United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Research and
Applications
Division

SRB Staff Report
Number SRB-90-10

June 1990

THE COMPARISON OF EMULATED MULTISPECTRAL SCANNER DATA SETS

James Mark Harris

THE COMPARISON OF EMULATED MULTISPECTRAL SCANNER DATA SETS by James Mark Harris, Research and Applications Division, National Agricultural Statistics Service, United States Department of Agriculture, Washington, D.C. 20250, March 1990. R&AD Staff Report No. SRB-90-10.

ABSTRACT

The National Agricultural Statistics Service used LANDSAT multispectral scanner (MSS) data to aid regression estimates of acreage for major crops during 1980-1987. Future LANDSAT satellites will not have the MSS scanner, but will have a thematic mapper (TM) scanner. TM data is more expensive to purchase and process for large areas because of its higher spatial and spectral resolution. In this study TM data was reduced to resemble or "emulate" MSS data. Four emulations, averaged versus sampled TM data for two different band combinations, were evaluated as possible replacements for the MSS data in crop acreage regression estimates.

KEYWORDS

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

****
*
* This paper was prepared for limited distribution to the research community *
* outside the U.S. Department of Agriculture, The views expressed herein are *
* not necessarily those of NASS or USDA. Use of company names in this *
* publication is for identification only and does not imply endorsement by the *
* U.S. Department of Agriculture. *
*
****

ACKNOWLEDGMENTS

The author thanks the following for their support: E. M. Jones, Jr. and 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 and Michael Craig for their supervision during this project.

TABLE OF CONTENTS

SUMMARY

The Domestic Crops and Land Cover Project (1980-87) of the National Agricultural Statistics Service (NASS) used LANDSAT multispectral scanner (MSS) data in regression estimators of major crop acreage. The MSS data will not be available from future LANDSAT satellites. The next LANDSAT satellite to be launched, LANDSAT 6, will carry the thematic mapper (TM) sensor which has seven times the information per unit area. Past research conducted by NASS indicated that TM data produced a more precise regression estimator of crop acreage. However, since both initial data costs and processing costs were greater for TM data, the cost/benefit ratio favored the MSS data for large area applications. One solution to this problem is to decrease the original TM data volume content to MSS levels (i.e. to "emulate" MSS data from TM data) before further processing. Four possible approaches to creation of emulated MSS (EMSS) data are evaluated in this study.

The TM scanner records seven readings, or bands, for each 30 square meter ground area (the ground area scanned is called a pixel). The MSS data has a 60 square meter pixel with four bands. Four TM pixels, each with seven bands, cover the same area as one MSS pixel with four bands (seven times more information per unit area). Two different data reduction techniques, sampling and averaging, were combined with two different subsets of thematic mapper bands to produce the four different emulated data sets. The sampling approach selected every other TM pixel in every other row (one in four) to represent the EMSS pixel. The averaging approach used the mathematical average of a 2x2 matrix of TM pixels to represent the EMSS pixel. One of the band combinations (or subsets) closely approaches original MSS bands, while the other subset includes more infrared and thermal infrared bands. The emulated data sets were produced by the Earth Observation Satellite Company (EOSAT) and processed through the NASS PEDITOR software.

The TM scene used was an early September date over Columbia, Missouri; the main crops were corn and soybeans. A stratified sample of ground surveyed areas (called segments) were located in the TM scene area. Strata are based on percent of cultivation. Regression relationships were calculated between the ground information and computer classified EMSS data from each segment. Specifically, the number of pixels for each segment classified to a given crop is used as an independent variable or "auxiliary variate" in a linear regression estimator of the planted crop acres. Selection of the "best" emulated data set was based on the precision of the linear regression estimates produced by the data sets.

Results showed some significant differences between emulated data sets at the strata level. When the strata were combined, the differences between emulated data sets would be significant at the .06 (soybeans) and .13 (corn) levels. Differences in EMSS data are attributed to differences in band subsets, and not to the difference between sampling and averaging pixels. The TM band combination closest to the original MSS bands produced the highest sample correlation coefficients for both data reduction techniques. The averaging data reduction technique produced slightly (but not significantly) higher sample correlation coefficients for both corn and soybeans when all strata were combined for the regression. The recommendation is to request the emulation with pixel averaging and band subset closest to original MSS bands. Due to the constantly decreasing cost of processing, it is also recommended to conduct a cost/benefit study comparing the emulated and original TM data.

INTRODUCTION

The National Agricultural Statistics Service (NASS) used LANDSAT multispectral scanner (MSS) data for the Domestic Crop and Land Cover project during the 1980 through 1987 crop seasons [1]. The MSS sensor data will no longer be available when the current generation satellites, LANDSAT 4 and 5, fail to operate. LANDSAT 6, scheduled to be launched in late 1991 by the Earth Observation Satellite (EOSAT) Company, will only carry the Thematic Mapper (TM) sensor. The TM and MSS sensors differ in spatial resolution (ground area imaged) and spectral resolution (wavelengths of reflected light measured).

The basic ground area unit for recording remotely sensed data is called a "pixel". The pixel size is 60 square meters for MSS and 30 square meters for TM. For each pixel, spectral measurements of reflected light are taken in one or more wavelengths ("colors") called bands or channels. The second difference between MSS and TM sensors is in 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 wavelengths for MSS and TM.

For a standard LANDSAT satellite scene, 185 km by 170 km, TM data has seven times the data volume of MSS data due to the sensor differences in pixel size and number of bands. It takes four TM pixels to cover 60 square meters, the ground area of 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 60 square meter ground area. Thus, for the same ground area, TM sensor data has 28 readings versus MSS sensor data's four readings. TM's 28 divided by MSS's four gives the seven-fold data increase for TM sensor data.

Past research performed by NASS indicated that TM sensor data produced a better regression estimator than MSS [8]. However, when costs are taken into account in a cost/benefit ratio, MSS sensor data was preferable. Purchase costs are higher for the raw TM sensor data ($3960 per scene) versus MSS ($1000 per scene). Processing costs are also higher for TM data given the seven-fold data volume increase.

One possibility for reducing costs was to reduce the amount of information in a TM scene. EOSAT agreed to provide NASS with data generated from the TM sensor that had the same number of bands and pixels as MSS. The raw TM sensor data was processed by EOSAT to imitate or "emulate" MSS data, thus the name Emulated Multispectral Scanner (EMSS) data. To create EMSS data, the number of TM pixels have to be reduced to one in four and the number of TM bands are reduced from seven to four.

EOSAT provided NASS with four different EMSS data sets. The data sets cover the Columbia, Missouri area, LANDSAT Path 25 and Row 33. The coverage date was September 5, 1985. Two different pixel reduction techniques were used, averaging and sampling. Averaging takes the average of four TM pixels to produce one EMSS pixel. Sampling took every fourth pixel to represent one EMSS pixel. For each pixel reduction technique, two subsets of TM bands were provided. The TM band subsets were 2,3,5,4 (corresponding to green, red, shortwave or near infrared, and near infrared spectra) and 1,7,6,4 (corresponding to blue, infrared, thermal infrared,

and near infrared spectra). For later reference, the four types of EMSS data were:

Pixel Technique	Spectral Subset
(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 acreage regression estimators for the major crops in the area (corn and soybeans). The study was conducted by processing the four data sets independently through the NASS PEDITOR software system. The PEDITOR system, through various clustering and classification steps, combines raw satellite data and ground gathered ("truth") data for corresponding areas. A regression relationship between computer classified pixels and ground information is calculated for each major crop. The regression estimator uses the number of pixels classified to a crop in a specific area as the independent variable and the crop acres reported by the ground survey as the dependent variable.

STUDY AREA and DATA SET

Ground information for this study was taken from the NASS June Enumerative Survey (JES) in Missouri. The JES design consists of a replicated, stratified sample of land areas called "segments". Enumerators visit the selected segments during June to determine crop acreages. Field boundaries of crops and other land covers within each segment are drawn on aerial photographs as a quality control measure.

There were four different agricultural strata in the Columbia, Missouri study area. The study area and September satellite image date correspond to those used in the 1985 Classifier Study [5]. Nonagricultural areas were not considered for this study nor for 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 [2]. 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 [3]. Within each stratum, replicated samples were drawn. The following table gives the number of segments in the each strata 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 similar judgments 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 and 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 be 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 [4]. 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 the Hotelling test.

Tables I and II present the correlation coefficients and test values by strata for corn and soybeans, respectively. The PROB column gives the probability of observing a greater F-value.

TABLE I: CORRELATION COEFFICIENTS FOR CORN

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	0.9319	0.9402	0.7900	0.7887	7.65	0.00
20	0.8171	0.7749	0.6913	0.6619	1.94	0.15
30	0.8162	0.8275	0.5194	0.2995	1.46	0.23

TABLE II: CORRELATION COEFFICIENTS FOR 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 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.

TABLE III: CORRELATION COEFFICIENTS FOR CORN
COMBINED STRATA

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

TABLE IV: CORRELATION COEFFICIENTS FOR SOYBEANS
COMBINED STRATA

STRATA	AVERAGED 2,3,5,4	SAMPLED 2,3,5,4	AVERAGED 1,6,7,4	SAMPLED 1,6,7,4	F-VALUE	PROB
COMBINED	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 probability 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 coefficient 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 versus 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

TABLE VI: TEST FOR DIFFERENCE BETWEEN BAND COMBINATIONS
SAMPLED 2,3,5,4 versus 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.

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 2,3,5,4 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 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 the cost of the sampled data set a cost/benefit analysis would be appropriate. .

REFERENCES

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

[2] Cotter, Jim. "Area Frame Design Information." U.S. Department of Agriculture, National Agricultural Statistics Service. June 1987.

[3] Cotter, Jim, and Jack Nealon. "Area Frame Design for Agricultural Surveys." U.S. Department of Agriculture, National Agricultural Statistics Service, August 1987.

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

[5] Jones, E. M., Jr. "Classifier Study." Unpublished Paper, U.S. Department of Agriculture, National Agricultural Statistics Service. 1987.

[6] Ozga, Martin. "USDA / SRS Software for LANDSAT MSS-Based Crop-Acreage Estimation." Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS'85), Amherst, Mass, pp 762-779. October 7-9, 1985.

[7] Ozga, Martin, and others, "PEDITOR - A Portable Image Processing System." Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS'86), Zurich, ESA-SP-254. September 8-11, 1986.

[8] Zuttermeister, John Paul. "Evaluating TM Data for SRS Acreage and Production Estimates." SRS Staff Report No. RSB-85-02. U.S. Department of Agriculture, Statistical Reporting Service, July 1985.

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.

Estimates of the total crop acres in the scene in a single stratum.

$\hat{y} = N([\bar{y} + b(\bar{X} - \bar{x})],$, **where**

N = The number of population units in the stratum.

$\bar{y}$ **= The JES sample average reported crop acres in the stratum.**

b = The slope in the regression model of the stratum

$\bar{X}$ **= The population pixel mean in the stratum.**

$\bar{x}$ **= The sample pixel mean in the stratum.**

Estimate of the Variance for each stratum in the scene.

$$\hat{\delta}^2 = N^2 \times \left(1 - \frac{n}{N}\right) \times \left[\frac{\sum (y_i - \bar{y})^2}{(n-2)}\right] \times (1 - R^2) \times \left[1 + \frac{1}{(n-3)}\right]$$

APPENDIX C

HOTELLING'S F-TEST [8]

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

The test is based on a specific crop and stratum.

y = reported acreage for crop and stratum

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

where i =

1, averaged bands 2, 3, 5, 4
2, sampled bands 2, 3, 5, 4
3, averaged bands 1, 6, 7, 4
4, sampled bands 1, 6, 7, 4

$a_{ij} = \sum\sum (x_i - \bar{x}_i)\ (y_j - \bar{y}_j)$, covariance of x_i and y_j,

a = $[a_{ij}]$, variance covariance matrix,

c_{ij} = cofactor of a_{ij} , in the determinant of **a**

$$w_i = \frac{\sum a_{ij}}{\sum\sum a_{ij}}, \sum w_i = 1$$

$$l_i = \sum (x_i - \bar{x})\, y_i \ , \quad l = \sum w_i l_i$$

$$s_1^2 = \frac{(\sum\sum c_{ij} l_i l_j - l^2 \sum\sum c_{ij})}{(p-1)}$$

$$s_2^2 = \frac{\sum (y_i - \bar{y}_i)^2}{(N-p-1)} \ ,$$

$$F = \frac{S_1^2}{S_2^2}, with\ (N-p-1)\ and\ (p-1)\ \text{degrees of freedom}$$