United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Research and
Applications
Division

SRB Research Report
Number SRB-91-02

February 1991

COMPARISON OF SENSORS FOR CORN AND SOYBEAN PLANTED AREA ESTIMATION

Michael Bellow

COMPARISON OF SENSORS FOR CORN AND SOYBEAN PLANTED AREA ESTIMATION, by Michael E. Bellow, Research and Applications Division, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C. 20250, February, 1991. NASS Research Report No. SRB-91-02.

ABSTRACT

This report compares the effectiveness of Landsat Thematic Mapper (TM) and French SPOT Multispectral satellite data as a supplement to ground survey data for estimation of corn and soybean planted area. Reference data from USDA's 1988 June Agricultural Survey were used in the estimation process and to check results. The survey data covered a sample of 30 segments in western Iowa. TM and SPOT scenes of the region, imaged during late July of 1988, were utilized. The ground and satellite data were processed through USDA's PEDITOR software system. For both TM and SPOT, each pixel within the sample segments was classified to a specific ground cover based on previously computed spectral signatures. Since the true cover for each pixel was known from the ground data, classification accuracy could be assessed. Statistical criteria used to evaluate sensor performance included percentage of pixels correctly classified, commission error, and regression determination coefficient. For both crops of interest, the TM data produced more accurate acreage estimates than the SPOT data. The entire TM scene was classified in order to generate region level estimates of crop area. These estimates were compared with the corresponding direct expansion estimates computed from survey data alone. Battese-Fuller estimates of crop acreage on a county basis were computed and compared with the official county estimates.

Key Words: Landsat, TM, SPOT, pixel, classification, regression

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The author would like to thank Don Allen, Paul Cook, Sherm Winings, and Mike Craig for their guidance and advice on the preparation of this report. In addition, he wishes to thank Martin Ozga and Bob Losa for programming assistance, Lillian Schwartz for graphical support, and the Iowa State Statistical Office for supplying important data and materials.

INTRODUCTION

The National Agricultural Statistics Service (NASS) used the Landsat Multispectral Scanner (MSS) for the Agency's operational crop area estimation program from 1980 to 1987 [1]. Since this sensor would not be on future Landsat satellites and the current ones were not expected to stay in operation for much longer, NASS decided in 1987 to pursue a research program. The objective of the research program was to compare the utility of two available sensors that could replace MSS. The two sensors are the Landsat Thematic Mapper (TM) and the French SPOT multispectral scanner. Estimation accuracy and cost efficiency were the main criteria for comparing the sensors. This report evaluates the quality of TM and SPOT based estimates of corn and soybean planted acreage, using ground and satellite data from a region in western Iowa.

A pixel (picture element) is the basic unit of a remotely sensed image. The pixel is represented by a d-component vector of reflectance measurements, where d is the number of channels or spectral bands. The spatial resolution of a sensor is the length of the sensor's pixel. The Landsat TM sensor features seven channels with a spatial resolution of 30 meters, while the SPOT sensor has three channels and a resolution of 20 meters. By comparison, the Landsat MSS sensor has four spectral bands with a spatial resolution of 80 meters.

In the NASS operational remote sensing program, MSS data were processed and combined with ground reference data from the area portion of the June Agricultural Survey (JAS) to produce crop acreage estimates. The PEDITOR software system processed the data. All pixels within a satellite scene were classified to a specific crop or ground cover. The system used linear regression to relate JAS reported acres for a given crop to the classified pixel counts for that crop, and to generate the Landsat based acreage estimates.

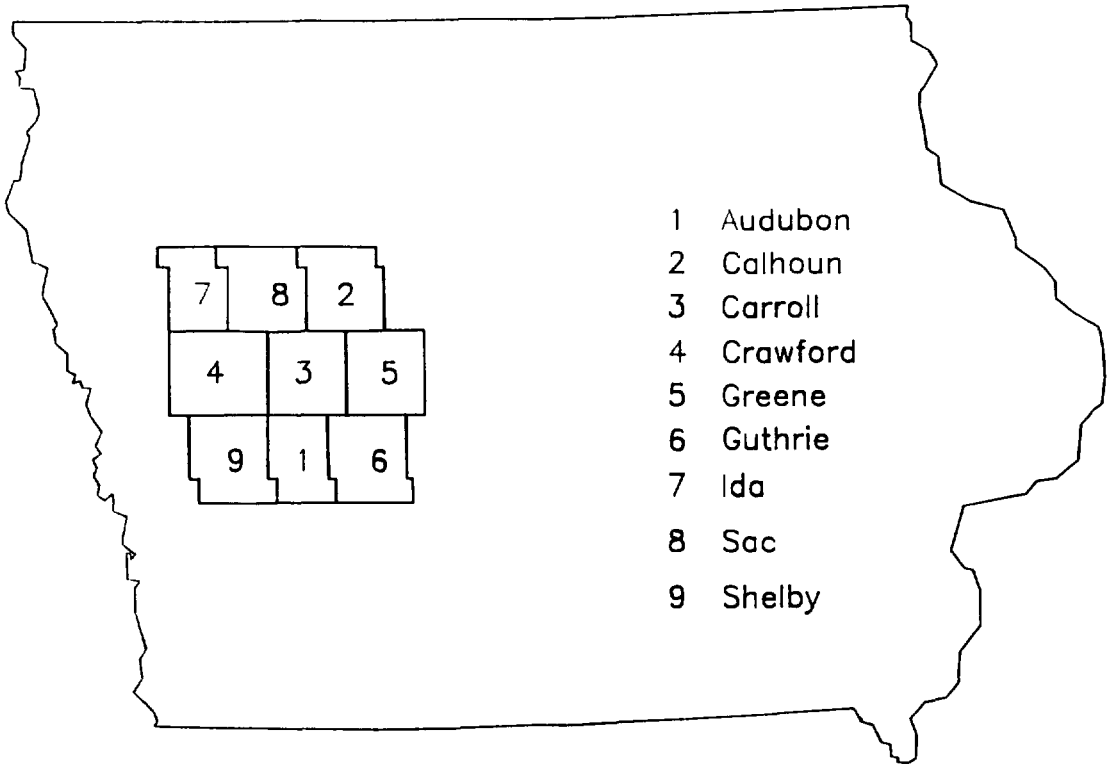
This study uses the statistical efficiency of the regression estimator as the key criterion for comparing the performance of the TM and SPOT sensors. However, other remote sensing studies often use percent correct classification and commission errors. In order to produce good results, the regression estimator requires accurate ground reference data.

The TM data used for this study were applied to a full scene classification to obtain large scale regression estimates of corn and soybean planted acreage in the region of interest. The large scale acreage estimates were compared with the corresponding JAS direct expansion estimates that use only ground data. In addition, the Battese-Fuller method was used to compute TM based county level estimates of acreage for the two crops. The study compared the estimates for each county with the official 1988 county estimates issued by the Iowa State Statistical Office.

RESEARCH AREA

The research site for this study was a nine county region in western Iowa, shown in Figure 1. Corn and soybeans are the predominant crops in this region. Ground reference data from the area frame portion of the 1988 JAS were used. The survey data covered a statistical sample of 30 land segments.

Figure 1: Map of Iowa Showing Research Site



The 1988 Iowa area sampling frame consisted of eleven land use strata. Nine of the strata comprised a geographic subdivision of all agricultural land in the state, while the other two covered agri-urban and residential-commercial areas. Stratum 14 included the agricultural land in the research area. Of the 30 segments used for the study, 28 came from stratum 14 and the other two from stratum 30 (agri-urban). There were no sampled segments from stratum 40 (residential-commercial). Some prominent covers in the region other than corn and soybeans were pasture, oats, and alfalfa.

The region was covered by one TM scene and four SPOT scenes. The overpass dates were July 25, 1988 for the TM scene and July 31, 1988 for the SPOT scenes. Cloud cover was insignificant for all of the scenes. Four sample segments were completely within the TM scene but not within any of the SPOT scenes. Two other segments were completely within one of the SPOT scenes but not

the TM scene. These six segments, which included one from stratum 30 (agri-urban), were not used for the sensor comparison. The remaining agri-urban segment contained no corn area and very little soybean area according to the ground reference data. This segment was included in the training process (supervised clustering) but excluded from classification and statistical analysis. The remaining 23 segments, all in stratum 14, were classified into categories. Use of this categorized data with the ground data permitted the establishment of a regression relationship between the ground and satellite data. Thus the small scale classifications for TM and SPOT used the same ground area. All available spectral bands for each sensor were utilized.

PROCESSING

All data processing associated with remote sensing crop area estimation is performed using PEDITOR, a special purpose software system developed at NASS [2]. PEDITOR is written mainly in PASCAL and maintained on a MicroVax 3500 computer at NASS, with many modules that also run on IBM compatible personal computers. At the time this study was done, satellite scenes were stored on tapes at the CRAY X-MP supercomputer facility operated by Boeing Corporation in Seattle, Washington. Portions of those scenes could be retrieved and transferred to the MicroVax in the form of a multiwindow file. The CRAY supercomputer was also used for large scale classification, estimation, and aggregation.

The required processing steps (up to classification) are discussed in Appendix A. Following classification, the options available to the user are small scale estimation, large scale estimation, and county estimation. Appendices B, C, and D describe these procedures. For the current study, large scale estimation was done only for TM. Small scale estimation was sufficient for the sensor comparison because measures of estimation accuracy could be obtained from processing at the sample level.

The ground reference data for this study required both internal and external editing before further processing. Internal editing detected and corrected errors within the ground reference data. External editing detected discrepancies between the ground data and registered satellite imagery that required corrective action. Some fields were labelled as bad and removed from the training data set. Fields having large discrepancies between field and planted size, field and harvested size, or planted and harvested size fell into this category. Fields for which the reported (survey) acreage differed too much from the digitized acreage were also labelled as bad.

In selecting TM or SPOT pixels for training, all covers containing fewer than 5 percent of the total number of pixels were combined into one category, labelled 'other'. The covers

lumped together in the 'other' category were farmstead, alfalfa, oats, idle crop, waste, woods, crop pasture, and water. This resulted in a total of four covers for the subsequent classification process: corn, soybeans, permanent pasture, and other.

Small scale classification was done in stratum 14 only, using both equal and unequal prior probabilities for the four covers (see Appendix A). With unequal priors, the probability for each cover was defined to be the percentage of total pixels in the appropriate packed file (TM or SPOT) belonging to that cover. The packed files used to calculate the priors were the original versions that included the outlier pixels not used for training.

SMALL SCALE ESTIMATION RESULTS

Small scale estimation refers to estimates and performance measures obtained from small scale (segment only) classification. Regression is used to relate the classified pixel counts within segments to the corresponding ground reference data from the JAS. The regression methodology and performance measures used to compare TM and SPOT are described in Appendix B.

Small scale estimation was done only in stratum 14, using 23 segments as discussed in Section 2. Separate estimates for each sensor were computed using the classification results obtained with unequal priors and equal priors. Table 1 shows, for both corn and soybeans, the values of the regression determination coefficient (R^2), relative efficiency, percent correct, and commission error for TM and SPOT. The performance measures used about the same ground area as did the training samples, so the results for both sensors probably display a higher level of accuracy than would be obtained from classifying other areas. Training pixel counts, prior probabilities, number of categories for each cover, and confusion matrices showing the number of pixels from each cover classified to each cover can be found in Appendix E, Tables A1 and A2.

Table 1 shows that for both corn and soybeans, higher values of R^2 occurred for TM than for SPOT. The discrepancies ranged from 0.051 for the soybeans/equal priors case to 0.142 for the corn/unequal priors case. In addition, percent correct was higher for TM than for SPOT in every case, while the commission error was lower. The table also shows that R^2 was usually higher with unequal priors than equal priors, with the SPOT results for soybeans being the exception.

For both TM and SPOT, the R^2 values for soybeans were higher than the corresponding ones for corn, while the commission errors for soybeans tended to be lower than those for corn. Conversely, corn showed higher values of percent correct than did soybeans. This illustrates that the three metrics are measuring different aspects of estimator efficiency.

Table 1: TM and SPOT Efficiency Comparison

	TM Priors		Spot Priors	
Description	Unequal	Equal	Unequal	Equal
Corn R^2	.928	.825	.786	.727
Soybeans R^2	.943	.913	.853	.862
Corn Rel. Efficiency	12.65	5.21	4.25	3.33
Soybeans Rel. Efficiency	15.90	10.44	6.18	6.61
Corn Percent Correct	85.67	88.23	84.27	81.47
Soybeans Percent Correct	84.13	77.78	74.20	72.35
Corn Commission Error	19.83	30.08	29.08	33.62
Soybeans Commission Error	20.01	24.27	25.93	31.38

The paired sample t-test for equality of means of the absolute residuals provides a formal method for assessing whether or not the TM sensor produced a significantly better regression fit than the SPOT sensor. A previous study on winter wheat used this test for the same purpose [3]. The test was performed for the 'with priors' case for each crop. The hypotheses are as follows:

$$H_0: \mu_{TM} = \mu_{SPOT}$$

$$H_1: \mu_{TM} < \mu_{SPOT}$$

where μ_{TM} and μ_{SPOT} are the means of the absolute residual distributions for TM and SPOT. The formula for the test statistic t^* is given in Appendix B. Assuming normality of the data, t^* has a t-distribution with 22 degrees of freedom under H_0.

The computed values of t^* were 2.182 for corn and 1.985 for soybeans. The null hypothesis of no significant difference can be rejected at the 2.5 percent level for corn and at the 5 percent level for soybeans.

LARGE SCALE ESTIMATION RESULTS

A large scale classification was performed on the TM data over the nine county region. Each pixel within the TM scene was classified to a specific cover. The classification program used the maximum likelihood method with the same discriminant functions as for small scale classification. Once the classification was complete, large scale regression crop acreage estimates were computed. Appendix C gives formulas for direct expansion, proration, and regression estimates as well as their sample variances.

The two analysis districts comprising the study area were labelled AD1 and ADDE. District AD1 contained all area covered

by the TM scene, and ADDE all other area. Stratum 14 was divided into two subsets, corresponding to AD1 (regression) and ADDE (proration). Regression was not performed in stratum 30 since it contained only two sample segments. Stratum 40 was not needed as it contained no segments in the region. The regression coefficients computed previously for small scale estimation were used again. They were determined based on the 23 segments in stratum 14 covered by both TM and SPOT imagery. To provide a valid comparison between the estimators, the three segments from analysis district AD1 in stratum 14 not used in the regression were also not used in the direct expansion estimates.

Table 2 gives the large scale estimation results for corn and soybeans. For comparison purposes, state office estimates of corn and soybean acreage for the region are also shown. These were computed by summing the official 1988 county estimates for the nine counties comprising the region (see Section 6). The official county estimates were issued by the Iowa State Statistical Office (SSO). Percent error values for the overall direct expansion and regression estimates were computed using the state office figures to represent "truth". Table A3 of Appendix E shows the sample and population data used to compute the regression estimates in stratum 14.

Table 2: Large Scale Acreage Estimates

	Corn	Soybeans
State Office Estimate	1,147,000	958,300
Direct Expansion Estimate	1,246,925	885,666
Regression Estimate	1,097,440	881,523
Direct Expansion Std. Dev.	70,884.5	93,189.0
Regression Std. Dev.	20,398.1	33,478.8
Direct Expansion C.V. (%)	5.68	10.52
Regression C.V. (%)	1.86	3.80
Direct Expansion Percent Error	8.71	7.58
Regression Percent Error	4.32	8.01

Table 2 shows that use of the regression estimator causes a significant reduction in overall variance over the direct expansion estimator for both crops. This was previously shown for the regression subset via the relative efficiency figures in Table 1. TM data reduced the overall coefficient of variation (C.V.) from 5.68 to 1.86 for corn and from 10.52 to 3.8 for soybeans. The percent error from the state office estimate was much lower with regression than direct expansion for corn, but was higher for soybeans. This was most likely due to the large difference between the sample mean acreages for soybeans in the AD1 and ADDE subsets of stratum 14, which caused the proration estimate to be too low.

COUNTY ESTIMATION RESULTS

Estimation of crop acreage at the county level is a topic of great interest to NASS. County estimates based on list frame control data and supplemental surveys are published periodically. The major obstacle to obtaining good county estimates from JAS survey data is the fact that a given county usually contains very few sample segments.

Current NASS county estimation research focuses on using additional available information such as total farm acreage and the previous year's county estimates [4]. TM data represents another possible supplementary data source. The potential for improved estimation accuracy with TM is based on the fact that with adequate coverage, all or most of the area within a county can be classified.

Several county estimation methods utilizing satellite data have been investigated by NASS [5]. Of these, the Battese-Fuller family of estimators [6,7] gives the best performance. Appendix D provides a description of Battese-Fuller estimation, as well as proration estimation for counties.

TM based estimates of corn and soybean acreage were computed for all nine counties in the study area. The computations used the large scale classification discussed in Section 5, with the same subdivision of the region into stratum/analysis district combinations. Battese-Fuller estimation was applied within the regression subset of stratum 14. Three counties (Calhoun, Crawford, and Ida) were not completely within the TM scene. Table A4 shows the number of frame units and sample segments in each county, broken down by stratum. Proration was used within stratum 14 for the parts of counties outside the scene, and in stratum 30 for all nine counties.

Table 3 gives the computed (TM based) county estimates by stratum and estimation method, and the official county estimates issued by the Iowa SSO. Table 4 shows the estimated standard deviations and coefficients of variation of the computed estimates, and the percent error between the computed and official estimates.

The tables show that the computed county estimates for corn were more efficient overall than those for soybeans. For eight of the nine counties, the C.V. for corn was less than 3 percent. No county had a C.V. of less than 3 percent for soybeans. The percent error ranged from 0.4 to 16.4 for corn, and from 2.2 to 29.6 for soybeans. For each crop, the computed acreage estimate was lower than the official estimate in seven of the nine counties.

Ida County showed much higher percent error values than the other counties. This was because a large portion of that county was outside the TM scene. The proration component of the

estimate, based on only two sample segments in stratum 14, was much too low for each crop. Table A5 in Appendix E gives the county estimates by stratum and estimation method.

The superior results for corn agree with the large scale estimation results presented earlier. Further investigation is needed to evaluate more fully the ability of TM data to improve county crop area estimation.

Table 3: County Estimates

	Corn		Soybeans	
County	Official	Computed	Official	Computed
Audubon	100,000	89,051	70,700	72,254
Calhoun	133,000	131,864	150,000	141,447
Carroll	141,000	140,453	117,000	108,656
Crawford	147,000	150,062	106,000	98,861
Greene	125,000	119,345	143,000	118,179
Guthrie	98,000	94,515	77,500	86,223
Ida	112,000	93,675	75,200	52,972
Sac	136,000	137,569	124,000	112,788
Shelby	155,000	140,906	94,900	90,143
Total	1,147,000	1,097,440	958,300	881,523

Table 4: Efficiency of County Estimates

	Corn			Soybeans		
County	Std. Dev.	% C.V.	% Error	Std. Dev.	% C.V.	% Error
Audubon	2,267.8	2.5	10.9	3,268.2	4.5	2.2
Calhoun	2,423.4	1.8	0.9	4,474.3	3.2	5.7
Carroll	2,519.2	1.8	0.4	5,261.2	4.8	7.1
Crawford	3,203.9	2.1	2.1	10,596.1	10.7	6.7
Greene	2,809.8	2.4	4.5	4,196.5	3.6	17.4
Guthrie	4,148.9	4.4	3.6	5,579.1	6.5	11.3
Ida	1,050.3	1.1	16.4	10,219.8	19.3	29.6
Sac	2,616.9	1.9	1.2	4,042.7	3.6	9.0
Shelby	2,533.4	1.8	9.1	4,260.4	4.7	5.0

CONCLUSIONS

The sensor comparison has provided strong evidence that TM data is preferable to SPOT data for estimating corn and soybean planted area. The use of prior cover probabilities appears to improve classification efficiency.

The large scale estimation results further illustrate the ability of TM data to improve survey based estimates. Overall, the corn estimates were more improved than those for soybeans. The county estimation results show the potential of TM in this important application.

In addition to the results presented in this report, the TM and SPOT systems have recently been evaluated for estimation of wheat, dry beans, rice, and cotton in several regions of the

country [3,8,9]. In all cases, TM produced more accurate estimates than SPOT.

The superior ground resolution of SPOT means that it may be the most useful sensor for land use mapping. However, by providing more spectral information, TM has become the preferred sensor for crop area estimation.

REFERENCES

[1] J.D. Allen and G.A. Hanuschak, "The Remote Sensing Applications Program of the National Agricultural Statistics Service: 1980-1987," U.S. Department of Agriculture, NASS Staff Report No. SRB-88-08, Aug. 1988.

[2] G. Angelici, R. Slye, M. Ozga, and P. Ritter, "PEDITOR - A Portable Image Processing System," in Proceedings of the IGARSS '86 Symposium, Zurich, Switzerland, Sept. 8-11, 1986, pp. 265-269.

[3] J.M. Harris, S.B. Winings, and M.S. Saffell, "Remote Sensor Comparison for Crop Area Estimation," in Proceedings of the IGARSS '89 Symposium, Vancouver, Canada, July 10-14, 1989, pp. 1860-1863.

[4] E.A. Stasny, P.K. Goel, and D.J. Rumsey, "County Estimates of Wheat Production," Ohio State University Technical Report, in preparation.

[5] G. Walker and R. Sigman, "The Use of LANDSAT for County Estimates of Crop Areas - Evaluation of the Huddleston-Ray and Battese-Fuller Estimators," U.S. Department of Agriculture, SRS Staff Report No. AGES 820909, Sept. 1982.

[6] W.A. Fuller and G.E. Battese, "Transformations for Estimation of Linear Models with Nested-Error Structure," Journal of the American Statistical Association, vol. 68, no. 343, pp. 626-632, Sept. 1973.

[7] G.E. Battese, R.M. Harter, and W.A. Fuller, "An Error-Components Model for Prediction of County Crop Areas using Survey and Satellite Data," Journal of the American Statistical Association, vol. 83, no. 401, pp. 28-36, March 1988.

[8] J.D. Allen, "Remote Sensor Comparison for Crop Area Estimation Using Multitemporal Data," U.S. Department of Agriculture, NASS Staff Report No. SRB-90-03, March 1990.

[9] C.L. Stup and J.D. Allen, "The Construction of a Dry Bean Area Sampling Frame in Michigan," U.S. Department of Agriculture, NASS Staff Report No. SRB-90-06, May 1990.

[10] G.H. Ball and D.J. Hall, "A Clustering Technique for Summarizing Multivariate Data," Behavioral Science, vol. 12, pp. 153-155, March 1967.

[11] P.H. Swain, "Pattern Recognition: A Basis for Remote Sensing Data Analysis," Information Note 111572 (1973), Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana.

[12] M.E. Bellow and M. Ozga, "Evaluation of Clustering Techniques for Crop Area Estimation using Remotely Sensed Data," in preparation.

[13] R.K. Lennington and M.E. Rassbach, "CLASSY - An Adaptive Maximum Likelihood Clustering Algorithm," in Proceedings of the Ninth Annual Meeting of the Classification Society (North American Branch), Clemson, South Carolina, May 21-23, 1978.

[14] Johnson, R.A. and Wichern, D.W., Applied Multivariate Statistical Analysis, Englewood Cliffs, New Jersey: Prentice Hall, 1988, ch. 11, pp. 501-513.

APPENDIX A: REMOTE SENSING PROCESSING STEPS

During the June Agricultural Survey, enumerators draw off all field boundaries within segments onto aerial photographs. These boundaries are later transferred to digital form. At the state offices, questionnaire data from the survey are key-entered, edited, and transmitted to NASS Headquarters. The Remote Sensing Section's cartographic unit registers the JAS photographs and satellite scenes to a map base in latitude/longitude coordinates. This allows the remote sensing analyst to identify and manipulate pixels corresponding in location to the JAS fields. The analyst then selects pixels to be used for training and creates a packed file containing only those pixels. A boundary pixel is one that "touches" the segment border or the within segment border between two fields. Since reflectance values of boundary pixels are assumed to represent a mixture of covers on either side of the boundary, these pixels are generally excluded from the packed file. The analyst can apply a clipping algorithm based on principal components to remove outlier pixels, i.e. those whose multidimensional reflectance vectors are too isolated from the others.

The next step is the training process, which performs supervised clustering on the sampled satellite data. Pixels in the packed file belonging to a specific cover are clustered to produce signatures. Signatures are discriminant functions defined by mean vectors and covariance matrices describing the multivariate normal distributions assumed to model reflectance patterns. The collection of these statistics for all covers in a satellite scene constitutes the scene classifier. The clustering program used in this study implements a modified version of the Isodata algorithm of Ball and Hall [10]. It involves repeatedly assigning pixels to moving cluster centers based on the Euclidean distances between pixel reflectance vectors and the centers. The algorithm periodically merges cluster pairs whose Swain-Fu distance is sufficiently small. Swain-Fu distance is a measure of intercluster separation that takes into account the covariance structure of the clusters [11]. If a cluster displays excessive heterogeneity as measured by the largest eigenvalue of its sample covariance matrix, then it can be split into two subclusters [12]. The number of clusters in the final output of the program is generally not known in advance. An alternate clustering program, known as CLASSY, is also available for this task [12,13].

Once clustering has been performed for each cover, another PEDITOR program allows the analyst to combine the clusters into one large file containing all required statistical information. Options exist for editing this statistics file via deletion of clusters based on certain criteria. However, recent improvements to the clustering programs may make this process less important. The final statistics file contains the defining information for all remaining categories (clusters). Each category is assigned a label corresponding to a cover from the ground data. Unequal

prior probabilities can be assigned to the categories based upon available information on relative acreage of the different covers in the region of interest. This information may come from a previous survey, the current ground reference data, or other sources. Each cover is assigned a prior probability reflecting its approximate percentage of the land area. If unequal priors are not assigned, then the prior probability for each cover is assumed to be the reciprocal of the number of covers (equal priors). In either case, the prior probability for each category within a given cover is computed by multiplying the prior probability for the cover by the ratio of the number of pixels belonging to that category to the total number of pixels in all categories associated with the cover. The intent of using unequal priors is to improve the accuracy of the subsequent classification process.

After a final statistics file has been created, classification can begin. The classification process uses a maximum likelihood rule that assumes multivariate normality [14]. For each pixel to be classified, quadratic discriminant scores are computed for all categories. The scores for a pixel having reflectance vector **z** are given by:

$$d_i^Q(\mathbf{z}) = -0.5\ \ln[\det(S_i)] - 0.5(\mathbf{z}-\bar{\mathbf{z}}_i)^T S_i^{-1}(\mathbf{z}-\bar{\mathbf{z}}_i) + \ln p_i\ ,\ i=1,\ldots,c$$

where:

c = number of categories

$\bar{\mathbf{z}}_i$ = mean reflectance vector for category i

S_i = sample covariance matrix for category i

p_i = prior probability for category i

Small scale classification assigns covers only to those pixels identified with the JAS sample segments, while large scale classification operates on all pixels within a TM or SPOT scene. Each pixel is assigned to the category for which its discriminant score is highest. For each segment, the pixel counts are summed over categories within covers to obtain the number of pixels classified to each cover. By summing these counts over segments, the analyst can determine the overall number of pixels classified to each cover.

APPENDIX B: SMALL SCALE ESTIMATION

Within the sample segments for a given stratum, the estimation procedure uses regression methodology to relate classified pixel counts to the ground reference data. Counts of pixels within each sample segment classified to a specific crop are regressed against the corresponding crop acreage values from the JAS enumeration. A separate first order model is applied for each stratum where regression is used. Assume that for a given stratum h, segments $j=1,\ldots,n_h$ are used for regression. The model is expressed as:

$$y_{hj} = \beta_{h0} + \beta_{h1}x_{hj} \ , \ j=1,\ldots,n_h$$

where:

y_{hj} = reported acres of crop in segment j of stratum h

x_{hj} = number of pixels classified to crop in segment j of stratum h

β_{h0}, β_{h1} = regression coefficients for stratum h

The regression parameters are estimated using the standard least squares formulas:

$$b_{h1} = \sum_{j=1}^{n_h}(x_{hj}-\bar{x}_h)(y_{hj}-\bar{y}_h) / \sum_{j=1}^{n_h}(x_{hj}-\bar{x}_h)^2$$

$$b_{h0} = \bar{y}_h - b_{h1}\bar{x}_h$$

where:

$$\bar{y}_h = (1/n_h)\sum_{j=1}^{n_h} y_{hj}$$

$$\bar{x}_h = (1/n_h)\sum_{j=1}^{n_h} x_{hj}$$

This study uses several performance measures to evaluate classification and estimation accuracy. The most important to NASS is the regression determination coefficient:

$$R_h^2 = [\sum_{j=1}^{n_h}(x_{hj}-\bar{x}_h)(y_{hj}-\bar{y}_h)^2 / \sum_{j=1}^{n_h}(x_{hj}-\bar{x}_h)^2 \sum_{j=1}^{n_h}(y_{hj}-\bar{y}_h)^2$$

R_h^2 is the square of the correlation coefficient between the independent and dependent variables. It measures the goodness of fit of the regression equation. Closely related is relative efficiency (R.E.), a measure of the effectiveness of satellite

data in improving upon the JAS estimates. The relative efficiency is the ratio of the variance of the direct expansion (JAS) estimate to the variance of the regression (satellite based) estimate. Equivalently, R.E. is the factor by which the JAS sample size would have to be increased in order to produce a direct expansion estimate with the same precision as the regression estimate. For a single stratum h, the relative efficiency can be computed directly from the determination coefficient:

$$R.E. = (n_h-3)/(n_h-1)(1-R_h^2)$$

Two other measures, percent correct and commission error, do not require the regression data for their computation. Percent correct is the percent of pixels reported for a specific crop that were classified to that crop. Commission error is the percent of those pixels classified to a crop that actually belong to a different cover according to the ground data. Percent correct reflects a classifier's ability to identify correctly pixels belonging to a crop of interest, while commission error measures its ability to avoid labelling to the crop of interest pixels belonging to other covers.

The following is a description of the paired sample t-test used in Section 4. For each segment j in stratum h, let $e_j^{(TM)}$ and $e_j^{(SPOT)}$ denote the regression residuals for TM and SPOT. The pairwise differences of the absolute residuals are:

$$D_j = |e_j^{(SPOT)}| - |e_j^{(TM)}|, \quad j=1,\ldots,n_h$$

The hypotheses given in Section 4 can be written as:

$$H_0: \mu_D = 0$$

$$H_1: \mu_D > 0$$

where:

$$\mu_D = \mu_{SPOT} - \mu_{TM}$$

is the mean of the distribution of the D_j's.

The test statistic is:

$$t^* = n_h^{1/2}(\bar{D} - \mu_D)/s_D$$

where:

$$s_D^2 = (1/n_h-1)\sum_{j=1}^{n_h}(D_j-\bar{D})^2$$

Assuming that the absolute residuals are normally distributed, t^* has a t-distribution with n_h-1 degrees of freedom under H_0.

APPENDIX C: LARGE SCALE ESTIMATION

Let h=1,...,L denote the land use strata represented in a given region. The direct expansion estimator of the total acreage Y for a crop of interest in the region is based on ground survey data only. It is given by:

$$\hat{Y}_{DE} = \sum_{h=1}^{L} N_h \bar{y}_h$$

where:

N_h = number of frame units in stratum h

n_h = number of sample segments in stratum h

y_{hj} = reported acreage of crop in segment j of stratum h

$$\bar{y}_h = (1/n_h) \sum_{j=1}^{n_h} y_{hj}$$

An estimator for the variance of $\hat{Y}_{DE}$ is:

$$v(\hat{Y}_{DE}) = \sum_{h=1}^{L} [N_h(N_h-n_h)/n_h(n_h-1)] \sum_{j=1}^{n_h} (y_{hj}-\bar{y}_h)^2$$

If satellite data are used for estimation, then each stratum represented in the region of interest can be further subdivided by analysis district. The analyst usually defines the analysis districts based on the extent of available cloud-free satellite imagery for the region. A separate estimate of crop acreage can be made for each stratum/analysis district combination (subset), using regression where feasible and proration elsewhere. For a regression estimator to be feasible, a subset must be covered by satellite data and contain a sufficient number of segments.

Suppose regression is to be performed in a subset A of stratum p, containing N_A frame units and n_A segments. For convenience, assume that frame units 1,...,N_A and segments 1,...,n_A of stratum p are the ones contained in A. The mean number of pixels per frame unit and pixels per segment classified to the crop of interest are:

$$\bar{x}_A = (1/N_A) \sum_{i=1}^{N_A} x_{pi}$$

$$\bar{x}_A = (1/n_A) \sum_{j=1}^{n_A} x_{pj}$$

where:

X_{pi} = number of pixels classified to crop in frame unit i of stratum p

x_{pj} = number of pixels classified to crop in segment j of stratum p

The regression estimator of the total crop acreage in A is:

$$\hat{Y}_{pA}^{(reg)} = N_A[\bar{y}_A + b_A(\bar{X}_A - \bar{x}_A)]$$

where:

$$\bar{y}_A = (1/n_A)\sum_{j=1}^{n_A} y_{pj}$$

$$b_A = \sum_{j=1}^{n_A}(x_{pj}-\bar{x}_A)(y_{pj}-\bar{y}_A) / \sum_{j=1}^{n_A}(x_{pj}-\bar{x}_A)^2$$

The estimated variance of this estimator is:

$$v(\hat{Y}_{pA}^{(reg)}) = [N_A(N_A-n_A)/n_A(n_A-3)](1-R_A^2)\sum_{j=1}^{n_A}(y_{pj}-\bar{y}_A)^2$$

where R_A^2 is the regression determination coefficient for A (see Appendix B for formula).

Proration is used to obtain estimates for subsets where regression cannot be done. For a subset B of stratum p containing N_B frame units and n_b segments, a proration total acreage estimate is given by:

$$\hat{Y}_{pB}^{(pro)} = N_B\bar{y}_B$$

where $\bar{y}_B$ is the sample mean acreage in B.

A stratum level estimate of crop acreage is computed by summing the appropriate subset estimates (regression and proration) within the stratum. The various stratum level estimates are summed to obtain a region level acreage estimate.

APPENDIX D: COUNTY ESTIMATION

The Battese-Fuller model for county estimation can be applied within a stratum/analysis district combination (subset) where classification and regression have been performed. The analyst computes Battese-Fuller estimates of crop acreage for all counties and subcounties within the subset. For those counties and subcounties within the region of interest but outside of usable satellite coverage, proration can be used to obtain estimates.

Assume that a given stratum p has been divided into subsets A and B, where A has usable satellite coverage and B does not. Of course, one subset or the other could be the entire stratum. The Battese-Fuller model assumes that, for a subset where regression is performed for a crop of interest, segments grouped by county have the same slope relationship as the analysis district but require a different intercept. For the sample segments in A, the following relation is assumed to hold:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{ij}$$

$$= \beta_0 + \beta_1 x_{ij} + v_i + e_{ij},\ i=1,\ldots,c;\ j=1,\ldots,n_{Ai}$$

where:

n_{Ai} = number of sample segments in subset A portion of county i

y_{ij} = reported acres of crop in segment j of county i

x_{ij} = number of pixels classified to crop in segment j of county i

The error terms v_i and e_{ij} are assumed to be independent and normally distributed, with mean 0 and variances σ_v^2 and σ_e^2, respectively. The covariance structure of the summed error terms u_{ij} is then:

$$\begin{aligned} \mathrm{cov}(u_{ij},u_{kr}) &= 0, && \text{if } i\neq k \\ &= \sigma_v^2 && \text{if } i=k,\ j\neq r \\ &= \sigma_v^2+\sigma_e^2, && \text{if } i=k,\ j=r \end{aligned}$$

The parameter σ_v^2 is both a within county covariance and a between county component of the variance of any residual, while σ_e^2 is the within county variance component. The county mean residuals are given by:

$$\bar{u}_{i.} = \bar{y}_{i.} - b_0 - b_1 \bar{x}_{i.}$$

where:

$$\bar{y}_{i.} = (1/n_{Ai}) \sum_{j=1}^{n_{Ai}} y_{ij}$$

$$\bar{x}_{i.} = (1/n_{Ai}) \sum_{j=1}^{n_{Ai}} x_{ij}$$

b_0, b_1 = estimated regression coefficients (from small scale estimation)

For a county or subcounty, the unadjusted estimator of total crop acreage is:

$$\hat{Y}_i(BF) = N_{Ai}[b_0 + b_1\bar{X}_i + \delta_i\bar{u}_{i.}]$$

where:

$\bar{X}_i$ = mean number of pixels per frame unit classified to crop in county i

N_{Ai} = number of frame units in subset A portion of county i

$0 \leq \delta_i \leq 1$

Different values of δ_i generate different Battese-Fuller estimates. If $\delta_i = 0$, then the estimate lies on the analysis district regression line. It can be shown that the mean square error of the estimator for a given county is minimized by using:

$$\delta_i^* = n_{Ai}\sigma_v^2/(n_{Ai}\sigma_v^2+\sigma_e^2)$$

If the variance components σ_v^2 and σ_e^2 are not known, then they can be estimated to approximate δ_i^* using the above formula. The unbiased estimators given by Fuller and Battese [6] were used to obtain the estimates presented in Section 6. These require that a county or subcounty contain at least two sample segments. For counties and subcounties containing fewer than two segments, estimates were obtained by using $\delta_i = 0$.

The unadjusted estimates of county totals generally do not sum to the corresponding analysis district totals. In order to get agreement, adjustment terms are added to the estimates. The formula for the resulting adjusted Battese-Fuller estimator is:

$$\hat{Y}_i(adj) = \hat{Y}_i(BF) - (N_{Ai}/N_A)\sum_{j=1}^{c} \delta_j\bar{u}_{j.}$$

where N_A is the number of frame units in A. The adjusted estimates thus generated will sum to the appropriate analysis district totals. A method for estimating the variance of these estimators is given in [5].

As mentioned earlier, proration can be used to obtain acreage estimates for counties or subcounties where usable satellite coverage is not available. These areas often contain very few segments if any, so the mean acreage over all of subset B is used to compute the proration estimates. The formula is:

$$\hat{Y}_i(\text{pro}) = N_{Bi}\bar{y}_B$$

where:

N_{Bi} = number of frame units in subset B portion of county i

$\bar{y}_B$ = mean reported crop acreage for subset B

For a county that includes more than one stratum, estimates are generated for each stratum and summed to obtain an overall county acreage estimate.

APPENDIX E: ADDITIONAL TABLES

Table A1: Training Pixel Counts, Number of Categories, and Prior Probabilities For Classification

TM

Cover	Number of Training Pixels	Number of Categories	Prior Probability
Corn	20,665	16	.441
Soybeans	15,093	12	.321
Permanent Pasture	3,171	9	.067
Other	8,080	5	.171

SPOT

Cover	Number of Training Pixels	Number of Categories	Prior Probability
Corn	45,335	24	.431
Soybeans	33,467	23	.317
Permanent Pasture	7,317	12	.069
Other	19,308	13	.183

Table A2: Small Scale Classification Summary

TM (unequal priors) ----------Pixels Classified To:----------					
From:	Corn	Soybeans	P. Pasture	Other	Total
Corn	25,155	1,600	440	2,166	29,361
Soybeans	1,634	18,533	349	1,514	22,030
Permanent Pasture	929	398	2,774	1,369	5,470
Other	3,660	2,639	1,825	9,397	17,521
Total	31,378	23,170	5,388	14,446	74,382

SPOT (unequal priors) ----------Pixels Classified To:----------					
From:	Corn	Soybeans	P. Pasture	Other	Total
Corn	50,014	4,008	1,403	3,926	59,351
Soybeans	7,994	33,034	773	2,719	44,520
Permanent Pasture	2,404	1,224	3,213	4,231	11,072
Other	10,109	6,331	4,197	14,551	35,188
Total	70,521	44,597	9,586	25,427	150,131

TM (equal priors) ----------Pixels Classified To:----------					
From:	Corn	Soybeans	P. Pasture	Other	Total
Corn	25,906	1,637	1,135	683	29,361
Soybeans	3,768	17,135	737	390	22,030
Permanent Pasture	1,330	501	3,355	284	5,470
Other	6,046	3,353	4,255	3,867	17,521
Total	37,050	22,626	9,482	5,224	74,382

SPOT (equal priors) ----------Pixels Classified To:-------					
From:	Corn	Soybeans	P. Pasture	Other	Total
Corn	48,353	5,563	3,458	1,977	59,351
Soybeans	8,504	32,209	1,748	2,059	44,520
Perm. Pasture	2,887	1,604	5,159	1,422	11,072
Other	13,101	7,559	8,335	6,193	35,188
Total	72,845	46,935	18,700	11,651	150,131

Table A3: TM Regression Data

	Corn	Soybeans
Sample Mean Pixels	1,364.26	1,007.39
Population Mean Pixels	1,202.8	996.13
Sample Mean Acreage	252.06	185.58
Regression Mean Acreage	219.4	183.4
Slope Coefficient	0.2023	0.1933

Table A4: Number of Frame Units and Segments for Counties

	Stratum 14		**Stratum 30**		**Stratum 40**	
County	F.U.'s	Segs.	F.U.'s	Segs.	F.U.'s	Segs.
Audubon	436	3	19	0	8	0
Calhoun	562	3	22	0	0	0
Carroll	566	1	39	0	0	0
Crawford	709	6	50	0	15	0
Greene	566	4	23	0	8	0
Guthrie	586	2	34	0	15	0
Ida	432	2	20	0	0	0
Sac	573	4	44	1	7	0
Shelby	579	3	31	1	14	0
Total	5,009	28	282	2	67	0

Table A5: Breakdown of County Estimates by Strata and Estimation Method

Corn

	Stratum 14		Stratum 30	
County	Battese-Fuller	Proration	Proration	Total
Audubon	88,712	-	339.2	89,051
Calhoun	129,466	2,005	392.7	131,864
Carroll	139,757	-	696.2	140,453
Crawford	130,923	18,246	892.5	150,062
Greene	118,934	-	410.6	119,345
Guthrie	93,908	-	606.9	94,515
Ida	43,995	49,323	357.0	93,675
Sac	136,784	-	785.4	137,569
Shelby	140,353	-	553.4	140,906
Total	1,022,832	69,574	5,033.9	1,097,440

Soybeans

	Stratum 14		Stratum 30	
County	Battese-Fuller	Proration	Proration	Total
Audubon	72,192	-	61.8	72,254
Calhoun	140,639	737	71.5	141,447
Carroll	108,529	-	126.8	108,656
Crawford	91,992	6,707	162.5	98,861
Greene	118,104	-	74.8	118,179
Guthrie	86,112	-	110.5	86,223
Ida	34,777	18,130	65.0	52,972
Sac	112,645	-	143.0	112,788
Shelby	90,042	-	101.0	90,143
Total	855,032	25,574	916.9	881,523