United States Department of Agriculture

National Agricultural Statistics Service

Research and Applications Division

SRB Staff Report Number SRB-90-06

May 1990

THE CONSTRUCTION OF A DRY BEAN AREA SAMPLING FRAME IN MICHIGAN

Cheryl L. Stup
J. Donald Allen

The Construction of a Dry Bean Area Sampling Frame in Michigan
by Cheryl L. Stup and J. Donald Allen, Research and Applications Division, National Agricultural Statistics Service, Washington, D.C. 20250, April 1990. NASS Staff Report No. SRB-90-06.

ABSTRACT

This project utilized Landsat Thematic Mapper data, French Spot Multispectral Scanner data, and Landsat Multispectral Scanner data to define the land use stratification for use in dry bean surveys. Ground truth was collected from 198 sites, each approximately one square mile. The classified data from the three sensors were pieced together to help lessen the effect of cloud coverage problems in the Lake Huron area of Michigan. The new dry bean strata were incorporated into a sampling frame, or list of units, comprised of areas of land. These strata were developed from a combination of the existing area sampling frame strata and the indicated percentage of dry bean acreage contained in those strata. This paper documents the procedures followed in the construction of this dry bean area sampling frame.

KEY WORDS

Area sampling frame, classification, percent correct, pixel, primary sampling unit, segment, strata

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The authors would like to thank Dale Atkinson and the entire Area Frame Section for their technical work and observations, Martin Ozga and Bob Losa for their programming assistance, Jennifer Kotch for her computer graphics, and the Michigan State Statistical Office for their support and assistance with this project.

INTRODUCTION

The National Agricultural Statistics Service (NASS), within the United States Department of Agriculture (USDA), is primarily responsible for collecting, preparing, and publishing crop and livestock estimates. Additionally, NASS continually conducts research aimed at improving the accuracy and efficiency of these estimates. The input data for the estimates is gathered during quarterly agricultural surveys (June, September, December, and March) and during the growing season by monthly objective yield surveys that utilize both crop and field characteristics. One source of data for the quarterly agricultural surveys is a stratified area sampling frame (ASF). These frames are constructed for each individual state with the stratification being based on the percent of land cultivated. The ASF covers the entire population (i.e., land area within a state) and can be used alone to produce an acreage estimate. Remotely sensed data was originally used in the 1950's in the form of aerial photography to aid in the construction of these frames. Currently, remotely sensed data from satellites is not only used for state ASF's, but also for creating ASF's for specialized crops such as dry beans.

Approximately one third of the dry beans produced in the United States are grown in an eighteen county area of Michigan. For most crops, an ASF which is stratified based on land uses (i.e., percent cultivation) will be sufficient for producing reliable acreage indications. However, in a state where a crop is grown only in a small area, the ASF may fail to yield precise acreage indications. The ASF is built upon the concept of homogeneity within the strata across the entire area. When the strata in this area are not homogeneous, the crop variances become inflated. These high variances produce widely varying crop acreage indications.

One way to address this problem is by using substratification; that is, within a stratum, a geographical substratum can be formed. At the substratum level, the sample can then be allocated to increase precision for the crop of interest. In addition, a multiple frame approach can be used. With this approach, a list frame of known growers of the crop of interest is maintained and a stratified sample is drawn from it. The incompleteness of the list is accounted for by the portion of nonoverlap found in the ASF; that is, a multiple frame encompasses both an area frame and a list frame. In 1976 and 1977, a substratified dry bean area sampling frame was used for estimating Michigan dry bean acreage. In 1978 this approach was supplanted by a multiple frame survey. The multiple frame survey resulted in an improvement in the coefficient of

variation (CV) for the acreage indication. However, over time the resulting CV's began to increase. Additionally, there was some concern over the timing of the ASF survey, which was part of the national June Enumerative Survey (JES). It was felt that the June survey, which was conducted annually around the end of May (and is currently conducted in early June), was too early in the season to correctly assess dry bean acreage and "intentions" to plant dry beans since normal dry bean planting dates ran from about May 25 to July 1 [1,2].

As a result, the decision was made to build a special ASF. This special ASF was to be stratified on the crop of interest (dry beans) as opposed to stratifying on general land usage. Such a project was undertaken in Michigan in 1980 and covered a sixteen county area. Initially the land in the area was divided into primary sampling units (PSU's) which were homogeneous as to soil type and percent cultivated. These units were assigned a subjective measure of the likelihood of containing dry beans. This measure was based on survey data from the previous year, climatological data, soil type, and county acreage estimates. Clusters were formed based on these variables and the strata were based on the resulting clusters. In the first year (1981) there were twenty strata used. A sample was allocated using previous ASF survey data to provide an estimate of the standard error to expect. The survey itself was conducted in mid July. The resulting CV was 8.2 as compared to a CV 13.0 for the area frame indication (see table 1) [3]. No comparison with a multiple frame indication was made since the list frame was not utilized. Over time, the special dry bean frame for Michigan has become outdated, so at the request of the Michigan State Statistical Office, the Remote Sensing Section and the Area Frame Sampling Section of NASS undertook the development of a new dry bean ASF.

TABLE 1: Acreage Indications (thousands) and Coefficient of Variations (percentages) from the Area Frame, Multiple Frame, and Dry Bean Area Frame from 1978 - 1989.

Year	Official Estimate	Area Frame		Mult Frame		DB Area Frame	
		Acreage	CV	Acreage	CV	Acreage	CV
1978	530	623	12.2	530	5.4		
1979	470	510	13.4	496	6.2		
1980	590	496	14.7	518	7.4		
1981	650	536	13.0			612	8.2
1982	560	383	14.8			522	7.9
1983	360	275	17.2			318	8.2
1984	400	367	16.2			360	7.2
1985	440	488	20.1			396	7.6
1986	480	421	15.5	419	11.7	417	7.2
1987	470	536	17.1	466	10.2	407	7.9
1988	260	307	21.8	321	14.8	204	10.6
1989	340	294	27.0	253	10.2	292	9.0

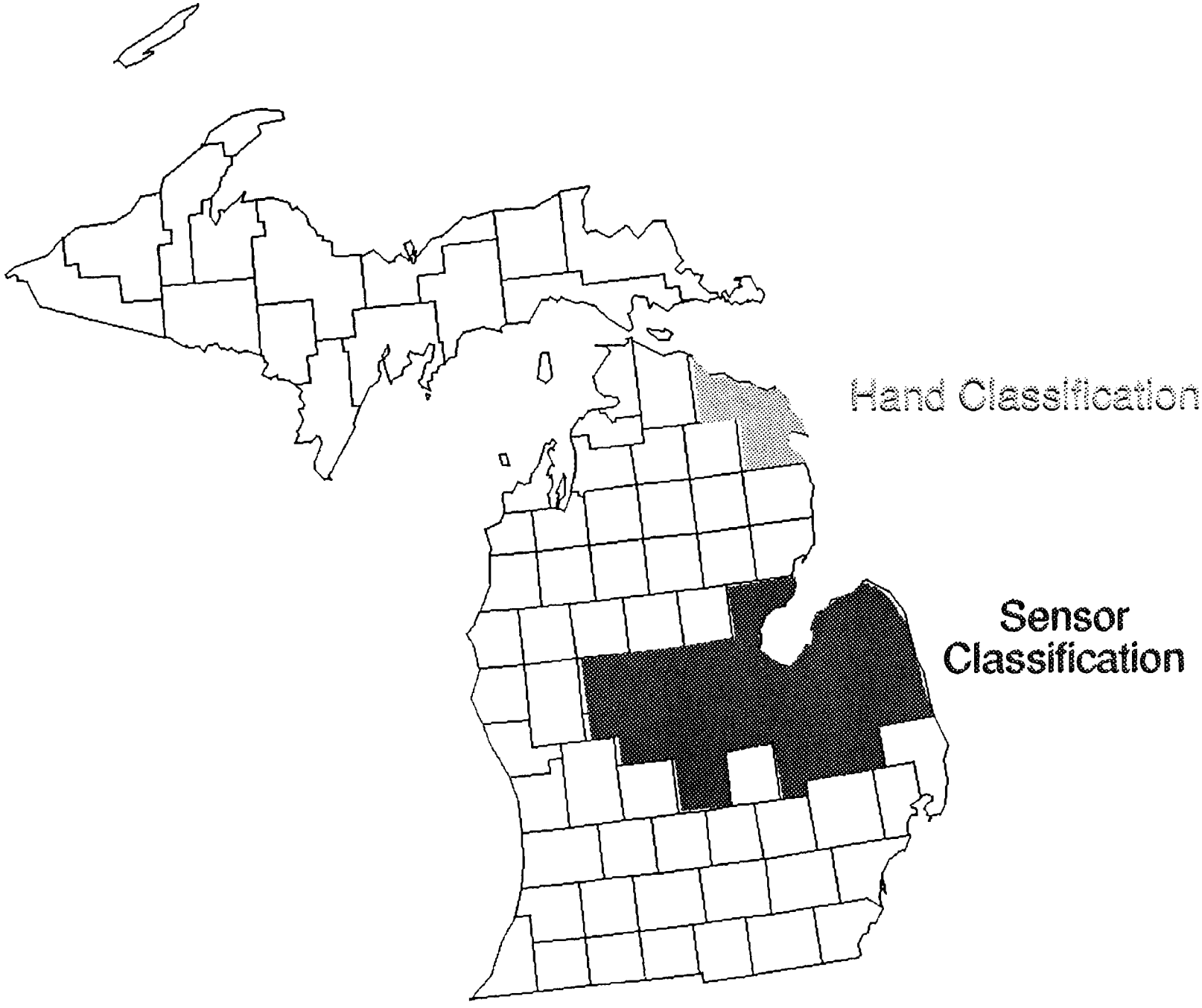
Multiple frame indications were obtained for all major crops beginning in 1986 as part of the national estimating program.

In developing the new dry bean ASF, data from the Landsat Thematic Mapper (TM), French Spot Multispectral Scanner (SPOT), and Landsat Multispectral Scanner (MSS) were used. Three sensors were used due to cloud coverage problems over part of the study area. The satellite data were at the "pixel" level, which is the ground area unit for recording remotely sensed data. Ground truth data were gathered for replicated, stratified samples of land areas called segments. For each sensor, pixel data were combined with ground truth data in a supervised clustering procedure to develop categories with statistical signatures (means and variances) for each crop. This information was then used in a maximum likelihood classification at the segment level, where the probability of each pixel being from each category was calculated. The crop category with the highest probability was then selected as the output category [4]. Upon achieving a satisfactory segment level classification, county wide classifications were made. These classifications identified dry beans and all other crops in each PSU in the study area. The results were unique to each sensor (TM, SPOT, and MSS) and, when pieced together, covered most of the sixteen county thumb area. The two remaining counties were manually classified at the PSU level using subjective measures and past survey data to determine the percent of dry beans in each PSU. Utilizing these classifications, the percent of dry beans was calculated for each PSU in the entire eighteen county study area. The new Michigan dry bean strata were then formed from a combination of the previous ASF land use strata and the percent of dry beans by PSU, as opposed to the previous method of forming strata solely on the percent cultivated by PSU.

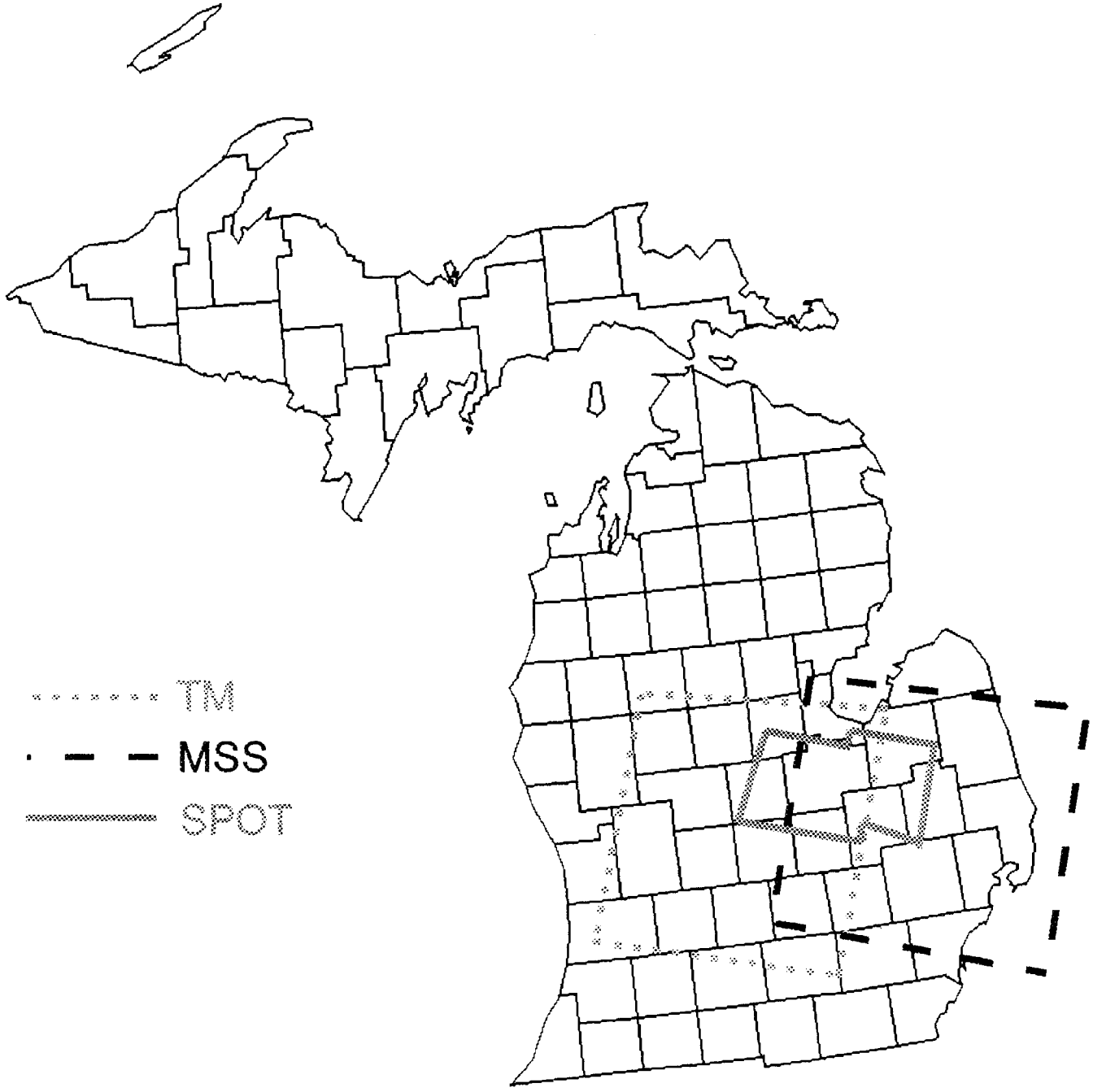
DATES FOR THE GROUND TRUTH AND SATELLITE DATA

The area of interest for sensor classification was the "thumb" area of Michigan, a sixteen county area bordered by Lake Huron. The crop under study was dry beans. Due to cloud coverage problems, data from the TM, SPOT, and MSS sensors were all pieced together. The overpass dates for the three sensors all occurred in the latter half of July 1987. The TM scene had an overpass date of July 29 with the MSS scene having an overpass date of July 22. There were two SPOT scenes, each having an overpass date of July 20. The ground information for this study was collected as part of the Michigan Dry Bean Acreage Estimation Survey conducted during July 1987. The ground truth consisted of 198 segments in the sixteen county region. (See map 1 and map 2). The data was processed using both the Computer-Aided Stratification (CAS) software system and the NASS PEDITOR software [5,6,7].

MICHIGAN CLASSIFIED COUNTIES



MICHIGAN SENSOR COVERAGE



PROCEDURES

A supervised clustering procedure was done independently for each of the three sensors [8]. Initially, the pixels were separated into k crop cover types based on the ground truth data. Then, n_i (i=1,...,k) categories were formed for each of the k crop cover types. In addition to the k crop covers, categories were also derived for water and clouds by using known water and cloud cover areas to develop statistical signatures. In the end, each cover (including water and clouds) had a set of signatures consisting of a mean vector and a covariance matrix for each of its categories. The length of the vector corresponds to the number of channels of data provided by the sensor (TM has 7 channels, SPOT 3 channels, and MSS 4 channels).

A statistics file was created by combining the n_i categories for all crop covers. A segment level classification was done and all pixels in the segment were assigned a specific category. Categories were added and deleted from the statistics file until a satisfactory classification was achieved. Categories within a specific crop cover were considered for deletion from the statistics file depending on the following criteria:

1) A small number of pixels (less than 75) classified to a given category.
2) Very high or very low channel means as compared to the corresponding elements across the set of mean vector values for a given crop.
3) High channel variances as compared to the corresponding channel variances across the set of covariance matrices for a given crop.
4) High covariances between channels in the covariance matrix.
5) A low degree of separability using the transformed divergence rule [8].
6) High generalized variances [9].
7) A large contribution to the dry bean commission error.

Any of these criteria, alone or combined, were valid reasons for the deletion of a category from the statistics file (see table 2). Once the statistics file was finalized, both no prior (np) and prior (p) probabilities were applied to the statistics file. An np probability assumes that any two crop categories have an equal probability of occurring; therefore, each crop category is assigned an equal weight in the statistics file. A prior probability assigns a weight to a crop based on the percentage of that crop actually reported in the ground truth. The weight is then proportioned among the various categories within a particular crop. This subweight is determined by the number of

pixels in each category of the crop. The result is the probability of a pixel being classified to a specific crop category.

TABLE 2: Original and Edited number of categories per crop cover

CROP	TM		SPOT		MSS		
	ORIGINAL	EDITED	ORIGINAL	EDITED	ORIGINAL	EDITED NP	EDITED P[a]
ALFALFA	13	9	19	8	7	5	7
BEETS	13	10	20	11	6	2	4
CLOUDS	1	1	1	1	2	2	2
COLORBNS[b]	12	9	17		3		
CORN	11	10	25	12	6	3	4
DRYBEANS[c]	12		18	9	9	5	7
FARM	11	6	10	3	4	4	4
NAVYBEANS[b]	17	12	18		7		
OTHER	18	18	18	7	17	8	7
OTHERHAY	3	3	8	2	3	2	3
PERMPAST[d]	13	10	10	8			
SORGHUM[e]	3	3					
SOYBEANS	13	13	16	5	7	7	7
WASTE	13	9	18	13	12	8	7
WATER	1	1	5	5	2	2	2
WOODS	16	13	12	9	5	5	5

a) The MSS sensor used different files for the no prior and prior statistics files, while TM and SPOT used the same statistics files.
b) The MSS and SPOT sensors obtained maximum results by combining color beans and navy beans and categorizing them as dry beans.
c) The TM sensor obtained maximum results by categorizing color beans and navy beans separately.
d) MSS grouped permanent pasture in the other category, while TM and SPOT categorized it separately.
e) MSS and SPOT grouped sorghum in the other category, while TM categorized it separately.

To determine the "best" classification for each sensor, several statistics were considered: the correlation coefficient (R) between the classified data and the reported data at the segment level, the apparent error rate (APER), and the percentage of correctly classified dry beans in conjunction with the commission error (see Appendix A for statistical definitions). The correlation coefficient was considered due to its importance as an indicator of how well the classification could be used in predicting actual dry bean acreage. In cases where the classification is incorrect in a consistent manner, the correlation coefficient can still be high, therefore lessening its importance in defining the new strata. The APER was a measure for determining the percentage of misclassification between the two categories, dry beans and not dry beans. While the APER provided valid information, by definition it reflects the misclassification of both dry beans and not dry beans, whereas the intent was solely for accuracy in the dry bean classification. In the end, the percentage of correctly classified dry beans along with the commission error were used as the most influential statistics in determining the best classification for each sensor.

When considering the percent correct of dry beans, one must understand that this is the percent of reported dry bean pixels actually classified to dry beans as opposed to being classified to one of the many other crops in the study area. Michigan has a very diverse agricultural community with the number of crops in the various dry bean counties being between 15 and 25. As a result, the "chance" classification to any single category ranged from 4 to 7 percent, with anything above that range due to the effectiveness of the process.

The correlation coefficient, APER, percent correct of dry beans, and commission error were calculated for both the np and prior statistics files for each sensor. Each statistics file was independently edited to the point where these four statistics were deemed to be maximized. In most cases, improvement in one statistic can only be achieved at the expense of another statistic, so there are "trade offs" in performing the editing (see table 3).

TABLE 3: The correlation coefficient, apparent error rate, percent correct dry beans, and commission error achieved by the final statistics file for each sensor

Sensor	# segs	R	APER	Dry bean % Cor	Dry bean % Com
TM					
NP	89	.92	5.1	74.21	43.56
PRIOR	89	.93	3.9	71.47	33.23
SPOT					
NP	43	.92	12.2	56.52	58.35
PRIOR	43	.89	11.6	54.73	56.62
MSS					
NP	102	.58	19.2	48.97	74.60
PRIOR	102	.56	18.9	51.63	73.55

Upon achieving satisfactory np and prior statistics files, the proportion of pixels classified to dry beans, other crops, noncrops, clouds, and water were determined for the study area. As alluded to earlier, an ASF is based on the percent of land cultivated, and the strata formed from the ASF served as the starting point for establishing the new dry bean ASF. The process called for bringing county boundaries and PSU boundaries together so that each county was essentially divided into PSU's. Next, the classification was performed at the PSU level for each county. This was done separately for each set of satellite data. Overlap across PSU's for the different satellites was not addressed at the classification stage; that is, if a single PSU was covered by multiple sensors, there were multiple indications for the percent of dry beans in that PSU. On the other hand, overlap across the two SPOT scenes was removed. This was done by assigning only unique portions of the PSU to each particular scene. At this point, each PSU even partially contained in a scene had associated with it percentages indicating the proportion of pixels classified to dry beans, other crops, noncrops, clouds, and water.

In determining which scene to give priority to, a combination of an overall acceptance criterion and the sensor quality was used. An overall acceptance criterion was developed uniquely for each PSU within a sensor. This criterion was based on the percentage of "real coverage", and was defined as follows:

% real coverage =

$$\frac{(\text{sensor coverage } PSU_j - \text{cloud coverage } PSU_j)}{\text{total area in } PSU_j} * 100$$

where j is the jth PSU and j=1,...,1639.

At least a 50% real coverage value was needed to ensure that the sampled PSU's were representative of the entire population of PSU's. At the 50% level, 83 PSU's (out of a total of 1639 PSU's) had either inadequate sensor coverage or excessive cloud coverage. These 83 PSU's were individually reviewed and manually imputed based on the percentages in neighboring PSU's. For those PSU's having at least 50% real coverage and contained in more than one sensor, dry bean percentages from the candidate sensor were selected based on the relative quality of the parent sensor. The following sensor priorities (highest to lowest) were used:

1.) TM
2.) SPOT
3.) MSS

The above priorities were superseded when a lower priority sensor had a real coverage percentage at least 25 percent higher than that of a higher priority sensor.

The percent of dry beans by PSU in the sixteen county area utilizing satellite data were then merged with the dry bean percentages by PSU in the two manually classified counties. The new dry bean ASF was then created based on the percent of dry beans by PSU from the combined classification results of these eighteen counties.

RESULTS

The lower than anticipated results of the statistical measures (especially MSS) can be accounted for by several problems that were inherent in the data:

1) Diverse agricultural community

There were 23 cover types classified with MSS, 24 with TM, and 21 with SPOT. This great diversity detracts from the separability between differing crop covers. Some of the crops causing the greatest confusion with dry beans were soybeans, idle crop, winter wheat, corn, and waste fields.

2) Combination of pixel size and small field size

The approximate pixel sizes for the three sensors are as follows: TM, .2 acres; SPOT, .1 acres; and MSS, .8 acres. The average dry bean field size in the study area was approximately 24 acres. For the "average" field, the following amount of data was available for each of the three sensors: TM, 120 pixels, 7 bands; SPOT, 240 pixels, 3 bands; and MSS, 30 pixels, 4 bands. Therefore, there were 840 pieces of information available to classify the "average" TM dry bean field, 720 for SPOT, and 120 for MSS. The impact of the smaller amount of MSS information available was apparent in its statistical measures.

In the construction of this new dry bean ASF, the idea was to leave the primary sampling unit boundaries utilized from the ASF intact while at the same time creating new strata boundaries. These new strata boundaries were determined by a combination of the ASF land use strata and the percent of dry beans within each PSU. The ASF land use strata utilized in the construction of the dry bean ASF were defined as follows:

Land Use Stratum	Definition
11	General Cropland, 75% or more cultivated
12	General Cropland, 50-74% cultivated
20	General Cropland, 15-49% cultivated
40	Range and Pasture, less than 15% cultivated

The new dry bean ASF consisted of all PSU's in the eighteen county study area in Michigan's ASF in strata 11, 12, 20, and 40. No other strata were included in the dry bean ASF due to their low probability of dry bean containment. The PSU's in stratum 40 were separated from those in strata 11, 12, and 20 because the target segment size in stratum 40 was two square miles as compared to one square mile in strata 11, 12, and 20. The final strata were defined as follows:

Dry Bean Stratum	Land Use Stratum	Definition
1	11, 12, 20	0.0 - 5.4 percent dry beans
2	11, 12, 20	5.5 - 10.4 percent dry beans
3	11, 12, 20	10.5 - 20.4 percent dry beans
4	11, 12, 20	20.5 - 30.4 percent dry beans
5	11, 12, 20	30.5 + percent dry beans
6	40	0.0 - 0.4 percent dry beans
7	40	0.5 + percent dry beans

CONCLUSION

Remote sensing was efficiently used in reducing the time involved in creating a new dry bean area sampling frame. The results from the new frame can be obtained and evaluated when the Michigan Dry Bean Estimation Survey is completed. In the future, it is hoped that CAS can be used more extensively. Although the viewing power of CAS has not been fully explored, it would be an excellent tool for viewing classified segments and for locating these segments on a county map. Also, in accordance with several other studies done in Kansas, Iowa, and Arkansas, the Landsat Thematic Mapper proved to be the "sensor of choice" for best performance [10,11,12].

REFERENCES

[1] Fecso, Ron, J. Geuder, B. Hale, and S. Pavlasek. "Estimating Dry Bean Acreage in Michigan," U.S. Dept. of Agriculture, Statistical Reporting Service, 1982.

[2] Fecso, Ron and J. Geuder. "Estimating Dry Bean Acreage in Michigan: Second Year Results," U.S. Dept. of Agriculture, Statistical Reporting Service, 1983.

[3] Cochran, W.G. Sampling Techniques. New York, New York: John Wiley and Sons, Inc., 1977.

[4] Hanuschak, G.A. "Precision of Crop-Area Estimates," in Proceedings of the Thirteenth International Symposium of Remote Sensing of Environment, Ann Arbor, Michigan, April 1979.

[5] Cheng, T.D., G.L. Angelici, R.E. Slye, and M. Ma. "Computer-Aided Boundary Delineation of Agricultural Lands," NASA Ames Research Center, NASA Technical Memorandum 102243, 1989.

[6] Ozga, M., "USDA/SRS Software for Landsat MSS-based Crop Acreage Estimation," Proceedings of the IGARSS '85 Symposium, Amherst, Massachusetts, 1985.

[7] Angelici, R., R. Slye, M. Ozga, and P. Ritter. "PEDITOR - A Portable Image Processing System," in Proceedings of the IGARSS '86 Symposium, Zurich, Switzerland, 1986.

[8] Swain, Philip H. and S.M. Davis. Remote Sensing: The Quantitative Approach. New York. McGraw-Hill Book Company, 1978.

[9] Johnson, R.A. and D.W. Wichern. Applied Multivariate Statistical Analysis, Englewood Cliffs, New Jersey: Prentice Hall, 1988.

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

[11] Bellow, M.E. (in press). "Comparison of Sensors for Corn and Soybean Planted Area Estimation," in Proceedings of the IGARSS '90 Symposium, Greenbelt, Maryland, 1990.

[12] Allen, J.D. (in press). "Remote Sensor Comparison for Crop Area Estimation using Multitemporal Data," in Proceedings of the IGARSS '90 Symposium, Greenbelt, Maryland, 1990.

[13] Cotter, J. and J. Nealon. "Area Frame Design for Agricultural Surveys," U.S. Department of Agriculture, National Agricultural Statistics Service, 1987.

[14] Hanuschak, G.A. and K.M. Morrissey. "Pilot Study of the Potential Contributions of Landsat Data in the Construction of Area Frames," U.S. Department of Agriculture, Statistical Reporting Service, 1977.

[15] Houseman, E.E. "Area Frame Sampling in Agriculture," U.S. Department of Agriculture, Statistical Reporting Service, 1975.

Appendix A: Definitions of Apparent Error Rate, Percent Correct of Dry Beans, Commission Error, and Correlation Coefficient

		CLASSIFIED PIXELS	
		DRY BEAN	NOT DRY BEAN
REPORTED PIXELS	DRY BEAN	a	b
	NOT DRY BEAN	c	d

Apparent Error Rate (APER)

--The percentage of reported pixels that are incorrectly classified with respect to dry beans.
--APER = $\frac{b+c}{a+b+c+d}$.

Percent Correct Dry Beans

--The percentage of reported dry bean pixels that are actually classified as dry bean pixels.
--Percent Correct = $\frac{a}{a+b}$.

Commission Error

--The percentage of non dry bean pixels that are incorrectly classified as dry bean pixels.
--Commission Error = $\frac{c}{a+c}$.

Correlation Coefficient (R)

--The correlation coefficient represents the degree of association between the classified dry bean pixels and the reported dry bean pixels.