USE OF REMOTE SENSING IN SAMPLING FOR AGRICULTURAL DATA

H. F. HUDDLESTON AND W. H. WIGTON

U.S.A.

ABSTRACT

The Statistical Reporting Service of the U. S. Department of Agriculture has been developing methods of using remote sensing as a basis for improving survey methodology. The paper discusses remote sensing techniques which have been developed to use scanner data or film acquired by ground personnel, aircraft, and satellite. Remote sensing information has been used in conjunction with conventional data sources and observations in the context of double sampling to estimate number of fruit on trees, the number of trees in an orchard, and the acreage of crops planted. All the techniques employed use remote sensing information in digital form and rely on computer analysis of magnetic tapes. The methods of analyses include discriminate functions, spectral and spacial clustering, and periodogram statistics.

Introduction

This paper reports on a number of research activities in the Statistical Reporting Service (SRS), USDA, directed at utilizing remote sensing information obtained by satellite, aircraft or man. These efforts are directed at increasing the efficiency of collecting current agricultural statistics. In addition, the resulting techniques also have application to sample census surveys.

The first major effort involving SRS was in California in 1961. A method was developed using high resolution aerial photographs to obtain an estimate of acreage and production of raisins as a basis for determining the grape acreage which would remain for wine. This application relied on photo interpreters to identify trays of grapes which were laid between the rows in vineyards for drying as a basis for an acreage inventory. Shortly, after the trays were identified and counted, ground crews would verify tray counts and sample trays to obtain an average weight of grapes per tray from which a yield per acre was derived. This application was highly successful in meeting the marketing objectives of the wine industry. However, the work was privately financed and was terminated when the industry's supply situation changed in 1968. While this application was successful, it was not cost effective in comparison to alternative data collection methods in terms of dollars, but it was undertaken because of the speed of acquiring the acreage information

(48 hours) and control of survey procedure through close supervision of photo interpreters under office conditions.

Beginning in 1967, SRS began exploring the use of information secured from films and scanner devices as sources for crop and livestock data. These efforts were markedly influenced and expanded by the pending availability of highly sophisticated photographic and scanner equipment from aircraft and satellites. Since the information provided from these systems are from sensors which provide only indirect measurement variables (i.e., secondary data) rather than the primary data sought and do not provide a source of information for many survey characteristics, such as, crop utilization, varietal data, prices, and farm income, our efforts have been directed at securing information which would provide suitable auxiliary data for use in sample frame construction, stratification, sample selection or in the method of estimation. In addition, our efforts have been largely directed at working with the information in a digital format and seeking sophisticated computer solutions to problems of identification and counting. Alternatively, greater reliance could be placed on photo interpretation or visual interpretation of recognizable patterns of digital data displayed in map form. Our efforts have been limited in this respect since resources for this type of application have not been available. The applications are discussed under the two general categories: (1) low resolution sensors, and (2) high resolution sensors.

Studies Using High Resolution Sensors

The earliest studies employed cameras as the sensors using a variety of films acquired with aircraft at relatively low altitudes. Consequently, it was possible to resolve or detect very small objects with the aid of suitable viewing equipment, but the area covered by a single photograph was typically less than 8 square miles (or 12 Km^2). Improvements in sensor technology, platforms, and high altitude photography produced high resolution images and covered larger areas. With these improvements, studies were started which employed high sensor resolving power so spacial properties as well as the spectral response properties of crops could be utilized.

One phase of our research has been to develop a system for identifying and counting fruit on photographs acquired by ground personnel using a 35mm camera. The same system has also been used to identify and count objects of interest acquired with high resolution aerial photography or digital scanner data.

The application of this system has been developed for counting clusters of oranges (Gleason and Hopkins 1975) using conventional 35mm color slides of trees, and counting fruit trees (Nealon 1975) in an orchard from aerial infrared photographs have been digitized using a scanning microdensitometer. A basic statistical tool used in this system is discriminate functions (Rao 1965) which classify data points as part of a data reduction step. In addition, a clustering technique based on the minimal spanning tree concept (Zahn 1971) uses the spatial properties (size and compactness) of the "target" of interest to develop homogeneous groups. These two tools are utilized sequentially to improve the accuracy of the system.

The classes of objects found in the scene on the respective film transparencies were as follows for these two applications:

(a) Counting Oranges	(b) Counting Trees	
(1) Sky	(1) Lake	Combined for classification purpose
(2) Ground	(2) Soil	
(3) Foliage	(3) Canal	
(4) Oranges	(4) Road	
	(5) Bushes	
	(6) Hedge (Trees)	
	(7) Citrus (Orange) trees	

The digital data consisted of four response variables for each picture element (pixel): red, green, blue and clear filter readings from color infrared film where the high resolution data were obtained through magnification. Figure 1 shows the spectral data and separation of the various classes of targets using the red and blue filtered readings obtained from a scanning microdensitometer for "training data" (i.e., a sample of pixels for the objects from the scene). The use of different data modes (transmission units versus density units) is possible since the output analog signal may be either logarithmic or linear with the microdensitometer used. For these two examples, two dimensional feature selection indicated obtaining the digital data in transmission values for oranges and density values for trees. In the first case, the object of interest (oranges) is "relatively light" and is better separated from other objects by a linear scale. In the second case, the object of interest (trees) is a "dark" object that is better separated from other background objects on a logarithmic scale. The relationship between these two units of measuring light intensity is: Density = Log_{10}

(1/transmission). The aperture size (i.e., pixel size) for the 35mm slide was 100 microns by 100 microns which represent about 1/84,000 of the total area of the slide. For the aerial photo transparency (9" x 9") the aperture size was 240 microns by 240 which represented about 1/1,000,000 of the total area of the transparency. Chart 1 shows the data reduction and final results for a limited amount of data. This same sequential system can be used with low resolution sensors where the objects to be detected are relatively large and possess distinct spatial characteristics.

A second approach (Gallant, Gerig, and Evans 1974) for field crops uses the spacing of data points at fixed distances along a straight line across fields as well as the low resolution sensor responses to identify crops or land use. The basic tool employed for the high resolution data is "spectral analysis" as commonly used in time series where dependencies may exist between data points acquired at regular intervals. One data point, along the straight line, is acquired corresponding to about every 4 inches on the ground. Four statistics were used to describe the periodogram fitted to the "clear" variable by a linear segmented model after the "spikes" were removed from the clear filter readings. These statistics were: (1) Relative amplitude (or intercept) at zero, (2) slope from 0 to .25 radians, (3) slope from .25 to 1 radian, and (4) slope from 1 to π radians.

Table 1 below shows the misclassification rates for a linear discriminate analyses using means, variances and covariances; periodogram statistics and combining both sets of statistics in a single discriminate analysis. The information obtained from high resolution optical scans (periodogram statistics) are shown to significantly improve crop classification results over the low resolution statistics (means, variances and covariances) acquired from digital optical density of an aerial photograph from a scanning microdensitometer.

Where the counts of the objects of interest are highly correlated with the primary data, several statistical techniques for using the information is practical. For counts of fruit on trees, double sampling needs to be employed since some fruit cannot be "seen" by any process of interpretation. In addition, the year to year changes in fruit set is likely to alter the fraction of fruit which is "visible" in the interpretative scheme. We can approximate the effect on the sampling error in double sampling using either a ratio or regression type estimator. The variance reduction expected is based on the formula below Table 1.

Chart 1.--Schematic diagrams showing data acquisition and classification of picture elements (i.e., pixels)

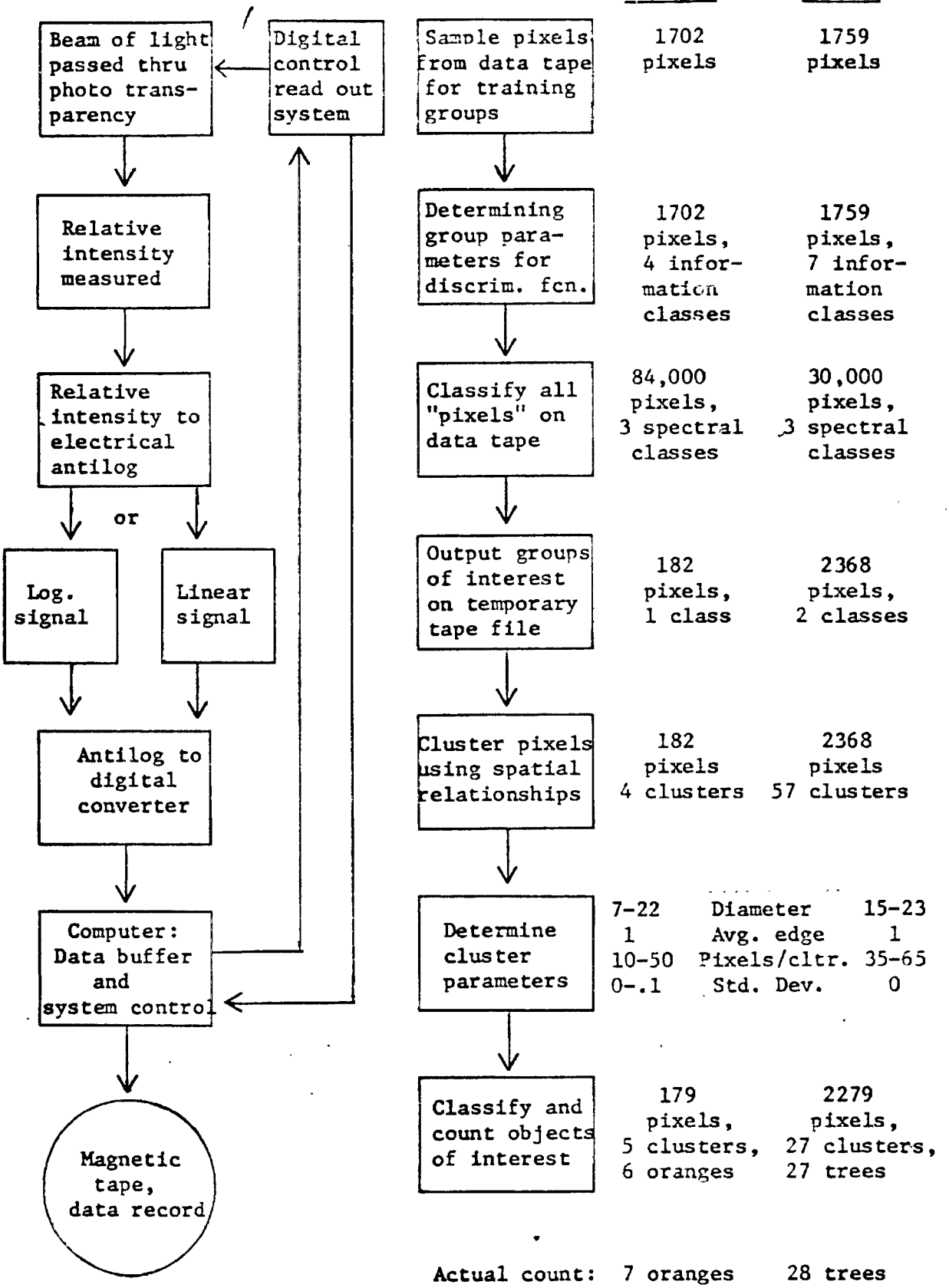

Table 1.--Misclassification errors using a linear discriminate function (No prior probabilities specified for crops)

Crop	Based on means, variances, covariances for four density variables	Four periodogram statistics from clear density readings	All statistics combined
	Percent	Percent	Percent
Cotton	57.3	17.1	4.2
Grapes	52.8	19.4	5.9
Oranges	58.0	14.3	5.8
Almonds	57.2	23.9	7.7
Alfalfa	47.4	9.5	1.5
Corn	35.5	8.5	2.5
Walnuts	41.9	17.3	5.3

$$S^2_{\bar{X}_n} \doteq \frac{S^2_X}{n'} [1 - \rho^2(1 - \frac{n'}{n})]$$

where: S^2_X is the variance per tree of the fruit counted using ground crews, ρ^2 is the correlation coefficient between the remote sensing count and the count obtained by ground crews, n is a random sample of trees for which remote sensing information is available, and n' is a random subsample to be counted by ground crews.

For the three fruits, oranges, apples, and peaches, for which experimental evidence is available (Huddleston 1971), ρ^2 varies from .5 to .9. Using a value of $\rho^2 = .7$ and $n' \div n = .2$, a variance reduction of at least 50 percent would be expected using remote sensing techniques as compared to only counting fruit on the n' trees by ground crews.

In counting fruit trees on aerial photographs by use of remote sensing techniques, three levels of sampling could be used with the remote sensing technique being employed at the second stage. At the first stage (or stage with large area coverage) n_1 photos would be selected that had one or more orchards shown. For the second stage, n_2 (or a subsample of the n_1) photos would be selected for counting trees. The

third stage would consist of selecting n_3 photos (from the n_2) with probability proportional to the number of trees counted which would be enumerated by ground crews to classify the trees by varieties (or kind) and possibly obtain production data.

By using information in the periodogram, the above procedure could be modified by the incorporation of another stage of sampling. The third stage of sampling could employ the periodogram statistics to classify the trees into varieties. The fourth stage would then consist of a smaller subsample of n_4 photos for which ground crews would obtain additional information on number, variety and production per orchard. Several important distinctions exist between the fruit counting and tree counting examples: (1) The estimation of tree numbers is more highly correlated with actual numbers than for fruit counting, and (2) the estimator for tree numbers appears to be approximately unbiased. Consequently, tree numbers in patterned orchards can be estimated using high resolution sensors as an independent information source.

Studies Using Low Resolution Sensors

Low resolution imagery studies have been based on satellite or high level aircraft which can provide coverage for large geographic areas. While the elementary unit of information, the pixel, corresponds to a relatively large area on the ground, there is still a very large number of data points which must be analyzed for a given scene. Consequently, a sampling scheme must be developed for estimating discriminate function parameters and classifying areas of interest into desired crop types.

Since the basic sampling frame used by SRS for major crops and livestock surveys is an area frame, remote sensing imagery can provide additional information for units in the sampling frame. The selected primary sampling units (count units) and secondary units (segments) represent a probability sample from the total land area of the United States. Consequently, the use of remote sensing information from satellites or aircraft which can be related to specific land areas provide a means for improving survey estimates.

Specifically, the use of low resolution satellite information obtained from ERTS-A provide the basis for several studies for agricultural sampling purposes. Much of this research was jointly sponsored by the National Aeronautics and Space Administration (NASA). One aspect of the SRS studies has been to examine either the classification matrix or the correlations between crop acreages obtained by field enumerators and by

classifying "pixels" using parametric quadratic discriminate functions for area sampling units. For these studies, enumerators made special visits to obtain ground data corresponding to the date of the imagery for estimating parameters and evaluating results. The basic task was that of differentiating between two or more crop (or land use) populations on the basis of multivariate measurements (See Table 2). The problem is to partition the measurement space in some optimal fashion so that the pixels are allotted, if not all correctly, at least sufficiently accurately to bear a close relationship to what is actually present. The method of linear discriminate functions has not been used due to the variance-covariance matrices by crops not being equal; instead the method of constructing contour "surfaces" in the measurement space has been used. A multivariate normal density distribution was estimated for each crop. The departure from normality was not a serious problem as long as the precaution was taken to insure unimodal data. In general, this entails verifying the marginal histograms are unimodel and when they are not to create two or more subclasses within crops. Or, alternatively to employ clustering techniques for each crop type to identify the existence of more than one group. At this point, mean vectors and variance-covariance matrices are calculated based on selected fields which constitute training data for each crop. The quadratic discriminate function was calculated based on the sample statistics. All unknown pixels are classified into one of the crops categories for which the crop mean vector is closest to the point based on the Mahalanolis distance. That is, the crop for which the probability is highest.

The measurement space consists of four sensor spectral bands shown in the table below:

Table 2.--Sensor spectral band relationships

Sensor	Spectral band no.	Wavelengths (micrometers)	Color	Band code
MSS.....	1	.5 - .6	Green	4
MSS.....	2	.6 - .7	Red	5
MSS.....	3	.7 - .8	Near infrared	6
MSS.....	4	.8 -1.1	Near infrared	7

The results of the quadratic discriminate function are presented in the form of a classification matrix. Frequently, the classification results are obtained using quadratic discriminate functions with equal prior probabilities. That is, it is assumed that the probability of occurrence of corn is the same as the probability for grain sorghum and each of the other classes. Since the assumption of equal prior probabilities is not consistent with the known facts for agricultural land use, the work undertaken by SRS has employed unequal prior probabilities based on historical and current estimates of the fraction of land in each crop. The use of equal prior probabilities has been shown to lead to highly biased estimates for individual crop types. Tables 3 and 4 show the classification results for two adjacent ERTS frames for the fields used to derive the parameters needed in the quadratic discriminate functions. Tables 5 and 6 show the classification results for the a random sample of area segments including some from which the fields were selected to derive the parameters.

The results shown in Tables 3, 4, 5 and 6 are somewhat better than we have obtained in other study areas. We believe these results are more favorable due to three factors: (1) large rectangular fields, (2) relatively few crops being grown, and (3) the date of the imagery corresponded to a time during the season favorable for discriminating between the particular crops studied. The use of temporal overlays has resulted in improved accuracy in the classification for most study areas. The correlation between acreage data obtained for ground verification and pixels classified into crops for area sampling units were as follows:

Total acreage vs. total pixels	$r^2 = .88$
Pasture acreage vs. pasture pixels	$r^2 = .84$
Corn acreage vs. corn pixels	$r^2 = .62$
Grain sorghum acreage vs. grain sorghum pixels	$r^2 = .58$

When r^2 values of this magnitude are realized, remotely sensed data are beneficial in stratification, varying probabilities of selection, and in estimators using supplementary information.

In order to use remotely sensed data in these ways, the task of extracting information for area sampling units must be achieved. We have been able to devise a first generation system for doing this. The boundaries of the sampling units are digitized from existing maps - 7-1/2 minute quadrangle maps are the most accurate for small areas such

Table 3.--Classification matrix for September 21, 1972 imagery (MSS bands 4, 5, 6, 7) using quadratic discriminant functions with unequal prior probabilities in Kansas for fields used to derive parameters

Class	No. of sample points	Percent correct	Number of pixels classified into				
			Alfalfa	Pasture	Corn	Grain sorghum	Thresh-old
Alfalfa.....	63	100.0	63	0	0	0	0
Pasture.....	172	98.3	0	169	2	1	0
Corn.....	51	90.2	0	1	46	4	0
Grain sorghum...	78	69.2	0	10	14	54	0
Total.....	364		63	180	62	59	0

Overall performance = 91.2%

Table 4.--Classification matrix for September 22, 1972 imagery (MSS bands 4, 5, 6, 7) using quadratic discriminant functions with unequal prior probabilities in Kansas for fields used to derive parameters

Class	No. of sample points	Percent correct	Number of pixels classified into				
			Alfalfa	Pasture	Corn	Grain sorghum	Thresh-old
Alfalfa.....	73	84.6	66	12	0	0	0
Pasture.....	230	93.0	0	214	11	5	0
Corn.....	337	65.0	0	93	219	25	0
Grain sorghum...	177	63.9	3	34	18	122	0
Total.....	822		69	353	248	152	0

Overall performance = 75.5%

Table 5.--Classification matrix for September 21, 1972 imagery (MSS bands 4, 5, 6, 7) using quadratic discriminant functions with unequal prior probabilities in Kansas for a random sample of area segments for specific crops

Class	No. of sample points	Percent correct	Number of pixels classified into				
			Alfalfa	Pasture	Corn	Grain sorghum	Thresh-old
Alfalfa.....	43	93.0	40	2	0	1	0
Pasture.....	6261	95.0	23	5949	121	139	29
Corn.....	332	37.7	38	110	125	59	0
Grain sorghum...	508	64.8	38	77	60	329	4
Total.....	7144		139	6138	306	528	33

Overall performance = 90.2%

Table 6.--Classification matrix for September 22, 1972 imagery (MSS bands 4, 5, 6, 7) using quadratic discriminant functions with unequal prior probabilities in Kansas for a random sample of area segments for specific crops

Class	No. of sample points	Percent correct	Number of pixels classified into				
			Alfalfa	Pasture	Corn	Grain sorghum	Thresh-old
Alfalfa.....	287	56.4	162	57	12	23	0
Pasture.....	4975	90.6	19	4508	45	44	23
Corn.....	1698	40.8	1	684	693	174	0
Grain sorghum...	2869	55.3	89	300	357	1586	4
Total.....	9829		271	5549	1107	1827	27

Overall performance = 70.7%

as primary units or individual area segments. The boundary points are stated in laditude/longitude divisions which are used with registration points from ERTS to restate the sampling unit boundaries in terms of row and column numbers on digital tapes. Figure 2 shows a primary sampling unit for Milam County, Texas along with the summary of the classification results for this unit. Similar crop information was obtained for all 105 primary sampling units in this county. In order to obtain a measure of the maximum reduction possible through stratification, the primary units were assigned to strata based on the square root of the pixel count for cropland and the two principle individual crops. Four strata were used in each case based on dividing the square root of the largest value of the stratification variable by four to obtain four equal intervals from zero to this value. For the stratification variable total cropland pixels, the reductions in variance were 27 percent for cotton and 35 percent for sorghum. When the stratification was based on the individual crop, the reduction was 60 percent for cotton and 58 percent for sorghum. Since the variable cropland is likely to be fairly constant over years for the primary units considered here, it should provide a useful stratification variable for important crops. While the pixel count for individual crops results in even greater gain for special purpose crop surveys when current crop year information is available, these gains are likely to diminish if based on information from prior years because of fluctuations in acreages planted to individual crops over years. Under the assumption current year crop information on pixel counts will be available before crops are harvested, post-stratification of sampling units or the use of estimators based on auxiliary variables would seem to be more likely uses for remote sensing information leading to improved estimates of acreages for harvest. The use of remote sensing information to improve early season planted acreage figures seems less likely due to the necessity of obtaining information coinciding with a time during the crop cycle when discrimination is satisfactory.

Another analysis which was made considered equalizing the primary unit "size" in terms of pixel counts. This study was made for three variables: (1) total pixels, (2) agricultural land pixels, and (3) cropland pixels. Table 7 shows the coefficients of variation using these variables on several crops.

Table 7.--Coefficients of variation of equal sized units using three measures of size - Milam County, Texas

Crop	Variable			Original frame units
	Total pixels	Agricultural land pixels	Cropland pixels	
Cotton	94.1	76.9	73.5	103.6
Small grains	79.6	84.6	77.6	76.9
Peanuts	107.0	103.0	100.5	117.9
Pasture	34.6	11.0	55.3	60.6
Hay	51.2	46.3	32.7	60.6
Sorghum	105.3	85.8	63.1	106.1

This table covers more crops as well as pasture, and also indicates important gains in efficiencies can be achieved using remote sensing information for variables which tend to be fairly stable over time.

Conclusions

The use of low resolution sensors to provide auxiliary variables for geographic areas appears to be promising for the collection of agricultural data since the information may be used during the current year or at a future time without regard to: (1) the sensor ability to provide data at or near the optimum period in the crop cycle for classification purposes because of cloud cover, (2) the necessity of processing large amounts of data rapidly during a critical survey period for which the information is needed, (3) the availability of sufficient information for calculation of prior probabilities by crop types to insure estimators which are approximately unbiased, and (4) the necessity for classification error rates by crops to be smaller than sampling errors.

If prior probabilities and proper sensor timing can be secured, approximately unbiased estimates can be realized for crops whose distributions do not materially overlap other crops. Where high resolution sensors are available, the use of multivariate normal statistics with periodogram statistics in discriminate functions, and clustering methods in a sequential manner can lead to marked improvements in classification results without prior probabilities. The potential for remote sensing is quite encouraging in terms of the possibilities for reductions in variances in agricultural surveys. However, the costs of acquiring remote sensing information is still too great for most users to justify.

References

Gallant, A. R., Gerig, Thomas M., and Evans, J. W., (1974) Time Series Realizations Obtained According to An Experimental Design, Institute of Statistics, Mimeograph Series No. 913, Raleigh, N. C.

Gleason, C. and Hopkins, P., (1975) Optimum Classification and Scanning Procedure for Mature Oranges for Point by Point Classification, Research Report, SRS, USDA, Washington, D. C.

Huddleston, H., (1971) Use of Photography in Sampling for Number of Fruit Per Tree, Agricultural Economic Research, Vol. 23, No. 3, Page 63.

Nealon, John (1975) The System of Sequential Classification, Clustering, and Counting of Fruit Trees from Digitalized Photography, Research Report, SRS, USDA, Washington, D. C.

Rao, C. R. (1965) Linear Statistical Inference and Its Application, John Wiley & Sons, New York.

Zahn, C. T. (1971) Graph - Theoretical Methods for Detecting and Describing Gestalt Clusters, I.E.E.E. Transactions on Computers, Vol. C-20, No. 1, Page 68.

ABREGE

Le Service des Statistiques du Département de l'Agriculture des Etats-Unis a mis sur pied certaines méthodes d'utilisation de la perception à distance en vue d'améliorer la méthodologie utilisée pour effectuer des travaux d'enquête. Le présent rapport fait l'exposé des techniques de perception à distance qui ont été élaborées pour permettre d'utiliser des données obtenues à l'aide d'appareils de sondage électroniques ou des films pris par le personnel au sol ou à bord d'avions ou de satellites. Les renseignements résultant de la perception à distance ont été utilisés en conjonction avec des observations et autres sources de données de caractère conventionnel dans le cadre d'échantillonages doubles utilisés pour arriver à une estimation du nombre de fruits sur certains arbres, du nombre d'arbres dans une plantation et de la superficie plantée de certaines cultures. Toutes les techniques employées pour ce faire font usage des données obtenues par perception à distance sous forme numérique et font appel à l'analyse par ordinateur de bandes magnétiques. Les méthodes d'analyse comprennent les fonctions discriminantes, le sondage spectral et spatial par grappes et les statistiques à périodogrammes.