81-20 DRAFT 5/1/81 May 1981

ABSTRACT

SOIL STRATIFICATION EFFECTS ON LANDSAT SIGNATURES

by Paul W. Cook

Results of the statistical analysis of the effects of soil stratification on LANDSAT crop classification in South Dakota are presented. Researchers from South Dakota State University's Remote Sensing Institute (RSI) implemented a soil stratification for eastern South Dakota from which U.S. Department of Agriculture's personnel sampled 252 one-quarter square mile areas to use in analyzing strata effects on crop signatures.

An ANOVA performed on the untransformed LANDSAT pixel data indicated that both corn and sunflower LANDSAT pixel data exhibit statistical differences among the soil strata. However, only sunflowers kept this distinction in R χ^2 test of the soil strata on r^2 of LANDSAT data regressed on ground data improved by the auxilary information from the soil stratification. This result presented the possibility that estimation of minor crops is most improved by the auxilary information from soil stratification. Further studies would be needed to assess this possibility.

ACKNOWLEDGMENTS

I want to thank Ron Steele for establishing the sampling plan for the segments drawn for the project. My thanks go to Barry Ford and Richard Sigman for suggestions used in the statistical analyses.

INTRODUCTION

The special reflectance properties of soil types is well documented, for example (1, 2, 3). This study's objective was to determine the feasibility of using photo-interpretation of LANDSAT data to define strata which would aid in crop discrimination by LANDSAT. This report presents the statistical analyses performed to determine what effects this stratification had on corn and sunflower signatures within a contiguous six county area of South Dakota. Both the ground data and LANDSAT data were obtained during the summer of 1979.

Soil associations derived from LANDSAT photo-interpretation of color-composites are actually homogeneous areas of similar cropping patterns (4). Westin and Lemme define soil associations as "Groups of defined and named taxonomic soil units occurring together in an individual pattern over a geographic area comparable in many ways to plant associations." The satellite derived areas often do follow boundaries which correspond with ground based soil surveys that delineate the standard soil associations as defined above.

LANDSAT reflectance values tend to be less influenced by soil differences on cultivated land since the crops grown there obscure the soil to varying degrees during the growing season. Such factors as soil drainage, profile and landscape do cause differences in planting dates and subsequent crop growth. Such variations in the plants' growth stage and their coverage of the soil should cause differences in the reflectance of light and thereby variations within crop type as observed by LANDSAT. The intention of this study was to evaluate whether or not these variations among strata by crop type were statistically significant for the corn and sunflower data obtained in South Dakota during 1979.

The data analyzed in this report were obtained from Berole, Clark, Codington, Hamlin, Kingsbury, and Spink counties in eastern South Dakota (see figure 1). Ten soil associations were determined by stratification of LANDSAT photographic data. Personnel at the Remote Sensing Institute (RSI) of South Dakota State University used three dates of LANDSAT photographic data to examine the study area under varying conditions of soil and vegetative cover. Prints at 1:500,00 scale from Band 5 were used for the following dates: 14 May 1978, 3 August 1978 and 8 September 1978. Regions of soil type homogeneity were photo-interpretated using tone, texture and qualitative drainage density as the major criteria for determining the soil boundaries.

SRS personnel used standard area sampling frame procedures to select a sample of 252 one-quarter square mile segments within the ten strata. RSI staff made field visits to all sampled segments during July 1979. They collected information on the location of each field, its size, crop type and crop maturity. All data were collected by field observation.

LANDSAT STRATIFICATION PROCEDURES

Interpretation of each LANDSAT scene was done independently to determine the strata definitions and boundaries. Because the 14 May print presented the maximum degree of contrast, all the overlays produced from the other dates were superimposed upon it. Quite clearly this date presented the optimal time for examining bare soil and consequently the most valuable for delineating soil boundaries.

The other two dates from August and September were more influenced by crop cover and land use. Rangeland, small grain, and row crops were the three land uses most prevalent in the study area. Such land use types seem strongly related to both soil capability and moisture regime. Tone and texture were the major two components used in photo-interpreting the strata. Both of these key features were largely determined by the land use types found in the study area. Strata determinations were thus made on a mixture of soil background information derived from the 14 May photo and soil capability obtained from 3 August and 8 September. Greatest weight was given to the 14 May soils information while less importance was given to 8 September and 3 August, respectively.

Although the procedures employed in determining the strata were not dependent upon any other available resource information, a careful examination of the strata revealed a rather close alignment with previously determined soils-geologic boundaries. Table 1 lists the major soils assocation found wihtin each stratum and lists the features by which each stratum is characterized.

THE GROUND DATA

Beadle, Clark, Codington, Hamlin, Kingsbury, and Spink counties were chosen for the study because they presented a rather diverse mixture of soil and crop types. Nearly all the land area was located within one LANDSAT scene so that analysis of the LANDSAT data would be greatly simplified.

A stratified random sample of 252 one-quarter square mile segments were selected within the ten soil strata. Not less than twenty segments were chosen within each to assure an adequate amount of ground data for each crop. A total of nine segments either were outside the scene boundaries or else had field problems which prevented their use.

To ensure the accuracy of the ground data, RSI researchers used current aerial photogrpahy or RBV imagery as necessary to check field locations and crop type. Field sizes were determined by digitizing each field with a coordinate digitizer connected to the software called EDITOR at Bolt, Barenek, and Newman (BBN) in Boston, Massachusetts. Network files containing all the edges which define the field boundaries were also obtained during this process. The computer files provided the means of determining which LANDSAT pixels belonged to each crop category as well as extracting the pixels for each crop type from the computer tapes.

LANDSAT DATA

Two LANDSAT scenes from the summer of 1979 were available for analysis. One scene was taken on July 26th and the other scene on August 25th. LANDSAT scene 21676-16321 taken on August 25th was chosen to be analyzed since it was the optimum time of year for crop should the soil stratification produce further changes in the LANDSAT data means for each crop in the study area, than it might be of value in improving the regression coefficient r^2 between the ground data and the LANDSAT pixel. If this is the case, then the coefficient of variation (r^2) should be reduced.

The LANDSAT scene used was processed in IDIPS format with 57 meter square pixels (picture elements). The area within the analysis area was free of clouds and the north-western corner of Spink county was not on the scene.

LANDSAT DATA ANALYSIS

Registration of the scene was done using computer-generated grey-scales and 7½ minute U.S. Geological Survey (USGS) maps. A total of twenty-six points well scattered throughout the scene were used to generate the coefficients of a bivariate full term third order linear polynomial. The polynomial coefficients were next used to predict the initial locations of the digitized segment data within the LANDSAT scene.

Plots were made of the digitized segment data and placed on grey scales that had been printed out for each segment's predicted location. By using aerial photographs, maps and plots of the segments, the researchers determined the correct location of each segment based on the lightness-darkness values of the greyscales and the crop types within each field. Maximum movements of about three pixels were necessary for segments located on 2 degree (1:250,000 scale) maps whereas segments located on 15 minute or 7½ minute (1:62,500 or 1:24,000 scale) maps required at most 1 pixel movements in row and/or column.

After segment movement was complete, a program was run in EDITOR which causes the creation of a computer tape which contained each LANDSAT pixel data value along with the identifying stratum, segment, and field number. This tape was then processed on the IBM 370-168 located at the Washington Computer Center (WCC). Programs incorporated in the Statistical Analysis System (SAS) were next used in analyzing the data statistically.

Calcualted sums of squares as determined by the SAS program are presented for both corn and sunflower in Appendix A. The model used for the test, however, is a nested analysis of variance in which segments and fields are random effects, the strata are fixed effects, and the data are unbalanced. This model is given in more detail in Appendix B.

The development of the F test used to determine which strata were significantly different from each other is also given in Appendix B. The results of the F test are given in Table 2. Quite clearly, both bands 4 and 5 show significant differences by strata. These results indicate reflectance values for both corn and sunflower within the study area during late August. were statistically different between strata.
~~System (SAS) were next~~ used ~~in analyzing the data statistically.~~

A significant difference between strata for corn and sunflower untransformed LANDSAT data is not sufficient for SRS to determine what effects the stratification will have on crop acreage predictions. Consequently, the next test of the strata information's usefulness is of its effect on the regression coefficient's r^2 since SRS uses a sampling regression equation to produce crop acreage estimates (6).

Since the regression estimator's variance is a function of $(1-r^2)$ times the variance of the direct expansion estimate as normally used with the JES segment data; larger r^2 values will result in reduced variance of the regression estimator (6, Appendix B). The relative efficiency of the regression estimator will thereby be increased. That is, the equivalent sampling size needed by the direct expansion estimator for the same degree of accuracy would be considerably larger.

The presence of strata differences would mean that the regression estimate would be more accurate for some strata. Detection of such differences among strata is thereby the next condition for which testing is of interest.

Training data to develop the classifier were obtained from the pixels interior to the fields of the segments. All the not background (that is, field interior and boundary) pixels were used to test the classifier for percent correct and to calculate r^2.

Clustering of the crop and land use pixels from the segments proceeded in two ways. Either each individual crop or land use was selected within the sample segments or else all the grasses and minimally represented crops were combined together as another category. The objective was to produce the highest possible r^2 using the sampled data.

Clustering by individual crop gave a total of 53 categories, whereas, the combined crop method gave 63 categories in the statistics files generated.

Both equal priors and prior probabilities (calculated from the proportions of pixels per crop in the segments) were applied to the statistics files created from clustering.

Classification of the not background file of pixels from all the segments was done using each of the resulting statistics files. The best results from this test of classifiers resulted from the use of the statistics file containing 53 categories with equal priors. Table III presents the results from this classification using the 53 category statistics file.

Test of the hypothesis that there is no difference among the strata was done using the χ^2 test as given in Appendix C. Preparatory work in performing the test is in Tables ~~IV~~ III and ~~Table III~~ IV. Natural logarithms of the r's calculated from the r^2's are in Table III while the transformed Z's are in Table IV. The number of segments within each stratum for each of the two crops under examination are listed below the Z's. Finally, the J_i's, i=1,...,I from which are calculated the J_i's ~~in using~~ for use of the test statistic of equation (1.0) are also presented.

Table V lists the results of the χ^2 test. The most prevalant crop, corn, does not show a difference among strata. Sunflower, which is considerably less available, however, does exhibit highly significant differences among strata r^2 values.

These results indicate that crop management practices become more important in determining the accuracy of crop estimation for relatively minor crops that for major crops. It is quite possible that soil stratification would provide better estimation for minor crops. It would be necessary, however, to test this hypothesis using other areas where different crops are grown to determine if this effect extends to other crops as well as it does to sunflower.

CONCLUSIONS

The Analysis of Variance test of the untransformed LANDSAT reflectance values indicates that the stratification criteria produced statistically significant differences in the LANDSAT scene acquired data during the crop growing season for corn and sunflower. Crop management techniques as well as soil variability (especially lightness, darkness caused by organic matter content) are the most probably causes of these differences. There is presently insignificant data available to decide the relative importance of such soil association variables in this study. It would be necessary to provide a more detailed soil stratification to assess the relative importance of each soil component.

The results of the χ^2 test indicate that spectral differences do not always produce statistically different r^2's between the classified LANDSAT data and the ground data. The corn data indicated a stronger rejection for the Analysis of Variance test that did the sunflower data, however, the χ^2 test indicated that the soil stratification would not aid in making estimates of the corn acreage. This quite possibly could be due to the near-infared reflectance being more a determination in distinguishing corn that it was in distinguishing sunflower.

Another possibility to consider is that soil stratification would be most useful in aiding the estimation of minor crops. Since minor crops occur in fewer areas, the soil stratification might allow for differences in crop management and soils to aid in separating minor crop data that would otherwise be confused with major crop information. This should allow for more accurage estimation of both crops by reducing the amount of special overlap between the crops.

APPENDIX A

Crop	Source	BAND 4 SS	BAND 4 df	BAND 4 MS	BAND 5 SS	BAND 5 MS	BAND 6 SS	BAND 6 MS	BAND 7 SS	BAND 7 MS
Corn	C	7142.955	9	793.662	7931.334	881.259	25961.033	2884.559	30969.374	3441.042
	B(within C)	4866.459	91	53.478	17233.582	189.380	103145.623	1133.468	173658.716	1908.338
	T(within B)	141.566	26	5.445	291.764	11.222	6753.344	259.744	9569.195	368.046
	Error	6491.879	4937	1.315	15553.508	3.150	49001.184	9.925	75751.288	15.344
TOTAL		18642.859	5063		41010.189		184861.184		289948.573	
Sunflowers	C	2227.774	6	371.296	2125.121	354.187	28198.753	4699.792	26631.376	4438.563
	B(within C)	639.140	13	49.165	1561.887	120.145	30059.815	2312.293	36926.083	2840.468
	T(within B)	38.439	5	7.688	236.017	47.203	3432.519	686.504	4480.652	896.130
	Error	1497.241	1437	1.042	245.915	1.702	38058.872	26.485	45393.887	31.589
TOTAL		4402.594	1461		6368.941		99749.959		113431.997	

DRAFT

APPENDIX B

A fully nested mixed model of the following form was used to test the untransformed ~~raw~~ LANDSAT pixel data:

$$Y_{ijkm} = \mu + \alpha_i + \beta_{ij} + \tau_{ijk} + \xi_{ijkm}$$

Where

$i = 1,\ldots I$ ~~where~~ and I is the number of soil strata for a given crop,

$j = 1,\ldots, J_i$ ~~where~~ and J_i is the number of segments within stratum $i, i = 1,\ldots I$,

$k = 1,\ldots, K_{ij}$ ~~where~~ and K_{ij} is the number of fields within segment $j, j=1,\ldots, J_i$ for stratum $i, i=1,\ldots I$.

$m = 1\ldots, M_{ijk}$ where M_{ijk} is the number of pixels for each segment,

We note that $\sum_i K_{ij} = K_i, \sum\sum_{ij} K_{ij} = \sum_i K_i = K$ where K is the total number of fields in all segments, M_{ijk} is the number of pixels for each segment and $\sum_i J_i = J$ is the total number of segments in all strata. The (α_i) are the strata, the (β_{ij}) are the segments within strata, the (τ_{ijk}) are the fields within segments within strata, and the (e_{ijkm}) are the error components derived from the LANDSAT measurements. The variables are independently normal with zero means and variances (6). Both the (β_{ij}) and (τ_{ijk}) are random effects in the model whereas the (α_i) are fixed effects.

Only two crops were still at a stage of maturation which would provide a meaningful analysis. Results for the corn and sunflower data are thereby presented in Appendix A. Because the model is nested, the segment and fields are random effects and the data is unbalanced by cell, an approximate F test developed by Satterhawaite was used to test the following hypothesis:

$H_{o1}: C_1 = C_1 = C_2 = C_3 = C_4 = C_5 = C_6$ by using

$$\frac{MS_c}{B_T\,(MS_T) + B_B(MS_B) + B_e\,(MS_e)} = \frac{MS_c}{\widetilde{MS}} = F_{I-1,\,\nu}$$

$$\text{where } \nu = \frac{[B_T\,(MS_T) + B_B\,(MS_B) + B_e\,(MS_e)]^2}{\frac{(B_T MS_T)^2}{\sum\sum_{ij}(K_{ij}-1)} + \frac{(B_B MS_B)^2}{\sum_i(J_i - 1)} + \frac{(B_e MS_e)^2}{\sum\sum\sum_{ijk}(M_{ijk} - 1)}}$$

and

$$B_T + B_B + B_e = 1$$

$$B_T \cdot \frac{n - \sum\sum_{ij} \frac{T_{ij}}{n_{ij}}}{\sum\sum_{ij}(k_{ij} - 1)} + B_B \cdot \frac{\sum\sum_{ij} \frac{T_{ij}}{n_{ij}} - \sum_i \frac{T_i}{n_i}}{\sum_i (J_i - 1)} = \frac{\sum_i \frac{T_i}{n_i} - T/n}{I - 1}$$

$$B_B \cdot \frac{n - \sum_i B_i/n_i}{\sum_i (J_i - 1)} = \frac{\sum_i B_i/n - B/n}{I - 1}$$

where $T_{ij} = \sum_k M^2_{ijk}$, $T_i = \sum_j T_{ij}$, $T = \sum_i T_i$, $B_i = \sum_i n^2_{ij}$, $B = \sum_i B_i$, $C = \sum_i n^2_{i.}$

NOTE: $\widetilde{MS}$ is known as Satterhawaite's Approximation.

APPENDIX C

To determine the effectiveness of the soil stratification in improving classification accuracy, an examination of the differences among the strata r^2_i value is made. First, the r^2 had to be converted to Normal $(\mu = \frac{1}{2} \ln \frac{1+r}{1-r}, \sigma^2 = \frac{1}{n-3})$ by taking square roots and then using the formula $z = 0.5 \ln (1+r_i) - 0.5 \ln (1-r_i)$ [8]. Since the variations of these normalized scores are $S_{z_i} = \frac{1}{\sqrt{n_i - 3}}$, the number of observations made was reduced by three for each stratum, that is

$J_i' = J_i - 3$ for $i = 1, 2, \ldots, I$ where

J_i = the number of segments in stratum i

$i = 1, 2, \ldots, I$, for I = the number of strata.

Consequently, we could then consider $J_i, i = 1 \ldots, I$; z_i, $i = 1, \ldots, I$, $\bar{z} = \sum_{i=1}^{I} J_i' z_i / I$, were I = number of strata for each individual crop, than

$$(1.0) \quad \sum_{j=1}^{I} \left(\frac{Z_i - \bar{Z}}{1/\sqrt{J_i'}} \right)^2 = \sum_{i=1}^{I} J_i' (Z_i - \bar{Z})^2 .$$

The sum is χ^2_{I-1} since Z_i is a Normal (0,1) random variable. Under the H_0 assumption that there is no difference among the strata r^2's, the calculated $\sum_{i=1}^{I} J_i'(Z_i - \bar{z})^2$ value should be less than the tabular χ^2_{I-1} value since larger values would indicate that for the crop under consideration the r^2's are not equal between strata.

REFERENCES

1. Dalsted, K. J., K. K. Worcester, and M. E. DeVries, Influence of Soils on LANDSAT Spectral Signatures of Corn, Eighth Annual Remote Sensing of Earth Resourses Conference. The University of Tennessee Space Institute, Tullahoma, 1979.

2. Stoner, Eric R. and Baumgardner, Marion F., Physiocochemical, Site, and Bidirectional Reflectance Factor Characteristics of Uniformly Moist Soils, Laboratory for Applications of Remote Sensing, Purdue University LARS Technical Report 111679, February 1980.

3. Weismiller, R. A., Kirschner, F. R., Kaminsley, S. A., Hinzel, E. J., Spectral Classification of Soil Characteristics to Aid the Soil Survey of Jasper County, Indiana. LARS Technical Report 040179. Laboratory for Application of Remote Sensing, Purdue University. West Lafayette, Indiana.

4. Westin, F. C., Lemme, G. D., LANDSAT Spectral Signatures: Studies with Soil Associations and Vegetation, Photogrametric Engineering and Remote Sensing, 44:315-325, 1978.

5. Gleason, Chapman, and others, The Auxilary Use of LANDSAT Data in Estimating Crop Acreages: Results of the 1975 Illinois Crop-Acreage Experiment, USDA, Statistical Reporting Service, SRS-21, October 1977.

6. Scheffe, Henry, The Analysis of Variance, Wiley, 1959.

7. Dunn, O. J., Clark, V. A., Applied Statistics, Analysis of Variance and Regression, Wiley, 1974.

8. Hanuschak, G. A., Sigman R., and others, Obtaining Timely Crop Area Estimates Using Ground-Gathered and LANDSAT Data, U.S. Department of Agriculture, Economics, Statistics, and Cooperatives Service, ESCS Technical Bulletin, No. 1609, August 1979.

9. Sterling, Theodor D., Pollack, Seymour V., Introduction to Statistical Data Processing, Prentice-Hall, pages 408-409, 1968.

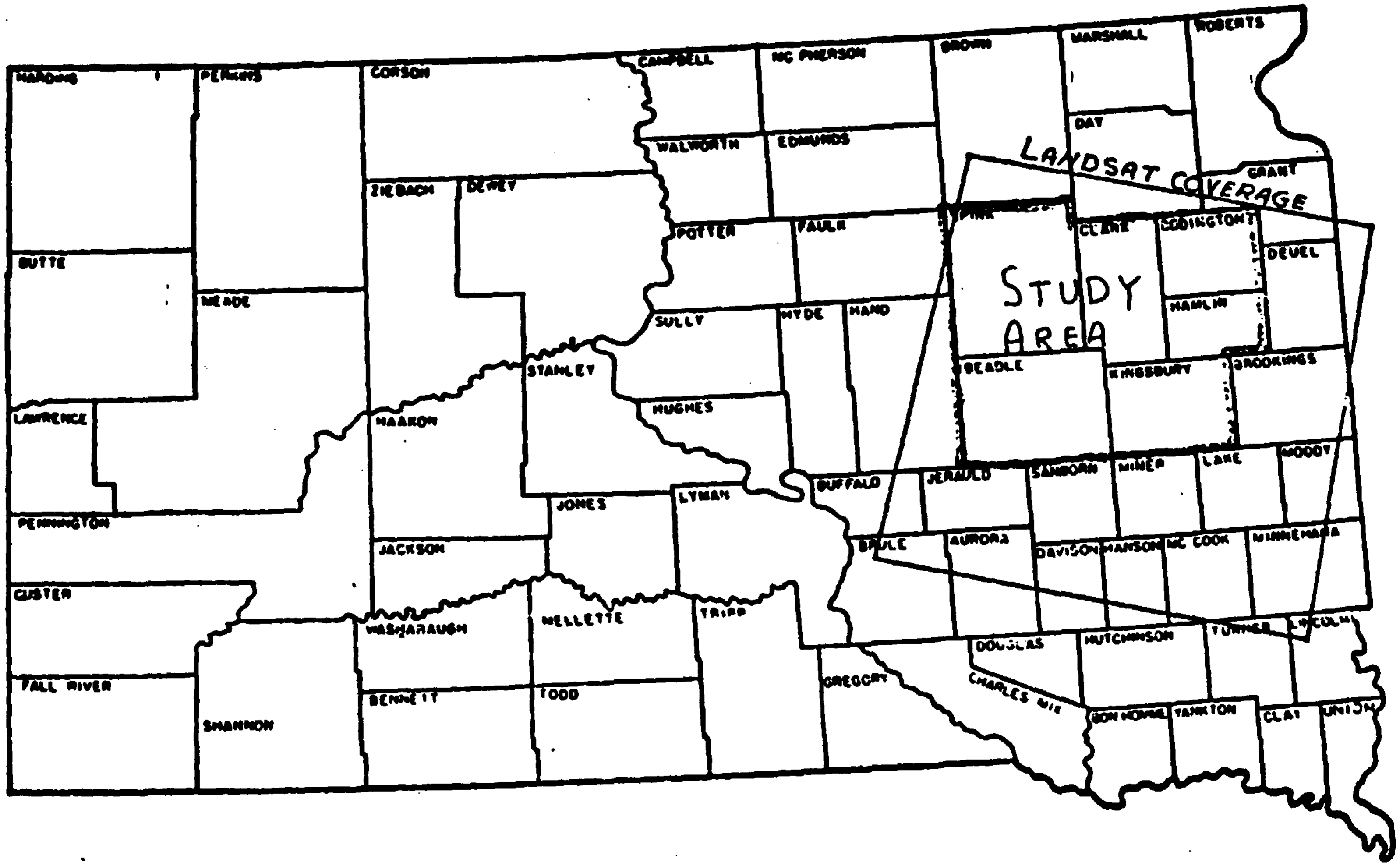

Figure 1: Location of study area in crop-soil spectral study. Boundaries of the Landsat scene and the counties included in the study are outlined.

Table I. Physiographic criteria for delineation of Landsat strata used in ESCS-RSI study.

Stratum	Major Physiographic Unit	Glacial Geology	Major Soil Associations	Major Characterizing Feature on Landsat†
1	James Basin	Late Wisconsin Ground Moraine (LWGM)	Houdek-Prosper Blendon-Enet	Land use is dominated by range; tones associated with range.
2	Lake Dakota	Lake Dakota (lake deposits)	Harmony-Aberdeen Beotia Beotia-Harmony	Light tones, intensified area of cropland, drainage pattern indicates very little slope.
3	James Basin	(LWGM)	Houdek-Prosper	Darker tones than Unit 1 and lighter than Unit 3. Has scattered speckling of potholes.
4††	James Basin	(LWGM)	Blendon-Enet	Light tones.
5	James Basin	(LWGM)	Beadle	Similar to Unit 6 with speckling of potholes, darker tones than Unit 3.
6	James Basin	(LWGM)	Peever-Forman	Similar to Unit 5 without potholes.
7	Coteau des Prairies	Late Wisconsin Stagnation Moraine	Poinsett-Waubay-Parnell	High density of potholes.
8	Coteau des Prairies	Late Wisconsin End Moraine	Vienna-Lismore	Similar to Unit 7 with lower density of potholes, but higher density of potholes than Units 3 or 5.
9	Coteau des Prairies	Early Wisconsin Drift (EWD)	Kranzburg-Brookings	Dark tones associated with floodplain and adjacent productive soils.
10††	Coteau des Prairies	(EWD)	Vienna-Lismore	Similar to Unit 8 with lesser density of potholes.

†Characteristics refer to 14 May 1978 Landsat band 5. Other dates were used to refine boundaries.

††Limited occurrence in the study area.

DRAFT

TABLE II

CROP	PIXELS	BAND	MS	df	$\widetilde{MS}$	df	F RATIO	$F_{\alpha=.05;I-1,\nu}$
Corn	5064	4	793.662	9	74.592	92	10.64*	2
		5	881.259	9	263.526	92	3.34*	2
		6	2884.559	9	1574.702	92	1.83	2
		7	3441.042	9	2650.334	92	1.32	2
Sunflower	1462	4	371.296	6	44.333	13	8.38*	2.92
		5	354.187	6	104.129	12	3.40*	3.00
		6	4699.792	6	2034.054	12	2.31	3.00
		7	4438.563	6	2490.989	12	1.78	3.00

* Indicates the F ratio is significant at the α=.05 level of significance.

DRAFT

TABLE III

R^2 by Stratum using 53 Categories with Priors Obtained Clustering Across Strata by Crop

CROP	STRATA										COMBINED EST.
	1	2	3	4	5	6	7	8	9	10	
Corn	0.79	0.74	0.70	0.77	0.58	0.42	0.61	0.68	0.64	0.50	0.62
wheat	—	0.51	0.69	0.22	0.03	0.32	0.37	0.35	0.03	0.16	0.39
Sunflowers	—	0.95	—	—	—	—	0.29	—	0.62	—	0.67
Rangeland	0.18	0.55	0.70	0.49	0.11	0.04	0.53	0.57	0.181	0.121	0.36

CROP	PERCENT CORRECT
Corn	66.0%
Wheat	38.5%
Rangeland	69.2%
Sunflowers	60.7%
Alfalfa	21.4%
Barley	0.9%
Flax	20.9%
Oats	21.5%
Hay Cut	0.6%
Grass	4.8%
Wetland	30.9%
Farmstead	4.2%
Fallow	18.4%
OVERALL	43.0%

TABLE IV

A. Strata r^2 as Transformed to Normal Z Scores

Stratum	Corn	Sunflower
1	1.41269	--
2	1.08230	2.10493
3	1.29334	--
4	1.47222	--
5	0.96846	--
6	0.96846	--
7	1.13299	0.77187
8	1.15384	--
9	1.13299	1.11566
10	1.08230	--

B. Number of Segments Within Each Stratum by Crop (J_i)

Stratum	Corn	Sunflower
1	20	--
2	19	19
3	19	--
4	20	--
5	18	--
6	40	--
7	48	48
8	20	--
9	19	19
10	18	--

DRAFT

TABLE V

Calculated χ^2 Test Values for the Four Land Cover in the Test Area

LAND COVER	$\Sigma J_i'(z_i-\bar{z})^2$	$\chi^2_{\alpha; I-1}$
Corn	4.96	$\chi^2_{9, \alpha = .10} = 14.70$
Wheat	7.83	$\chi^2_{8, \alpha = .10} = 13.40$
Sunflowers	23.91 *	$\chi^2_{2, \alpha = .005} = 10.60$
Rangeland	25.23 *	$\chi^2_{9, \alpha = .005} = 23.60$

* = Significant at the α=.005 level