72-02

R-40

Paper A-5

(2)

CROP IDENTIFICATION AND ACREAGE MEASUREMENT UTILIZING ERTS IMAGERY

William H. Wigton and Donald H. Von Steen, Statistical Reporting Service, U.S. Department of Agriculture.

1972

ABSTRACT

The Statistical Reporting Service of the U.S. Department of Agriculture as a principle investigator for NASA, is evaluating ERTS-1 imagery as a potential tool for estimating crop acreage. The Statistical Reporting Service makes crop and livestock forecasts and estimates throughout the year, across the U.S. A main data source for the estimates is obtained by enumerating small land parcels that have been randomly selected from the total U.S. land area. These small parcels are being used as ground observations in this investigation. The test sites are located in Missouri, Kansas, Idaho, and South Dakota. The major crops of interest are wheat, cotton, corn, soybeans, sugar beets, potatoes, oats, alfalfa, and grain sorghum. Some of the crops are unique to a given site while others are common in two or three States. This provides an opprotunity to observe crops grown under different conditions. Results for the Missouri test site are presented in this report. Results of temporal overlays, unequal prior probabilities, and sample classifiers are discussed. The amount of improvement that each technique contributes is shown in terms of overall performance. The results show that useful information for making crop acreage estimates can be obtained from ERTS-1 data.

INTRODUCTION

SRS of the U.S. Department of Agriculture is the main fact-gathering agency of the USDA. The name of the agency has changed several times, but the objective of collecting and disseminating primary data on agriculture has remained the same for more than 100 years. Crop acreage and production as well as livestock, prices, labor, and farm expenditures are estimated.

Many of these estimates are generated from a general purpose land area sample survey conducted in June and based on 17,000 segments selected at random from the total U.S. land area. This is a sample stratified by States and within states by land use. Segments for a State are defined within each category of land use or stratum and a sample of these segments is selected. Stratification by land use has made it possible to sample more efficiently for all items because sample segments are allocated to each stratum individually. At the time of field enumeration, the interviewer must be able to identify the boundaries of the sample segment and collect information which applies to the land inside these boundaries. ERTS imagery may also be helpful in stratification and in the segment selection process; we have not used ERTS for these purposes yet, but plan to try this soon.

Keep in mind that we use these segments to generate livestock and price estimates as well as crop acreages, and for this reason, ERTS will not replace our present system for major items. Secondly, our estimates have sampling errors between 2 percent and 5 percent at the U.S. level, and between 5 and 10 percent at the state level for major commodities. We do not go much below the state level for our probability survey since the sample was not designed to provide estimates below the state level.

PROCEDURES

Twenty-nine segments of approximately one square mile size were located in two ERTS frames covering most of Crop Reporting Districe No. 9 in Southeast Missouri. The segments are located over a 10,000 square mile area. Information on the crop and acreage of each field was obtained by SRS enumerators during the summer of 1972; this data has been used for training the classifier and testing its performance. ERTS data from three dates was included in the analysis. Data collected September 14 and October 2, 1972 was registered (overlaid) to data collected August 26, 1973. The temporal overlay alleviated the necessity of locating fields in three different data sets, as well as permitted a test of the utility of temporal data in the classification.

The ERTS data was also geometrically corrected to facilitate locating the coordinates of segments and fields. In the geometric correction process the MSS data is rotated, deskewed, and scaled to 1/24,000 scale. The geometrically corrected data was overlaid on 1/24,000 scale topographic maps on which the segments had been outlined. The individual segments were then classified (clustered) using the ono-supervised classifier in LARSYS. Field coordinates were located on the map output from this classification. Final classifications were carried out using the supervised classifier in LARSYS.[1]

RESULTS

The results are presented in the form of a classification matrix. Table 1 shows the classification results obtained when using quadratic discriminant functions with equal prior probabilities. That is, it is assumed that the probability of occurence of corn is the same as the probability for cotton, and so forth. Because of the small size of the data set the whole data set was used in training the classifier. This is a nine channel classification with data from three ERTS passes. The four major classes, cotton, corn, soybeans, and grass were classified 74, 59, 40, and 57 percent correctly, respectively. Overall performance was 59 percent.

[1] We gratefully acknowledge the assistance of Dr. Marvin Bauer, Purdue, LARS, for technical assistance and in reviewing this paper. Also, LARS for giving support in providing the temporal overlays for this analysis. However, the authors accept responsibility for the suitability of the methods employed and interpretation of the results.

The assumption of equal prior probabilities is many times not valid, but is frequently used because of lack of information. The prior probabilities used in this study came from an earlier survey, the June 1972 Enumerative Survey. Other sources of prior probability information are historic data, for example, last year's farm census. Classification results using unequal prior probabilities are shown in Table 2. Comparing the results in Table 1 to those in Table 2 it is seen that the overall performance has been increased from 59 to 71 percent; and secondly, that the total number of points classified into each class is much closer to the actual number of points present. For example from Table 2, the total number of points classified as cotton is 906 which is considerably closer to 927, the actual number present. The total number of corn points, 43 is rather close to the actual 58 present. For soybeans, the total of 866, is very close to the actual 852 present. Two hundred seventy-seven (277) points were classified as grass compared to 240 actual points of this crop. Further, the statistical properties of estimates made on this basis are better since, if the assumption of normality for the data set is correct, and the prior probabilities are correct, we obtain unbiased estimates.

Most classifications reported by other researchers have not used prior probabilities. While the overall error rate reported here is higher than reported by some researchers, this study was based on a statistical sampling of the entire land area in the study areas rather than on purposely selected test sites.

Table 3 shows results of using a sample classifier rather than a point classifier used in the above work. In a point classifier system each point in a field can be assigned to any of the groups. With the sample classifier all points in the field are assigned to the same class or crop. One drawback to this procedure is that there were a large number of fields that were not classified because the technique requires p+1 data points in order to form the statistics necessary to assign it to a crop (where p is the length of the vector of measurements). However, if enough points are present, classification performance has generally been found to be better than for the point classifier.

In the work we have done in Missouri using the sample classifier, about 40 percent of the fields were not classified because the required number of points for the classifier (10 in this particular case) exceeded the number of points present within the defined fields. Of the total number of fields 33 percent were correctly identified. Considering only those fields which were classified, 54 percent were classified correctly.

In Missouri 71 percent of the fields were less than 20 acres, but account for 32 percent of the total area. In our Kansas site, 20 percent of the fields were less than 20 acres, but account for only 1.5 percent of the total land areas. In South

Dakota, 40 percent of the fields were less than 20 acres, and account for 15 percent of the area. In Idaho, 74 percent of the fields were less than 20 acres, and account for 25 percent of the area. If 20 acres is a critical field size for the classifier, we would expect to do well in making acreage estimates, in Kansas, but in Missouri only a little more than 50 percent of the acreage would be accounted for.

Next, the information gained from the temporal overlay is evaluated. In Table 4 classification results for single dates are compared to the multitemporal classifications already presented. The overall classification performance was improved about 10 percent by the addition of temporal data with even greater improvement for several of the individual classes.

DISCUSSION

The results presented do not show the classification accuracy to be as high as that found by other investigators. The lower performance level is premarily attributed to the greater variation in crops, soils, and weather over a 10,000 square mile area than is found over smaller areas. And, secondly to the kind of crops which were being discriminated. Still, the classifications contain enough information to be useful in estimating crop acreages over large areas, particularly if regression or some other technique is used to improve the estimate.

A regression estimator can be used to reduce the variance of the estimate. For example, if a large area is classified and there is an r^2 of .50 between the discriminant function classification and what the ground acreage data shows. We can adjust our area sample estimates by the complete classified data and obtain a reduced variance of $\Sigma y^2(1-r^2)/n(n-2)$ where r^2 is the correlation coefficient squared. The estimate of the variance of the comparable statistic without using ERTS data is $\Sigma y^2/n(n-1)$ which would be nearly twice as large when $r^2 = .50$.

If we were to classify a sample of points we would have a double sample and the variance would be:

$$\frac{\Sigma y^2 (1-r^2)}{n(n-2)} + \frac{\Sigma y^2 (r^2)}{n\,m}$$

where n = the sample size from JES and m = the sample size from ERTS.

Another possible approach is to consider the classification a mixture problem. An unbiased estimate of the classification matrix based on a large sample and a count of the number of points of each crop from a second data set such as ground observations are needed. The point count for crop A contains not only crop A, but also potentially all the other crops in the classification. On the basis of the classification matrix the data can be unmixed as follows: The transpose of the classification matrix is multiplied by the vector of the total number of points in each crop. These estimates are unbiased if the classification matrix is correct.

To illustrate, assume the classification matrix below:

	Classification	
Ground Obs.	Corn	Soybeans
Corn	80%	20%
Soybeans	30%	70%

$$CM = \begin{bmatrix} .80 & .20 \\ .30 & .70 \end{bmatrix} \qquad (CM)' = \begin{bmatrix} .80 & .30 \\ .20 & .70 \end{bmatrix}$$

$$(CM)'^{-1} = \begin{bmatrix} \frac{7}{5} & \frac{3}{5} \\ -\frac{2}{5} & \frac{8}{5} \end{bmatrix}.$$

Assume that we have classified a county based on training sample and that 10,000 points were classified into corn and 5,000 points were classified as soybeans. We know that in the 10,000 points classified as corn some may be soybeans and of the 5,000 soybean points some may be corn. To unmix or obtain unbiased estimates, we use this properties of the classification matrix, multiply $(CM')^{-1}$ times $\binom{10,000}{5,000}$

$$\begin{bmatrix} \frac{7}{5} & -\frac{3}{5} \\ -\frac{2}{5} & \frac{8}{5} \end{bmatrix} \begin{matrix} 10,000 \\ \\ 5,000 \end{matrix} = \begin{bmatrix} 11,000 \text{ corn points} \\ \\ 4,000 \text{ soybeans points} \end{bmatrix}$$

If each point was an acre then the estimates would be 11,000 acres of corn and 4,000 acres of soybeans in the county.

This procedure "improves" the classifications of remotely sensed data.