12/12/88 88-04

DRAFT

THE COMPARISON OF EMULATED MULTISPECTRAL SCANNER DATA SETS by James Mark Harris, Research and Applications Division, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C. 20250, xxxxxxxx 1988. RAD Staff Report No. SRB-88-xx.

ABSTRACT

The study goal was the selection of a satellite data set to replace the multispectral scanner satellite data from Landsat 4 and 5. Multispectral scanner satellite data was used in the National Agricultural Statistics Service's Domestic Crop and Land Cover Project. The satellite data was processed to produce an independent variable which was used in a regression estimator of planted crop acres for corn and soybeans. Four data types were evaluated as a possible replacement for the multispectral scanner data. The four data types evaluated were generated from thematic mapper satellite data. The selection criteria was based on the precision of the regression estimator. Results showed significant differences in the evaluated data sets in some strata within a given crop, corn or soybean. Differences in data sets are attributed to differences in thematic mapper satellite band combinations, not in data reduction techniques.

KEYWORDS

Multispectral Scanner, Thematic Mapper, Scene, Pixel, R-Square, Bands, Auxiliary Variate

ACKNOWLEDGMENTS

The author thanks the following persons for their support:

E. M. Jones, Jr., Martin L. Holko for their general guidance and suggestions on statistical analysis.

Robert C. Hale, Sherman B. Winings, and Mickey Yost for their technical support on remote sensing analysis.

Martin Ozga and William Daugherty for their computer programming and hardware support.

Brian Carney for his supervision during this project.

CONTENTS

SUMMARY

The purpose of this study was to select a satellite data set to replace the multispectral scanner satellite data, which was used in the National Agricultural Statistics Service's Domestic Crop and Land Cover project. The multispectral scanner satellite data will not be available from the next LANDSAT satellite. The next LANDSAT satellite to be launched, LANDSAT 6, will carry the thematic mapper sensor. Past research conducted by the National Agricultural Statistics Service indicated that thematic mapper satellite data produced a more precise regression estimator. However, the cost / benefit ratio favors the multispectral scanner satellite data for large area applications. Both initial data costs and processing costs are greater for the thematic mapper data than for the multispectral scanner data.

The thematic mapper satellite scanner records seven readings, or bands, for each 30 square meter ground area. The ground area resolution is also called pixel size. Pixel is a term derived from "picture element", which has been generalized to mean the basic unit for recording satellite acquired remotely sensed data. Pixel sizes vary among types of satellite scanners. Multispectral scanner satellite data has a 60 square meter pixel with four bands. So, four thematic mapper pixels, each with seven bands, cover the same area as one multispectral scanner pixel with four bands.

The four data sets evaluated are described as emulated multispectral scanner data. The adjective "emulated" was used to describe the data sets, because the number of pixels and the number of bands of the thematic mapper satellite data were reduced to imitate the number of bands and pixels of multispectral scanner satellite data. Two different data reduction techniques, sampling and averaging, were combined with two different thematic mapper band combinations to produce the four different emulated data sets.

The emulated data sets were processed through the USDA's PEDITOR software. The output from the PEDITOR software is the classified number of pixels for each segment and crop. The classified number of pixels for each segment and crop is used as an independent variable or "auxiliary variate" in a linear regression estimator of the planted crop acres. Selection of the emulated data set was based on the precision of the linear regression estimates produced by the data sets.

Results showed significant differences in the emulated data sets in some strata, depending on the crop. Differences in data sets are attributed to differences in band combination, not in data reduction techniques. The thematic mapper band combination of 2, 3, 5, 4 produced the highest sample correlation coefficients for both data reduction techniques. The averaging data reduction technique produced slightly higher sample correlation coefficients for both corn and soybeans when all strata were combined for the regression. However, there were no significant differences between sampling and averaging data reduction techniques. The recommendation is to request the averaged bands 2, 3, 5, 4 emulation. Due to the changing cost of processing, it is also recommended to conduct a

cost / benefit study comparing the emulated and thematic mapper data. Given that averaging was not shown to be significantly better than sampling and that the data processing costs of averaging are slightly higher than for sampling, further research into the two data reduction techniques is warranted.

INTRODUCTION

The National Agricultural Statistics Service (NASS) used multispectral scanner (MSS) satellite sensor data for the Domestic Crop and Land Cover (DCLC) project. The MSS satellite data will no longer be available when the current generation of LANDSAT satellite, LANDSAT 4 and 5, fail to operate. The next satellite scheduled to be launched in 1991 by the Earth Observation Satellite (EOSAT) Company, LANDSAT 6, will carry the Thematic Mapper (TM) sensor. Therefore, MSS satellite data will no longer be available. TM satellite data will be available from LANDSAT 6 however, there are some differences in the TM and MSS satellite data sets. Basic differences in the two types of remotely sensed satellite data is in the ground resolution and number of bands or channels recorded. The ground area resolution or "pixel size" is 60 square meters for MSS and 30 square meters for TM. A pixel is the basic unit for recording satellite acquired remotely sensed data. Pixel sizes vary for different satellite sensors. The second difference of MSS and TM satellite data is the number of bands recorded for each pixel. MSS has four bands per pixel while TM has seven bands recorded for each pixel. Appendix A gives a comparison of the band wave length for MSS and TM. The differences in MSS and TM satellite data pixel size and number of channel make TM satellite data have a seven-fold data volume increase over the MSS satellite data. It takes four TM pixels to cover ~~represent~~ 60 square meters, the same ground area as a MSS pixel. The seven-fold data increase can be calculated by multiplying four TM pixels (60 square meters) x seven bands which equals 28 readings for a 60 square meter ground area. MSS has four readings per sixty square meter ground area. Thus, for the same ground area TM satellite data has 28 readings

versus MSS satellite data's 4 readings. TM's twenty eight divided by MSS's four gives the seven-fold data volume increase for TM satellite data.

Past research performed by NASS indicated that TM satellite produced a better regression estimator. However, when cost are taken into account in a cost / benefit ratio, MSS satellite data was preferable. Costs were higher for the raw TM satellite data $3300 per scene versus MSS $660 per scene. Processing costs were also higher for TM data given the seven-fold data volume increase.

EOSAT agreed to provide NASS with a satellite data set generated from TM satellite data with the same number of bands and pixels as MSS. Thus, TM satellite data will be processed by EOSAT to emulate MSS data. EOSAT provided NASS with four different data sets that were generated from TM satellite data to emulate MSS data, thus the name Emulated Multispectral Scanner (EMSS) data. To emulate MSS data from TM satellite data, the number of pixels have to be reduced to one in four and the number of bands are reduced from seven to four. Two different pixel reduction techniques were used, averaging and sampling. Averaging takes the average of four TM pixels to produce one EMSS pixel. Sampling takes every fourth pixel to represent one EMSS pixel. Two band selections were used, TM bands 2,3,5,4 and 1,7,6,4. EOSAT provided NASS with the four different emulated data sets. The data sets cover the Columbia, Missouri area, LANDSAT Path 25 Row 33. Coverage data was September 5, 1985. The four types of EMSS data are:

1) Averaged Method Data Bands 2, 3, 5, 4

2) Sampled Method Data Bands 2, 3, 5, 4

3) Averaged Method Data Bands 1, 7, 6, 4

4) Sampled Method Data Bands 1, 7, 6, 4

The goal of this study was to select the EMSS satellite data set which would provide NASS with the best regression estimator. The study was conducted by processing the four data sets independently through NASS's PEDITOR software. The PEDITOR software is a group of programs that process the raw satellite data into a classified number of pixels for each crop in each June Enumerative Survey segment. The regression estimator uses the number of pixels classified to a crop as the independent variable and the acre reported as the dependent variable.

STUDY AREA and DATA SET

The Columbia, Missouri study area and September 5, 1985 date corresponds to the study area and date used in the 1985 Classifier Study [3]. The ground data set for the study consisted of three independent replications of the June Enumerative Survey (JES). There were four different agricultural strata in the study area. Nonagricultural areas were not considered either for this study or the Classifier study. Strata definitions are as follows:

STRATA	DEFINITION
10	50% or more cultivated
20	50% or more cultivated
30	50% or more cultivated
35	15% - 49% cultivated

The target segment size for the strata 10, 20, and 30 is 0.5 square miles and the target segment size for strata 35 is 1.0 square miles [4]. Strata 10, 20, and 30 are unique due to their geographic location. Each of the strata 10, 20, and 30 are made up of geographically contiguous primary sampling units. For a complete description of NASS's area frame procedures see Cotter and Nealon [5]. The following table gives the number of segments in the each stratum and replication for the study area.

	STRATA				
REPLICATION	10	20	30	35	TOTAL
A	8	17	37	6	68
B	12	18	29	7	66
C	12	16	32	2	62

NASS's PEDITOR software was used for the digital processing of the four data sets to produce the auxiliary variate, classified number of crop pixels [6,7]. Parallel processing of the four data sets was undertaken in order to minimize the analyst effects on signature development and classification. In other words, the analyst ran each data set through a PEDITOR program at about the same time and used sumilar judgements about the processing of each data set. Also, an attempt was made to be consistent with the processing of the MSS data set in the Classifier Study. Replication A was used in signature development and replication B and C were classified. Therefore classification was independent of signature development. Stratum 35 was used in signature development and was classified, but due to the small number of segments, the difference in stratum definition, and target segment size, the stratum was excluded from the statistical analysis.

ANALYSIS

Selecting an emulation to replace the MSS data requires a selection criterion. In defining the criterion for selection, it should noted that all regression estimates for the emulated data sets have the same statistical properties. The approach taken was to select the emulation with the maximum correlation. Maximizing correlation, of course, maximizes the R^2, which minimizes the variance of the regression estimator. Another way to view the selection is: If you were given four estimates of a crop and informed that all four estimates had the same statistical properties, you would choose the one with the minimum variance. Appendix B describes the regression estimator which uses the classified number of pixels as an auxiliary variate. It should be noted that for each segment x_i, the auxiliary variate, changes between the four emulated data sets, but the y_i, the crop acreage remains the same. Thus when calculating the variance for each of the regression estimators only the R^2 changes.

A test for choosing an auxiliary variate with the maximum correlation was worked out by Harold Hotelling [8]. The limitation of the test is it is conditional on the observed x's, the auxiliary variates, in the sample. This sample is large in comparison with other remote sensing studies, and the limitation is not seen as a problem. Appendix C gives a summary of Hotelling's test.

Tables I and II present the correlation coefficients and test values by crop and stratum. The PROB column gives the probability of observing a greater F-value.

TABLE I

CORRELATION COEFFICIENTS

CROP: CORN

AVERAGED	SAMPLED	AVERAGED	SAMPLED	F-VALUE	PROB
2,3,5,4	2,3,5,4	1,6,7,4	1,6,7,4		
0.9319	0.9402	0.7900	0.7887	7.65	0.00
0.8171	0.7749	0.6913	0.6619	1.94	0.15
0.8162	0.8275	0.5194	0.2995	1.46	0.23

CORRELATION COEFFICIENTS

CROP: SOYBEANS

STRATA	AVERAGED 2,3,5,4	SAMPLED 2,3,5,4	AVERAGED 1,6,7,4	SAMPLED 1,6,7,4	F-VALUE	PROB
10	0.7784	0.7494	0.7123	0.6256	1.27	0.30
20	0.9133	0.9110	0.8943	0.8783	5.20	0.01
30	0.7820	0.7713	0.7281	0.7064	1.41	0.25

Only twice was there a significant difference between the sample correlation coefficients at the five percent level, once in stratum 10 in the corn crop and once in stratum 20 in the soybean crop. In strata 10 for corn the sampled 2, 3, 5, 4 emulation the sample correlation coefficient was only slightly above the averaged 2, 3, 5, 4 emulation coefficient, while the converse was true in stratum 20 for soybeans. Sample correlation coefficients were higher for emulations with bands 2, 3, 5, 4 than for emulations with bands 1, 6, 7, 4 for all strata and crops.

As noted earlier Strata 10, 20, and 30 have the same land use strata definition. The strata are differentiated only by geographical location. Although these strata are independent, it seemed appropriate to combine them to increase the power of the test.

BLE III

CORRELATION COEFFECIENTS

CROP: CORN

	AVERAGED	SAMPLED	AVERAGED	SAMPLED	F-VALUE	PROB
	2,3,5,4	2,3,5,4	1,6,7,4	1,6,7,4		
MBINED						
ATA	0.8396	0.8346	0.6254	0.5254	1.93	0.13

CORRELATION COEFFICIENTS

CROP: SOYBEANS

	AVERAGED 2,3,5,4	SAMPLED 2,3,5,4	AVERAGED 1,6,7,4	SAMPLED 1,6,7,4	F-VALUE	PROB
MBINED RATA	0.8485	0.8385	0.7992	0.7635	2.55	0.06

There was no significant difference between correlation coefficients at the five percent level for the combined strata. For soybeans, however, the probablity of the observed F was only six percent. The tendency of bands 2, 3, 5, 4 to have a higher sample correlation coefficient than bands 1, 6, 7, 4 continued with the strata combined. In both corn and soybeans the averaged 2, 3, 5, 4 emulation had a slightly higher correlation coefficient than sampled 2, 3, 5, 4 emulation.

A comparison of sample correlation coefficients between band combinations for the same data reduction techniques by strata are presented in TABLE V and VI.

TABLE V

TEST FOR DIFFERENCE BETWEEN BAND COMBINATIONS

AVERAGED 2,3,5,4 -- AVERAGED 1,6,7,4

CROP	STRATA	T-VALUE	P-VALUE
CORN	10	0.079	not significant
CORN	20	0.260	not significant
CORN	30	2.327	0.03
SOYBEANS	10	0.065	not significant
SOYBEANS	20	0.008	not significant
SOYBEANS	30	0.157	not significant

BLE VI

TEST FOR DIFFERENCE BETWEEN BAND COMBINATIONS

SAMPLED 2,3,5,4 -- SAMPLED 1,6,7,4

CROP	STRATA	T-VALUE	P-VALUE
CORN	10	0.065	not significant
CORN	20	0.486	not significant
CORN	30	4.901	0.00
SOYBEANS	10	0.232	not significant
SOYBEANS	20	0.020	not significant
SOYBEANS	30	0.028	not significant

Corn stratum 30 for both data reduction techniques showed significance between band combination. Tests for differences between data reduction techniques show no significant differences. As can be seen from table I, II, III, and IV there are only slight numerical differences between the averaging and sampling coefficients for the same band combinations.

As noted earlier, Hotelling's test is valid for the sample observations. Sample correlation coefficients can be influenced by leverage points and outliers. Figures 1 through 6 in Appendix D presents the data by crop, strata and emulation. Figures 7 through 12 combine the strata and presents the data by crop and emulation. In general graphs for the same band combinations are very similar. The graphs give a visual check of the data and show that any outlier points are not overly influencing the sample correlation coefficient and the test results.

CONCLUSION

Tables I and II showed the observed sample correlation coefficients for bands 2, 3, 5, 4 were greater than the observed sample correlation coefficients for bands 1, 6, 7, 4 in all strata and all crops. Tables V and VI showed two cases where the sample correlation coefficient for bands 2, 3, 4, were significantly higher than for bands 1, 6, 7, 4. For these reasons the recommended band combination is 2, 3, 5, 4.

The selection of a data reduction technique is more difficult given the mixed results shown in tables I and II. However, in tables III and IV, where strata were combined, the averaged data reduction technique had a slightly higher sample correlation coefficient. Therefore, a recommendation of using the averaged data reduction technique is given.

The bands and data reduction technique recommended is the averaged bands 2, 3, 5, 4 emulation. The recommendation is based on the interpretation of the of statistical analysis of the author. Because Hotelling's test is conditional on the observed variates, and there was limited statistical significance, others might draw different conclusions when combined with other factors or information. One factor which could affect the recommendation is the cost associated with producing the emulations. If the cost of the averaged data reduction set is greater than the cost of the sampled data set, a cost / benefit analysis would be appropriate.

Other work on the selection of channels is recommended. One possible area of investigation is principal components analysis. Principal components analysis on several scenes may reveal a consistent ranking of bands different than the two sets reviewed for this study. Other data reduction criteria have been recommended by different authors. Stakenborg, EEC Joint Research Center, recommends selecting the median pixel value of the 2 by 2 TM pixel window. This criteria is especially important for contextual classification.

REFERENCES

Zuttermeister, John Paul. "Evaluating TM Data for SRS Acreage and Production Estimates." SRS Staff Report No. RSB-85- 02. U.S. Dept. Agr., Stat. Rep. Serv., July 1985.

Hanuschak, George, and others. "Obtaining Timely Crop Area Estimates Using Ground-Gathered and LANDSAT Data." Technical Bulletin No. 1069. U.S. Dept. Agr., Stat. Rep. Serv., Aug. 1979.

Jones, E. M., Jr. "Classifier Study." Unpublished Paper, U.S. Dept. Agr., Nat. Agr. Stat. Ser., 1987.

National Agricultural Statistics Service. "Area Information Design." June 1987.

Cotter, Jim, and Jack Nealon. Area Frame Designs for Agricultural Surveys. U.S. Dept. Agr., Nat. Agr. Stat. Ser., Aug. 1987

Ozga, Martin. "USDA / SRS Software for LANDSAT MSS-Based Crop- Acreage Estimation." IGARSS, Amherst, Mass. 7-9 Oct. 1985, 762-779.

Ozga, Martin, and others, "PEDITOR - A Portable Image Processing System." Proceedings of IGRASS '86 Symposium, Zurich, 8-11 Sept. 1986, ESA SP- 254.

Hotelling, Harold. "The Selection of Variates for Use in Prediction with Some Comments on the General Problem of Nuisance Parameters." Annals of Math. Stat., Vol. 11 (1940), pp 271-283.

APPENDIX A

The size of the picture element, or pixel, describes the resolution of the sensor. For TM the pixel size is 30 square meters while MSS has a pixel size is 60 square meters. Each TM pixel has a vector of seven reflectance values associated with the pixel, while each MSS pixel has a vector of four reflectance values associated with the pixel. It takes four TM pixels to cover the same ground area as a MSS pixel. Thus, a MSS pixel with four reflectance values covers the same area as four TM pixels with a total of 28 reflectance values. The seven fold increase in data from MSS to TM is the twenty eight TM reflectance values divided by four MSS reflectance values. The TM and MSS reflectance values are observations from different spectral band wavelengths. The band wavelengths for TM and MSS are listed below.

MSS		TM	
Band	Microns	Band	Microns
1	0.5 - 0.6 (green)	1	0.45 - 0.52 (blue)
2	0.6 - 0.7 (red)	2	0.52 - 0.60 (green)
3	0.7 - 0.8 (near IR)	3	0.63 - 0.69 (red)
4	0.8 - 1.1 (near IR)	4	0.76 - 0.90 (near IR)
		5	1.55 - 1.75 (middle IR)
		6	10.4 - 12.5 (thermal)
		7	2.08 - 2.35 (middle IR)

The EMSS sampled and averaged data sets with channels 2, 3, 5, 4 have the band combination closest to the MSS data. The other two data sets with band combinations 1, 7, 6, 4 have the critical crop detection band 4 in common.

APPENDIX B

REGRESSION ESTIMATOR

The formulas listed below are used in the DCLC estimates for each strata.

Estimate of the total acres in the scene in a single stratum.

$$\hat{y}_H = N\left[\bar{y} + b\left(\bar{X} - \bar{x}\right)\right] \quad \text{where:}$$

APPENDIX C

HOTELLING'S F-TEST [8]

The selection of an auxiliary variate from among three or more variates is based on maximum correlation of the variates and is conditional on the variates in the sample.

The test is for a specific crop and stratum.

y = reported acreage for crop and stratum

x_i = classified number of pixels for variate i, for crop and stratum

$$i = \begin{cases} 1, \text{ averaged bands } 2, 3, 5, 4 \\ 2, \text{ sampled bands } 2, 3, 5, 4 \\ 3, \text{ averaged bands } 1, 6, 7, 4 \\ 4, \text{ sampled bands } 1, 6, 7, 4 \end{cases}$$

aij= -x^H-i)(yj-y^H-j), covariance of xi and xj

a = [aij], variance covariance matrix

cij = cofactor of aij in the determinant of a

$$wi = / \quad , \quad 1$$

li = covariance y and xi

$$1 = i$$

$$S2\text{^}H1 = (li\ lj - l2\ ) / (p - 1)$$

$$S2\text{^}H2 = (\text{^}H\text{^}i) / (N - p - 1)$$

$$F = S1\text{^}H2 / S2\text{^}H2 ,$$

with (N - p - 1) and (p - 1) degrees of freedom. ^Z