has value "1"), finding a neighbor of it with value "1", and so on until all points in the cluster have been found. A search is then begun for another point with value "1", but which has not yet been classified.

The process was done for 65 different cutoff values in each of the two histograms. The output was the number of clusters, the number of pixels in each cluster and the number of 4-tuples in each cluster in 4-space.

To compare the effects of the frame-shift in making the two histograms I examined three properties of the data: two involving the gross structure of the histograms and one of their finer structure.

Results and Discussion

In Figure 6, 15 digital units (pixels) were compared pairwise in all combinations. Vegetation differences for the corresponding ground sampling were measured using the Matusita distance function. Spectral signature differences were measured using the taxi-cab (absolute value) distance function. A point on the graph indicates a single pairwise comparison of two pixels with x-coordinate taxi-cab distance between their spectral signatures and y-coordinate Matusita distance between their vegetation. These distance measures are abstract measures of the

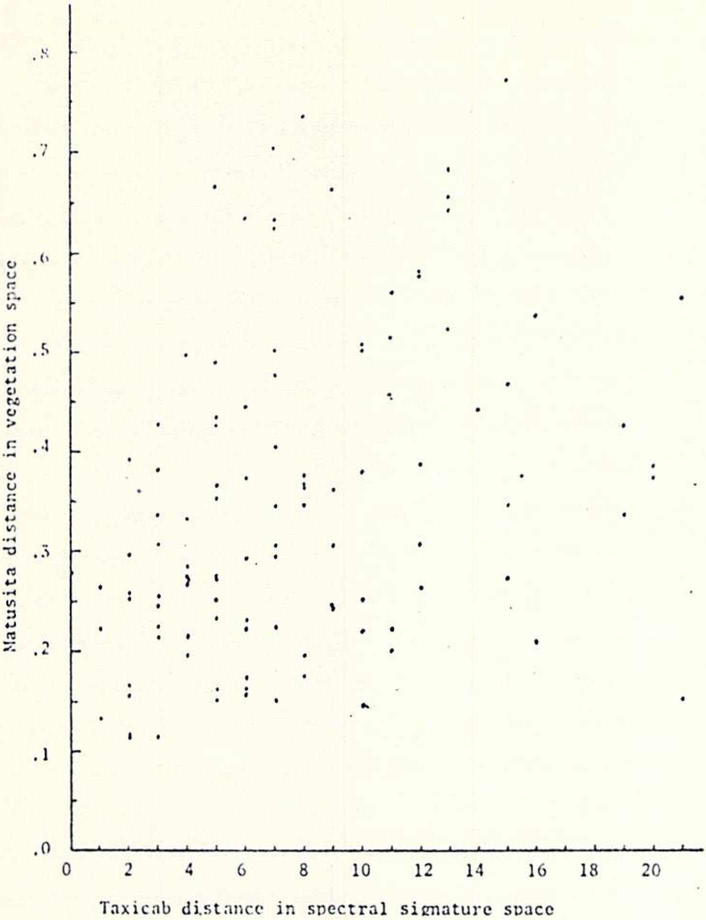


FIGURE 6 Graph showing lack of correspondence between Vegetation differences, as measured by Matusita distance; and spectral signature differences, as measured by taxicab distance.

difference between a pair of observed data sets. Matusita distance is used as a distance in n-dimensional "vegetation space," each dimension representing a species, while the taxi-cab distance is used in a 4-dimensional "spectral signature space," each dimension representing a spectral band. The scatter of points shows that this naive form of assumption (1) is wrong. Spectral signature differences do not have a simple relationship to vegetation differences. Small differences in spectral signature may relate to larger differences in vegetation, and vice-versa. In this regard it is unfortunate that my transect had the degree of homogeneity that it did. It might have been nice to cross a more heterogeneous tract of tundra but of the three legs of the triangle defined by the three ERTS-visible lakes on the plateau the segment chosen was the most heterogeneous. However, ERTS is capable of distinguishing differences here: of the 41 pixels along my transect there were 38 different spectral signatures (all similar but different). It is possible that the disparity in correspondence is due to excessive sampling variability in the ground sampling in spite of over 300 data points in each pixel. This suggests that careful study is necessary to determine the vegetation characteristics which contribute to the observed differences. One important consideration is the date of the ERTS scene: 21 August 1972. It is not likely that the vegetation had changed significantly in the 2 years between the time the ERTS data were taken and the time my sampling was done, but 21 August is late enough in the season that senescence of some plants is under way. Many plants (willows especially) have begun to drop leaves or die back so the cover estimates made several weeks earlier in the season are no longer so good if exact correspondences are desired. Cloud cover made more recent scenes or ones from earlier in the season useless.

These difficulties point out some important considerations in the use of satellite imagery. Available imagery may not always be ideally suited to the type of information desired. A person interested in surface features is hindered by cloud cover. As such, it is desirable to investigate ways which permit available imagery to be used as fruitfully as possible. Satellite and ground information are almost never contemporaneous and important changes may transpire. The seasonal changes at issue here are one example. In one sence the composition of perennial vegetation does not change. However, the composition in terms of cover, which is what a satellite "sees," may change through the seasons. My sampling was aimed directly at the cover relationships a satellite monitors by reflectance. As such it is particularly sensitive to such changes. Estimates of species "importance" in terms of cover will change drastically with senescence of plants when one of the "species" is dry dead plant material. In order to gain a detailed insight into the correspondence with satellite imagery, it seems necessary to have more information chronicling the nature of such changes on the ground.

Ground sampling itself is an important problem in heterogeneous vegetation communities. Differences from place to place can nearly always be found, but whether their magnitude is due to population differences or sampling variability is often open to question. The Matusita distances observed here are not large. Considerations as outlined in Part I for detecting significant differences indicate that, in general, the sampling intensity and number of species involved conspire to prevent statements even asserting the observed differences to be significant, to say nothing of their magnitudes.

Another difficulty is the mathematical one concerning the nature of a correspondence between distance measures. Simplistically it seems that such a correspondence would be straightforward. It is not. Even considering the plane, using two distinct distance measures, the correspondence is poor. Consider the taxi-cab and Euclidean distances of length 1 around the origin. The taxi-cab metric "unit circle" is a diamond with vertices (0,1); (1,0); (0,-1); (-1,0) while the Euclidean unit circle is a circle of unit radius. These distance measures agree at only 4 points (the vertices of the diamond) and have varying amounts of disagreement everywhere else.

The virtual coincidence of data points in figures 7 and 8 gives an indication that there is very little grouping effect on the large scale structure of density in the histograms. Figure 7 shows a very nearly lognormal distribution of occupancy of points in the 4-dimensional histograms by picture units (pixels). The logarithmic probability paper yields a straight line when the distribution is lognormal. The x-axis indicates the cutoff value (the minimum number of pixels which must be associated with a point [4-tuple] in the 4-space). The y-axis is the percent of the pixel-bearing 4-tuples that have a pixel density exceeding the cutoff value. By analogy with the work of Preston (1948), 4-tuples were treated as species and pixels were treated as individuals to be distributed into species. The distributions from the two histograms are virtually indistinguishable.

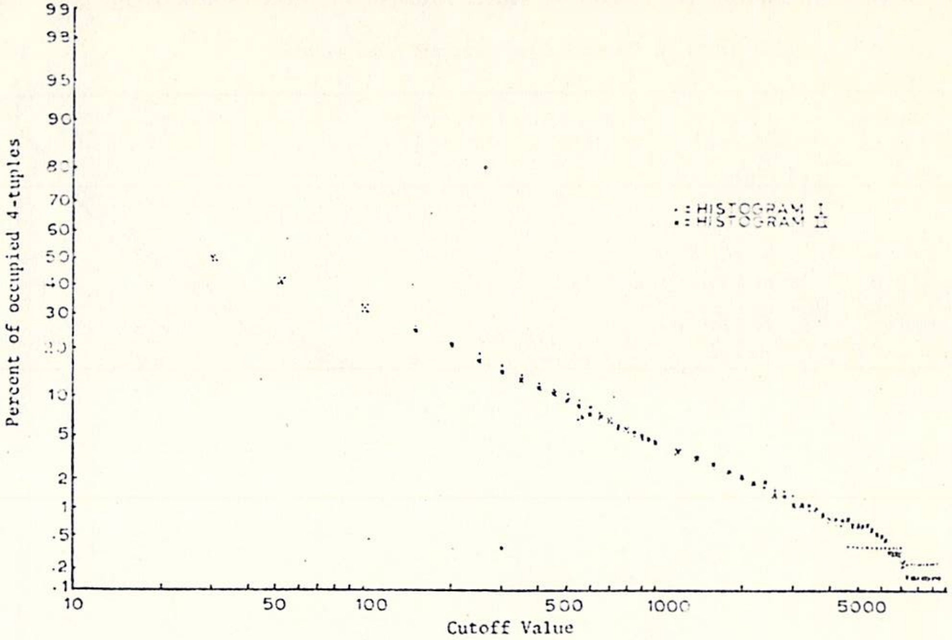


FIGURE 7 Graph showing the very nearly lognormal distribution of occupancy of 4-tuples. Histograms I and II and virtually indistinguishable.

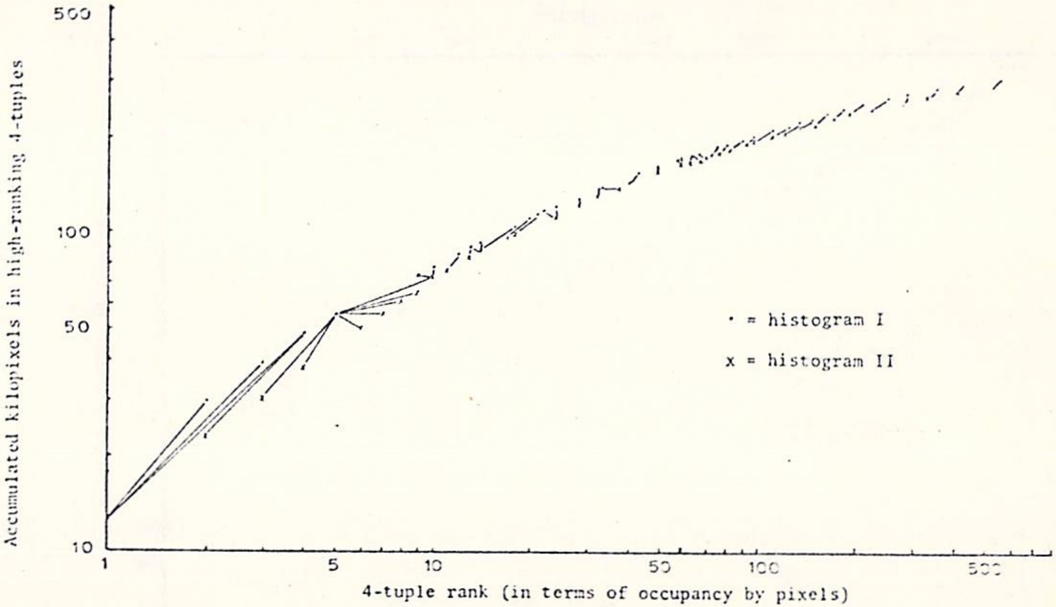


FIGURE 8 Graph showing another similarity of the two histograms: accumulated pixels in high ranking 4-tuples vs. 4-tuple rank. Line segments connect data points from the same cutoff value.

Figure 8 shows another similarity of the pair of 4-dimensional histograms. In each histogram the points in the 4-space (4-tuples) were ranked. The one having the largest number of picture units (pixels) associated with it was ranked #1, the second largest #2, etc. The y-axis then indicates the total number of kilopixels (units of 1000 pixels) accumulated in all the 4-tuples of rank less than or equal to the rank indicated on the x-axis. The ranking was accomplished by the clustering procedure at different cutoff values. The points resulting from the same cutoff value in the two histograms are connected by line segments. The data points for histogram II are consistently lower than those for histogram I since histogram II contains fewer pixels. The overall shape of the curves, however, is much the same.

Neither of these treatments gives any indication of the structure of the clusters through different cutoff values (i.e., changes in the apparent terrain as a result of the two different grouping procedures). To indicate this effect I traced the histories of the different clusters from one cutoff level to another based on the information about number of pixels and number of 4-tuples. I was nearly always able to determine unambiguously how much a cluster shrank going to a higher cutoff level. I then used this information to make a stylized cross-section of the mountain range.

Figures 9 and 10 help to illustrate the "topography" of the 4-dimensional histograms. The vertical axis indicates the height of the various peaks over the 4-space. Separate peaks join at the cutoff value which no longer permits them to be separated. A line parallel to the base in general will have several segments lying below the curve. Each

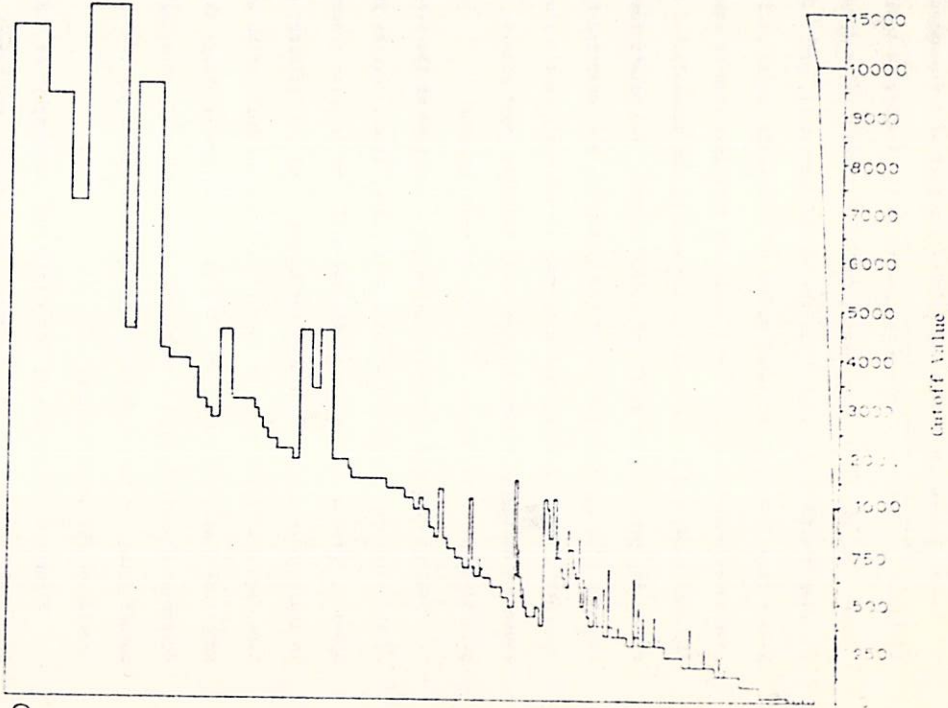


FIGURE 9 Stylized "cross-section" of 4-dimensional histogram I. (see text).

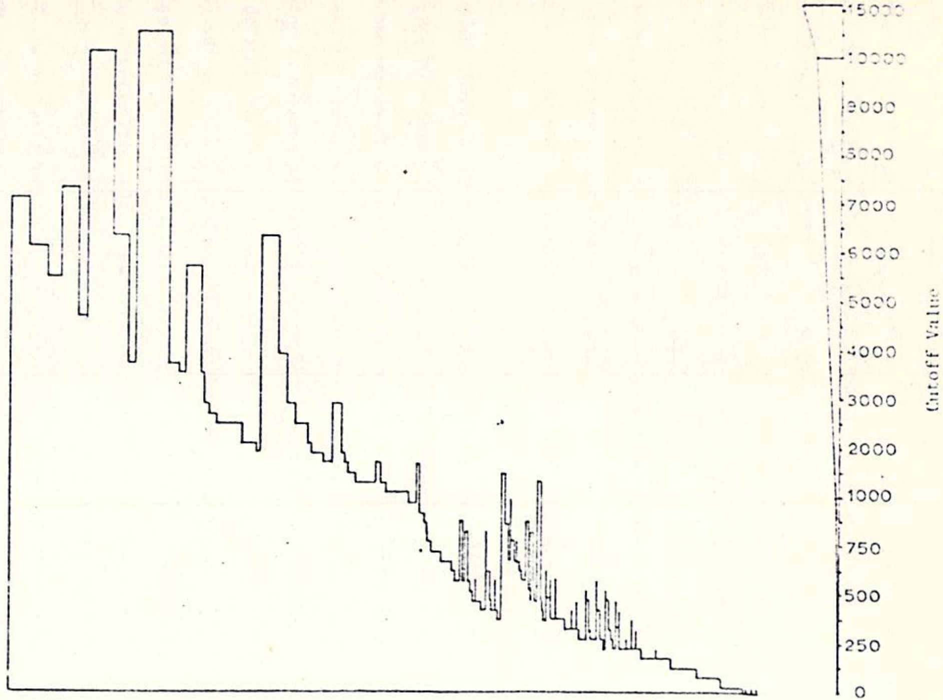


FIGURE 10 Stylized "cross-section" of 4-dimensional histogram II (see text).

such segment represents a separate cluster distinguishable at that cutoff value. The length of such a segment indicates the number of picture units (pixels) in the cluster. The inverted pyramid on the right is a scale for the length of such segments. A horizontal segment in it shows the length of a segment representing c pixels where c is a cutoff value. At the same time it shows the length of a segment representing a single point in the 4-space having c pixels associated with it. Note that there are three different scales for the cutoff value axis. In constructing the graph from the base, whenever there was a choice about the positioning of a cluster, the largest one was put on the left.

These graphs indicate the terrain over the 4-space. They show that there is basically one big mountain with a number of spires coming off it. Many of the spires are quite well defined and maintain their integrity across a wide range of cutoff values. For the most part they are extremely localized in the 4-space, consisting of only one or two 4-tuples. Comparing the two histograms it appears that the clustering is sensitive to different association of units in the grey scale but even so there is a rough correspondence between the two graphs. It is clear, however, that no one cutoff value will serve to distinguish all the spires that do exist. This suggests a revision of the notion of cluster to involve these peaks and their locations in the 4-space with less importance attached to a particular cutoff value. Successive cutoff values may be used, however, to indicate affinity between peaks yielding an hierarchy of relatedness and the levels at which these connections exist.

This suggests another way of thinking about these graphs, relating them to the dendrograms of numerical taxonomy. The "peaks" are now "branches." The graph contains more information than just level of relatedness, however. The width of a branch indicates the number of individuals (pixels) or population size in the branch.

It was possible to check the classification of pixels on my transect at only a few cutoff values. For all of these values the pixels were either unclassified or in the same (the biggest) cluster.

One further point of interest is that in each histogram of 10000 points 300,000 pixels were tallied. Only about 13-1/2% of the points in 4-space had any pixels associated with them at all. More than 95% of the 4-tuples were below the average occupancy of about 30 pixels per 4-tuple. This is consistent with the extreme skewness of the lognormal distribution which describes this aspect of the data so well (Figure 7). It indicates a high correlation between the reflectances in the four spectral bands and suggests possibilities for economizing in computer memory allocations if the correlation could be described well analytically (see Appendix II for further modifications and details of the clustering technique).

The notions of classification used here differ from those used by other investigators (e.g., Cibula 1972, Colwell <u>et al</u>. 1970). The more useful approach, particularly in agricultural applications, has been to observe "test plots," find their spectral signatures, and assign the corresponding signatures to a classification associated with the identity of the test plots. For much agricultural work and some wildlands vegetation the assignment of such a classification unit from the ground is reasonably clearcut. Natural vegetation is usually not the monoculture of agriculture and there is much more opportunity for variation from one "community type" to another.

Under some form of assumption (1), the existence of natural taxonomic units defined solely by ERTS data would produce an indication of the extent to which such variation between community types does exist on a geographical scale, even in a non-agricultural setting.

Conclusions

The correspondence between different reflectances and differences in vegetation is not a simple correspondence between distance measures in abstract "spectral signature space" and "vegetation space." The interplay between vegetation and spectral signature changes appears to be more complex. The problems take several forms. There is some question about the adequacy of the ground truth obtained; it appears that the intensity of sampling (300 data points per pixel) may have been inadequate to delimit vegetation differences sufficiently, since the indicated differences are not large and cannot be readily separated from possible sampling variability. Of 41 digital picture units (pixels) along the transect sampled, there were 38 different spectral signatures. Other complications to quantifying any correspondences are the senescence of plants between the time of my sampling and the season of the satellite data several weeks later, and a mathematical problem of the sort of correspondences such distance measures might be expected to give. These problems raise questions about the value of more detailed quantitative consideration of these data.

It appears that while ERTS does distinguish differences in the vegetation of the Macomb Plateau, a "natural" classification of the data from the plateau and surrounding areas does not distinguish any vegetative units from along my transect.

The ERTS data do subdivide into natural units. These units may be conceived of as spires in a mountain range of pixel frequency over a 4-dimensional space defined by the 4 spectral bands that ERTS monitors. These units have a high degree of integrity but are basically spires on a single big mountain. Their existence does not seem to be an artifact of the data reduction procedure used although their detailed form is sensitive to particular choice of reduction procedures.

As a method of data summary, the frequency of occurrence of pixels having given spectral signatures may be approximated very well with a lognormal distribution.

APPENDIX I.

A SEQUENTIAL WAY OF CONDUCTING A SIMPLE RANDOM SAMPLING OF AN ORDERED SET

Simple random sampling is often a desirable sampling strategy because the resulting data usually conform well to assumptions for subsequent statistical analysis procedures. One of the drawbacks of the procedure, however, is the prior organization that it requires. A simple random sample of points along a line, for example, usually involves either much back-tracking or a preliminary process of generating random numbers and putting them in order, followed by the actual sampling procedure. Here I describe a technique for generating the points in order from the start. This may be done in the field as sampling is conducted and consequently eliminates both prior organization of the sampling scheme and repeated back-tracking.

The technique will generate a series of points on a line. Some preliminary calculations will permit the expected number of points sampled to be predetermined. If an exact number of points is required, the technique may still be used to sample most of the points; if too many are sampled some may be randomly removed; if too few, some back-tracking for the remainder of points becomes unavoidable.

The method is based upon waiting times between random events. When a random variable is uniformly distributed on an interval, the number of points in a subinterval of fixed length is Poisson distributed while the distance between events is exponentially distributed (Parzen 1960). By generating random numbers which are exponentially distributed and accumulating their sum, a simple random sample of points may be gathered from the interval.

Specifically, the density function for the exponential distribution is given by $f(x) = \frac{e^{-x/a}}{a}$ and has an expected value of a (the mean spacing between points). In order to generate random numbers which are exponentially distributed, uniformly distributed random numbers may be transformed using the inverse cumulative exponential function (Abramawitz and Stegun 1965:950):

density function $f(x) = \frac{e^{-x/a}}{a}$, $a>0 \quad x>0$

cumulative distribution fn. $F(x) = 1 - e^{-x/a}$

$u = 1 - e^{-x/a}$

$1 - u = e^{-x/a}$

$\log_e (1-u) = -x/a$

$-a[\log_e (1-u)] = x$

inverse cum. exp. function $x = -a\ [\log_e(1-u)]$

If u is a uniformly distributed random number $0<u<1$, then x will be exponentially distributed on $(0,\infty)$, but $(1-u)$ is also uniformly distributed on $(0,1)$, so the required transformation is the negative of the mean value desired times the natural logarithm of the uniformly distributed r.v. To generate sequential random numbers then proceed as follows:

(1) Generate a set of random digits (as in a table of random

numbers) and consider it as a random number between 0 and 1.

(2) Look up its natural logarithms and multiply it by the negative of the desired mean spacing.

(3) Add the resulting value to the position of the previous sample point.

(4) Repeat the process in this fashion. The segment will be traversed and the points sampled will be a simple random sample of the points on the segment.

In summary:

Let L = length of segment to be sampled

N = total number of sample points desired

$a = L/(N+1)$ = mean spacing between points

$u_i = i^{th}$ uniformly distributed random variable $0<u_i<1$

$x_i = i^{th}$ exponentially distributed r.v. $0<x_i<\infty$

then

$$S_j = -\sum_{i=1}^{i} a\,[\log_e(u_i)] = -\sum_{i=1}^{j} \frac{L}{N}\,[\log_e(u_i)] = j^{th} \text{ r.v., in}$$

order, of a uniformly distributed sample of expected size N on segment of length L.

Further notes

1. In practice three significant figures is a typical precision of measurement for the average spacing. The operation of looking up natural logarithms and multiplying by a constant may be conveniently circumvented by making a graph using semilog paper. Three cycles is probably sufficient.

The graph of the desired transformation is then a straight line. The log scale is the u scale and two points for defining the line may be conveniently taken as $u=1$, $x=0$ and $u=.006735$, $x=5a$.

The resulting graph may then be used to read off the transformed values x_i. The running total may be kept in an inexpensive mechanical adding machine.

The use of the graphical method also permits generating u_i of a constant number of significant figures and deriving from them x_i of a constant precision. This permits each x_i to have the same number of decimal places and eliminates the feature of a fixed large value of x_i when u_i is small--variation is more continuous.

2. This technique may also be used to systematize 2-dimensional (3-dim) point sampling. By generating the points for one of the coordinates in order, a systematic path is generated through the sample points which, while it may not be a minimal path, will be a shorter path than many and represents a "real-time" systematic procedure for accomplishing the sampling.

3. Particularly in the case of sampling on a line, the resulting data contain more information than in a back-tracking procedure. Since the points are sampled in order, notes are also sequential and as a result there is some information about neighboring points. In some applications this auxiliary information may be of interest and is a bonus made possible by the technique described here.

APPENDIX II.

THE CLUSTERING ALGORITHM AND MODIFICATIONS TO INCREASE ITS VERSATILITY

The key to clustering algorithms is the identification of points which may be connected by paths. A tree search procedure was used to solve this problem.

Given an array of points already identified as classifiable (1) or unclassifiable (0), the demon doing the search (computer CPU) examines points and keeps track of where it is using a push-down stack. Each neighbor is examined and if classifiable it becomes the initial point of a new search with its origin stored in the stack. When it runs out of classifiable neighbors it returns to the stack for information about its origin and resumes that search. As I had hoped, the classification procedure could be executed very quickly.

Several modifications of the basic technique are potentially desirable. The large amount of data presented by ERTS means that there is an inherent smoothing of occurrence of spectral signatures. Further smoothing $\rho(P)$ might be desirable. For this purpose a weighting function $W(P)$ which fell off rapidly from the point P would work:

$$\bar{\rho}(P) = \sum_{P \epsilon \text{space}} W(d(P,P_0)\ (P)$$

or more generally

$$\sum_{P \epsilon \text{space}} W(P,P_0)\rho(P)$$

If K(P) fell off very rapidly, only points near to each P_0 need be considered in the calculations. No smoothing at all corresponds to

$$K(P) = \begin{matrix} 1 & P = P_0 \\ 0 & P \neq P_0 \end{matrix} \quad .$$

Another problem with ERTS imagery is the effect of sun angle. Different sun angles (as on different slopes) may yield different spectral signatures even though the reflectance might be identical under like illumination. As a first order correction to this I suggest approximating the effect of lower sun angle as that of a neutral density filter causing lower illumination but of the same spectral quality. An effect of this sort could be removed from the data by conversion from the rectangular coordinate system provided by the 4 grey scales to a polar (hyperspherical) coordinate system. The 3 direction angles would then define sets of points having the same spectral quality but differing only in intensity. Identification of all such points (ignoring the radius) might be too rash a move but sun angle considerations could be readily handled in that form since most of its effect would be in that one coordinate.

LITERATURE CITED

Abramawitz, M. and I. A. Stegun. 1965. Handbook of Mathematical Functions. Dover Publications, Inc., New York. 1046 p.

Cibula, W. 1972. Application of remotely sensed multispectral data to automated analysis of marshland vegetation. ERL Report 020, NASA. Earth Resources Laboratory, Mississippi Test Facility.

Colwell, R. N. *et al*. 1970. Application of remote sensing in agriculture and forestry. *In* Remote Sensing. National Research Council of Nat. Acad. Sci., Washington, D.C. 424 p.

Diem, K. (ed.). 1962. Documenta Geigy Scientific Tables. Geigy Pharmaceuticals, Division of Geigy Chemical Corp., Ardsley, N.Y. 778 p.

Grieg-Smith, P. 1964. Quantitative Plant Ecology. Butterworth, London. 256 p.

Holmes, G. W. and H. L. Foster. 1968. Geology of the Johnson River Area, Alaska. Geol. Survey Bull. 1249. U. S. Government Printing Office, Washington, D.C.

Hulten, E. 1968. Flora of Alaska and Neighboring Territories. Stanford University Press, Stanford, California. 1008 p.

Ostle, B. 1963. Statistics in Research. Iowa State Univ. Press, Ames, Iowa. 585 p.

Parzen, E. 1960. Modern Probability Theory and Its Applications. John Wiley and Sons, Inc., N.Y. 464 p.

Preston, F. 1948. The commonness and rarity of species. Ecology 29:254-283.

Walsh, J. E. 1962. Handbook of Non-parametric Statistics. Van Nostrand, Princeton, N.J. 549 p.