AUXILIARY USE OF LANDSAT DATA
IN ESTIMATING CROP YIELDS

R. S. Sigman
G. A. Larsen

Statistical Research Division
Economics and Statistics Service
U. S. Dept. of Agriculture

December, 1980

AUXILIARY USE OF LANDSAT DATA IN ESTIMATING CROP YIELDS. By R. S. Sigman and G. A. Larsen; Statistical Research Division, Economics and Statistics Service, U. S. Dept. of Agriculture, December 1980.

ABSTRACT

In this report a regression-like estimator is investigated as a method to use LANDSAT data to improve ESS objective yield (OY) estimates for corn and soybeans. The estimator's primary variable, which is required to be known only for sampled fields, is estimated field-level yield computed from observed plot data. The estimator's auxiliary variables are field-level means of MSS radiometric values and/or various MSS vegetative indices. By definition, auxiliary variables must be known over the entire population, which in this case is all land planted to the crop of interest within some target area. Since a pixel's population membership is not known for pixels exterior to June Enumerative Survey (JES) segments, the set of all pixels classified to the crop of interest is used to define a pseudo-population for the estimator. This creates an estimator bias which is estimated from labeled LANDSAT data coinciding with JES segments.

Evaluation of the developed estimator with 1978 unitemporal Iowa data produced mixed results in sub-state analysis areas. In some areas of Iowa, no yield estimation improvements from LANDSAT were indicated. In other parts of Iowa, yield estimation improvements were moderate for soybeans and marginal for corn. Haze correction was used to develop entire-state estimators. Entire-state estimation improvements were modest for both corn and soybeans. However, if LANDSAT data are available from the acreage estimation program at no additional cost then some savings can be realized in the OY program through reduced sample size.

CONTENTS

BACKGROUND

Yield Data Collection

Annually from late May to early June, ESS conducts a nationwide agricultural survey referred to as the June Enumerative Survey (JES). This survey consists of interviews with farm operators in randomly sampled areas of land called segments. These interviews include a field by field enumeration of land use and acreage within segments. These segments are selected by stratified random sampling from the population of all segments in a given land area. The strata are land use categories determined by visual interpretation of aerial photography or LANDSAT imagery and deliniated on county highway maps. The segments are typically one square mile in size depending on the availability of distinguishable boundaries. For Iowa, the JES includes a sample of approximately 300 segments.

Another type of survey conducted by ESS is the Objective Yield Survey (OYS). The OYS begins around the time of initial fruit development for the major crops and is repeated at monthly intervals until harvest. The purpose of the OYS is to collect objective plant counts and measurements to aid in making large-area yield forecasts and estimates with measurable precision. The OYS sample is selected by systematic random sampling of individual acres from all JES acres identified as planted to the crop of interest. The selected acres thus identify particular fields with selection probability proportional to size (PPS). Very large fields may be selected more than once. (Strictly speaking, OYS fields are selected only approximately PPS since very large fields are selected with probability 1, which is not PPS.)

An OYS-sampled field (with multiple selections of the same field being considered as different sampled fields) is called an individual "sample" by the OYS operational program. The OYS sample size in Iowa in 1978 was 240 for corn and 170 for soybeans. Each sample consisted of two randomly and independently selected plots. For corn, each plot was two rows in width and 15 feet in length. The soybean plots were 2 rows in width and 3.5 feet in length.

The OYS is designed to provide precise yield forecasts and estimates (approximately 2% coefficient of variation) at the state level. Since more variation in grain yield occurs between fields than within fields, the most precise state level estimates are generally obtained by having a large number of samples relative to plots within each sample. While this method is well suited to large-area yield estimation, it does not provide the data necessary to make good field-level estimates since each field typically contains only two OYS plots. The OYS pre-harvest data consists of counts, measurements, and weights for plants within OYS plots. These data permit estimation of gross grain yield and certain components of yield. For corn, the yield components are number of ears per unit area and grain weight per ear. Soybean yield components are number of plants per unit area, number of pods per plant, and bean weight per pod.

During the OYS post-harvest interview, the farm operator is asked for harvested acreage and production. For soybeans the operator is also asked to specify the moisture content of the beans he harvested. The corn interview does not ask for the moisture content. The post-harvest interview therefore provides a separate estimate of net grain yield. A disadvantage of the farmer reported yield is that its accuracy cannot be measured. The post-harvest interview is usually conducted in a quarter of the samples for corn and half of the samples for soybeans but in 1978 the interview was performed for all corn and soybean samples.

An additional type of data collected during the OYS is for harvest loss determination. Harvest loss samples were laid out in one-fourth of the corn sample fields and one-half of the soybean sample fields. The harvest loss data combined with the pre-harvest data provide an OYS estimate of net grain yield.

Since participation in OYS is voluntary, there are always a certain number of OYS samples which are lost due to the refusal of the farm operator to cooperate with the survey. In addition to refusals, weather and economic factors may cause the loss of samples. As long as the operator continues to cooperate, OY counts ard measurements are made without regard to the condition of the crop. In some cases the farmer may decide to plow up his field or, in the case of corn, cut the crop for silage. In these cases, OY data for the final pre-harvest visit is lost. Also, occasionally OY data are obtained but farmer yield is not reported.

LANDSAT Data

The basic element of LANDSAT data is the set of measurements by the satellite's multispectral scanner (MSS) of a 0.4 hectare (1.1 acre) area of the earth's surface. The MSS measures the amount of radiant energy reflected and/or emitted from the earth's surface in four different regions, called bands, of the electromagnetic spectrum. The MSS bands are designated MSS4 (green), MSS5 (red), and MSS6 and MSS7 (near infrared).

The individual 0.4 hectare MSS resolution areas, referred to as pixels, are arrayed along east-west running rows within the 185 kilometer wide north-to-south pass of the LANDSAT satellite. A given point on the earth's surface is imaged at least once every eighteen days by the same LANDSAT satellite. At the time of this study, there were two LANDSAT satellites in orbit with a nine day separation so that any given point on the earth would have been imaged at least every nine days. Satellite passes which are adjacent on the earth's surface are at least one day apart with respect to their dates of imagery and have a 35% or more sidelap.

Crop Areas from LANDSAT Data

The utility of LANDSAT data in improving crop hectarage statistics for multi-county areas and individual states has been demonstrated by a number of ESS remote sensing projects. These studies have been conducted for entire states in Illinois, Kansas, Iowa, and Arizona and for sub-state areas in California, Arkansas, South Dakota, and Missouri (4, 5, 6, 7, 8, 13, 15).

For estimating crop areas, ESS's approach to using LANDSAT data is as a supplementary data source to farmer-reported crop-type and field-size data collected by ESS's June Enumerative Survey. The JES data, which indicate crop type by field, plus the corresponding LANDSAT data determine discriminant functions that are used to classify LANDSAT pixels as to probable crop type. Crop areas are then estimated by a regression estimator having farmer-reported acres as the primary variable and the LANDSAT classification results as the auxiliary variable. The geographical region constituting the domain of a regression estimator is called an analysis district.

Yield and Yield Indicators from LANDSAT Data

A number of researchers have found that LANDSAT data, or various derived LANDSAT indices, are significantly correlated with yield or, alternatively, with yield indicators such as leaf area index and green biomass (3, 16, 18). The majority of these studies, however, have been conducted with winter wheat. Some of the derived LANDSAT indices that have been used in such studies are listed in Table 1.

Table 1 - LANDSAT indices.

Name	Definition
Sum	SUM = MSS7 + MSS5
Difference	DIFF = MSS7 - MSS5
Ratio	RATIO = MSS7/MSS5
Vegetative Index	VI = DIFF/SUM
Transformed Vegetative Index	TVI7 = $(VI + .5)^{1/2}$
Green Vegetative Index	GVI = -.290*MSS4 - .562*MSS5 +.6*MSS6 + .491*MSS7
Perpendicular Vegetative Index	PVI = $((RGG2 - MSS5)^2$ $+ (RGG4 - MSS7)^2)^{1/2}$

	where
	RGG2 = .851*MSS5 + .355*MSS7
	RGG4 = .355*MSS5 + .148*MSS7

DATA SOURCES

LANDSAT and JES Data

For purposes of estimating crop areas from LANDSAT data, ESS in 1978 processed twelve LANDSAT scenes covering the entire state of Iowa (Figure 1 and Table 2). In support of this effort, a data set of farmer-reported field sizes and cover types for fields within 1978 corn and soybean objective-yield data were used to evaluate yield estimation with the biased regression estimator.

Table 2 - Dates of LANDSAT imagery, Iowa project, 1978.

Path	Row	Date	Percentage Iowa cloud cover	Scene ID
30	30	August 19	0	30167-16274
	31	August 19	0	30167-16280
29	30	August 9	0	21295-16013
	31	August 9	40	21295-16020
	32	August 18	0	30166-16224
28	30	September 4	60	30183-16162
	31	September 4	0	30183-16164
	32	September 4	0	30183-16171
27	30	August 7	10	21293-15500
	31	August 7	15	21293-15502
	32	August 7	10	21293-15505
26	31	August 6	0	21292-15444

OYS Data

For the yield and LANDSAT study, more precise field-level yield estimates were required than available from the OYS. Consequently, for each OY sample three additional samples (six plots), called research samples, were established on the final pre-harvest visit.

The OY sample is generally laid out by counting random numbers of rows and paces from the most accessible corner of the field. The rows and paces are based on field size and determined in such a way that plots are randomly selected from the set of all plots contained in a quarter of the field. The research samples were similarly located from the other three corners of the field. (Variations on this theme occurred in the few fields which were irregular in shape.) The research plots were identical to the OY plots with the exception that the 6-inch counts in the soybean plots were not done since the 6-inch counts pertain primarily to early-season yield forecasts and were thus not relevant to the research study's objective of improving end-of-season yield estimates.

Of the initial 170 soybean samples, 126 had both OY data and farmer reported yield. Out of these 126 fields, 72 had harvest loss data. Of the 240 initial corn samples, 166 had OY data and farmer reported yield. Of these, 45 fields had harvest loss data.

All of the yield data were hand and machine edited for keypunching accuracy and reasonableness. The machine edit checked numerous relationships among the data to make sure that counts, measurements and weights were not unreasonable.

Yield Estimation

The most direct way to estimate the field-level gross yield is to take the weight of grain and the corresponding ground area and compute a weight per unit area for each sample in the field. With just the OY sample, the sample level yield would also be an estimate of the yield for the field. A simple average of the OY sample yield and the three research sample yields provides another, more precise field level estimate. When only the OY sample is used, the estimate is called the OY estimate. When all four samples within a field are used, this is referred to as the research yield estimate.

Field level variance was estimated by considering each quarter of the field to be a stratum and each sample to be a random sample within the strata. Since each sample was restricted to lie within a particular quarter of the field, the stratified variance estimate was preferred to assuming the eight units were a simple random sample. Since the probability of an OY sample falling within a certain field was proportional to the size of the field, the gross yield was estimated at higher aggregate levels by taking the simple average of the individual field level means. These mean yields are referred to as direct expansion estimates. The associated variance was estimated by the usual variance formula for a simple random sample. Within and between field components of variance could also be calculated if desired.[1/]

The direct expansion yield estimates could be made with either the OY samples or with the research samples as well. At the state level, these two estimates were very close. Net yield was estimated by subtracting the mean harvest loss. A third yield estimate was obtained by averaging the farmer reported yields. For soybeans, the net direct expansion yield was close to the farmer yield. For corn, this was not the case. However, the farmer yield could not be adjusted to the same standard moisture level which made comparison difficult. Table 3 shows means and variances for corn and soybeans at the state level.

[1/] Specific details concerning the estimation of field and higher aggregate level means and variances from OY data are contained in an internal report (12). Copies may be obtained from the Yield Research Branch.

Table 3 - Means and variances for various state-level yield estimators (bu/acre).

	Number of Fields	Mean	Variance of Mean	Coefficient of Variation
Soybean Yield				
Net Research	126	38.13	.77	2.3%
Net OY	126	37.63	1.28	3.0%
Farmer	126	38.04	.73	2.2%
Corn Yield				
Net Research	166	125.48	4.58	1.7%
Net OY	166	124.04	6.04	2.0%
Farmer	166	121.03	3.27	1.5%

Yield - LANDSAT Data Set

Because of unusable LANDSAT data due to cloud cover, only 144 corn fields and 98 soybean fields had OY data, farmer reported yield information, and also LANDSAT data. After yield and LANDSAT data were obtained for the same set of fields, an additional edit was performed. The number of acres in each field was estimated using associated LANDSAT pixel counts. This estimate was compared to the farmer reported planted and harvested acres to spot possible problems with the LANDSAT data not corresponding closely with the same area the yield was estimated from. Several mismatches were discovered which could sometimes be explained by a large difference between farmer reports of planted and harvested acres. Other mismatches could not be explained. As a result several fields were omitted from the final yield-LANDSAT data set used for this study. The final number of soybean fields totaled 96 while the corn fields dropped to 139. Table 4 summarizes the OY fields by type of field-level data.

Table 4 - Distribution of samples for yield and LANDSAT data

	Number of Soybean Samples	Number of Corn Samples
Initial	170	240
Types of Data1/		
OY, FY	126	166
OY, FY, HL	72	45
OY, FY, LS	98	144
Final OY, FY, LS	96	139

1/ OY - Objective yield
FY - Farmer reported yield
HL - Harvest loss
LS - LANDSAT

Haze Correction

A number of observation conditions can significantly alter LANDSAT data by changing the relationship between the actual reflectance at the crop canopy and the reflectance represented by the LANDSAT digital counts. These observation conditions include viewing and illumination geometry, amount and distribution of haze in the atmosphere, amount of water vapor, and amount and height distribution of cirrus clouds. Varying degrees and combinations of these conditions make LANDSAT data difficult to interpret and can obscure or distort any true relationship between yield and spectral reflectance which under uniform observation conditions might otherwise be discernable. The XSTAR haze correction algorithm, developed by Lambeck (10, 11), attempts to correct LANDSAT data to a standard set of observation conditions. Because of the different Iowa LANDSAT dates and hence different observation conditions, XSTAR correction was investigated as a means of improving LANDSAT-yield relationships over multiple image dates. Refer to Appendix C for a discussion of atmospheric effects and correction techniques.

STATISTICAL METHODOLOGY

A regression estimator makes use of supplementary data having a known population mean. However, since the population for yield estimation is all analysis-area fields planted to the crop of interest, it is not possible to calculate a LANDSAT regression estimator for crop yield. This is a result of lack of knowledge of field boundaries outside of JES segments, which precludes the calculation of the proper population means for LANDSAT variables.

Consequently, an alternative estimator, called a biased regression estimator, is derived in Appendix A. For LANDSAT variables this estimator defines a pseudo-population, consisting of all analysis-area pixels classified to the crop of interest (17). The form of the biased regression estimator is as follows:

$$\hat{\bar{\bar{P}}}_{reg} = \text{biased regression estimate of population-level yield}$$

$$= \hat{\bar{\bar{P}}}_{DE} + \underline{\hat{B}}' \, (\underline{\bar{\bar{W}}} - \underline{\bar{x}}^{*})$$

where

$\hat{\bar{\bar{P}}}_{DE}$ = direct expansion estimate of population-level gross yield calculated from plot data only

$$= \sum_{i=1}^{n} \hat{y}_i/n$$

$\hat{y}_i$ = estimate of yield for sampled-field i calculated from plot data only

n = number of sampled fields in analysis area

$\underline{\bar{x}}^*$ = vector of sample means of sampled-field means per pixel for LANDSAT measurements and indices

$$= \sum_{i=1}^{n} \underline{x}_i^*/n$$

$\underline{x}_i^*$ = vector of sampled-field i means per pixel for LANDSAT measurements and indices

$\underline{\hat{B}}'$ = vector of estimated regression coefficients for regression of $\hat{y}_i$ on $\underline{x}_i^*$

and $\underline{\bar{\bar{W}}}$ = pseudo-population mean per pixel for LANDSAT measurements and indices.

Since the number of pixels used to calculate $\underline{x}_i^*$, the vector of sampled-field i means per pixel for LANDSAT measurements and indices, is a measure of size of sampled field, the expectation of $\underline{\bar{x}}^*$ is

$\underline{\bar{\bar{X}}}$ = population mean per pixel for LANDSAT measurements and indices.

The quantity $\underline{D} = \underline{\bar{\bar{X}}} - \underline{\bar{\bar{W}}}$, called the bias of the population mean or simply M-bias, can be estimated from segment data (see Appendix B). As shown in Appendix A, this permits the estimation of estimator bias,

$$\text{Bias}\ (\hat{\bar{\bar{P}}}_{reg}) \doteq \underline{\hat{B}}'\underline{\hat{D}}$$

and estimator variance,

$$V(\hat{\bar{\bar{P}}}_{reg}) \doteq [V(\hat{\bar{\bar{P}}}_{DE})]\ (1 - R^2)\ (n-1)/(n-p-1) + \underline{D}'\ [V(\underline{\hat{B}})]\ \underline{D}$$

$$\doteq [V(\hat{\bar{\bar{P}}}_{DE})]\ [1 + n\underline{D}'(\underline{z}\ \underline{z}^t)^{-1}\ \underline{D}]\ (n-1)/(n-p-1)$$

where

$V(\hat{\bar{\bar{P}}}_{DE})$ = estimated variance of $\hat{\bar{\bar{P}}}_{DE}$

R^2 = coefficient of determination for the regression of $\hat{y}_i$ on $\underline{x}_i^*$

p = dimensionality of $\underline{x}_i^*$

and $\underline{z} = (\underline{x}_1^* - \underline{\bar{x}}^* \mid \underline{x}_2^* - \underline{\bar{x}}^* \mid \ . \ . \ . \ \mid \underline{x}_n^* - \underline{\bar{x}}^*)$

Note that the estimator variance is an increasing function of both the inability of LANDSAT data to predict yield, as measured by $1 - R^2$, and of the LANDSAT bias of the mean. The influence of the latter on estimator variance is indicated by

VIF = variance inflation factor

$$= \frac{\text{estimator variance with estimated D}}{\text{estimator variance with } \underline{D} = 0}$$

and on the mean square error,

$$\text{MSE}\ (\hat{\bar{\bar{P}}}_{reg}) = V(\hat{\bar{\bar{P}}}_{reg}) + [\text{Bias}(\hat{\bar{\bar{P}}}_{reg})]^2$$

by

$$\text{MSEIF} = \text{mean-square-error inflation factor}$$

$$= \frac{\text{estimator MSE with estimated D}}{\text{estimator MSE with } \underline{D} = 0}$$

The improvement, if any, of the biased regression estimator over the direct expansion estimator is measured by the mean-square-error relative efficiency:

$$\text{RE} = V(\hat{\bar{\bar{P}}}_{DE}) / \text{ MSE } (\hat{\bar{\bar{P}}}_{reg}).$$

The biased regression estimator is only one of several different regression estimators applicable to LANDSAT-based yield estimation. Alternative estimators are discussed in Appendix D.

ANALYSIS

Sub-state Analysis

Due to various acquisition dates of the Iowa LANDSAT imagery, associated cloud-cover problems, and the different times at which ESS received LANDSAT data, Iowa was partitioned into 10 separate areas, called analysis districts, for the Iowa crop-area project (Figure 2). Analysis district yield estimation with the biased-regression estimator using non-haze-corrected data was evaluated in four of the analysis districts -- 1, 2A, 3C, and 4. Summary statistics for these four analysis districts are given in Table 5.

Table 5 - Summary statistics for fields with both yield and LANDSAT data

Crop	analysis district	number of counties	number of fields with both data	avg. yield[1/] (Bu/A)	std. dev. of avg. yield (Bu/A)	C.V.(%)
Corn	1	20	49	136.9	3.16	2.3
	2A	12	25	145.5	3.40	2.3
	3C	13	19	126.6	7.67	6.1
	4	19	39	134.4	4.91	3.7
Soybeans	1	20	25	45.2	2.17	4.8
	2A	12	33	42.9	1.88	4.4
	3C	13	12	44.0	2.85	6.5
	4	19	16	39.9	2.11	5.3

[1/] Mean per field of field-level gross yields estimated from 8 plots per field

Eleven different LANDSAT variables (four MSS bands plus the seven indices in Table 1) and two different methods of computing LANDSAT-variable field means were evaluated. The two methods of computing LANDSAT field means were by using only *interior pixels* -- that is, pixels completely interior to the field -- or *all pixels* having pixel center points inside the field.

Analysis district estimates of single-variable M-biases are summarized in Table 6 for corn and in Table 7 for soybeans.

Table 6 - Analysis district estimates of LANDSAT relative biases of the mean for corn

analysis district	type of pixel	Estimated Relative Biases (%)		
		Distribution[1]	Median	
		+6 +4 +2 0 -2 -4 -6 -8 -10 -12 -14 -16	absolute value	sampling std. dev.
1	all	S G | T R 5 6 47P | D V	1.4	0.5
	interior	| RT G | 7D 56S | 4PV	0.2	0.5
2A	all	S | 6G | R 5 47 | PTD V	1.1	0.6
	interior	S | 5P | 47 | V RG6 | DT	0.1	0.7
3C	all	G SPD | T 6 5 74 | RV	1.2	0.9
	interior	| T G R | 65P V | 574 D	0.7	1.1
4	all	5 4 S | 6 T7 G P R DV	3.2	0.6
	interior	| S T G V 5 4 | 6 7 P R D	2.5	0.6

[1] 4=MSS4, 5=MSS5, 6=MSS6, 7=MSS7,
P=PVI, S=SUM, D=DIFF, V=VI, T=TVI,
R=RATIO, G-GVI

Table 7 - Analysis district estimates of LANDSAT relative biases of the mean for soybeans

		Estimated Relative Biases (%)		
		Distribution[1]	Median	
analysis district	type of pixel	+6 +4 +2 0 -2 -4 -6 -8 -10 -12 -14 -16	absolute value	sampling std. dev.
1	all	¦ G V 5 4 ¦ S T67 P R D	3.4	1.1
	interior	RD ¦ SP ¦ 547G ¦ T V	0.5	1.5
2A	all	¦ G 5 4 ¦ ST 76 P RV D	3.8	0.8
	interior	¦ 4VD ¦ SPG 5R ¦ T76	0.2	0.9
3C	all	¦ T P G 54 ¦ S76 VDR	1.2	1.5
	interior	G PR T ¦ D V 76S 45 ¦	1.8	1.6
4	all	¦ 5 V ¦ T 4 RS 76 PG D	8.7	1.0
	interior	¦ 5 ¦ T 4 R V S 76 P G D	6.1	1.4

[1] 4=MSS4, 5=MSS5, 6=MSS6, 7=MSS7,
P=PVI, S=SUM, D=DIFF, V=VI, T=TVI,
R=RATIO, G=GVI

Items of note in these tables are that:

1) the rankings of LANDSAT variables according to estimated relative absolute M-bias (ERAMB) differ considerably under different conditions of analysis district, crop, and field-mean calculation;

2) TVI7 is the only variable having an ERAMB of less than 2% across all levels of all conditions;

3) if analysis district 4 is ignored, then TVI7 and SUM are the only variables with less than 2% ERAMB across all remaining levels of all conditions and;

4) for the interior-pixel case, TVI7, SUM, AND MSS4 are the only variables with less than 1% ERAMB for both corn and soybeans across all analysis districts other than analysis district 4.

The algebraic definitions of SUM and TVI partially explain the low ERAMB performances for these variables. Note in Tables 6 and 7 that MSS5 and MSS7 often have large relative M-biases of opposite sign. Hence, SUM=MSS5+MSS7 has a small M-bias. Note also that nearly always

ERAMB(SUM)   ERAMB(VI)   ERAMB(DIFF).

This is expected, at least for the upper bound, because SUM and DIFF nearly always have M-biases of like sign and VI=DIFF/SUM. Finally, note that in all cases

ERAMB(TVI)   ERAMB(VI).

This also is expected because the TVI square-root transformation has the limiting property that

ERAMB(TVI) = ERAMB(VI)/2 as ERAMB(VI) approaches zero.

One of the summary measures displayed in Tables 6 and 7 is the median of ERAMB over variables for each analysis district and field-mean computation method. Figure 3 compares these median ERAMB's with the acreage-study percentages of correct classification. Items of note in Figure 3 are that ERAMB, at least as measured by the median of individual-variable ERAMB's:

1) has a decreasing relationship with percentage of correct classification;

2) is less for interior pixels than for all field pixels in all cases except one (analysis district 3C, soybeans); and

3) is extremely small (less than 0.4%) for interior pixels when the proportion of correct classification exceeds 75%.

Estimator properties--that is, variance, bias, VIF, MSEIF, and relative efficiency were estimated for a large number of different biased-regression estimators. The primary estimator variable was in all cases the field-level yield estimated from 8 plots per field. The auxiliary variables were all combinations of one, two, or three of the Table 1 indices and/or LANDSAT bands. The estimators were then ranked by relative efficiency. Tables 8 and 9 list the "winners". Appendix E lists winners plus runners up.

For the estimators in Tables 8 and 9, the relative efficiency when field-level yield is estimated from two plots per field was also calculated. Appendix F describes how the difference in results between two and eight plots per field can be used to estimate relative efficiency for the case of known field-level yield. Relative efficiencies for two plots, eight plots, and known field yields are listed in Tables 10 and 11.

Table 8 - Regression-like estimator "winners" for corn with field yields estimated from 8 plots per field.

analysis district	LANDSAT variables[1/]	R^2	estimated relative bias (%)	VIF	C.V.(%)	MSEIF	estimated relative root-MSE(%)	RE[2/]
1	-(S,R,G)	.19	-0.04	1.003	2.2	1.003	2.2	1.15
	-R	.11	-0.18	1.001	2.2	1.007	2.2	1.1±.02
	-(4,7,R)	.20	0.67	1.026	2.2	1.127	2.3	1.1
2A				no significant correlations				
3C	-(6,7)	.41	-0.5	1.005	4.9	1.015	5.0	1.6±(.35-.39)
	-(R,G)	.40	-0.5	1.002	5.0	1.011	5.0	1.6±.21
4	-5	.11	-1.4	1.034	3.5	1.185	3.7	0.9±.09
	-D	.11	-1.7	1.052	3.5	1.283	3.9	0.9±.08

[1/] See footnote, Table 6. Symbols: - and + indicate interior and field pixels, respectively.

[2/] The relative efficiency entry is of the form RE ± SD(RE), where SD(RE) is the one standard deviation uncertainty of the estimated relative efficiency due to sampling error in estimating LANDSAT biases of the mean. A range for SD(RE) corresponds to bounding from above and below the covariances among bias components in lieu of (properly) estimating the needed covariances.

Table 9 - Regression-like estimator "winners" for soybeans with field yields estimated from 8 plots per field.

analysis district	LANDSAT variables[1]	R^2	estimated relative bias (%)	VIF	C.V.(%)	MSEIF	estimated relative root-MSE(%)	RE[2]
1	-(4,S)	.66	0.007	1.012	2.9	1.012	2.9	2.7±.04
	-(4,7)	.65	-0.09	1.012	2.9	1.013	3.0	2.7±.12
	-D	.64	0.07	1.00002	3.0	1.0006	3.0	2.7±.12
2A	-(6,G)	.52	0.7	1.019	3.2	1.073	3.2	1.9±(1.4-1.7)
	-V	.48	-0.1	1.0001	3.2	1.001	3.2	1.9±.05
	-4	.48	0.2	1.0002	3.3	1.004	3.3	1.9±.05
	-T	.48	-0.2	1.0002	3.3	1.0005	3.3	1.9±.05
3C	+4	.42	0.6	1.002	5.3	1.015	5.3	1.5±.11
	+(S,G)	.40	0.6	1.010	5.4	1.023	5.4	1.5±.21
	+(S,T)	.38	-0.4	1.003	5.4	1.009	5.4	1.5±.10
	+(7,V)	.38	0.07	1.005	5.4	1.005	5.4	1.5±.02
	+(P,V)	.38	0.09	1.005	5.5	1.005	5.5	1.5±.04
4			no significant correlations					

[1] See Footnote, Table 8.

[2] See Footnote, Table 8.

Table 10 - Comparison of 2-plots, 8-plots, and predicted known-field-yield results for corn

analysis district ($CV(\hat{P}_{DE})$)	LANDSAT1/ variables	R^2 2/			estimated3/ relative root-MSE (%)			RE3/			f_{field} 4/		max R^2 5/	
		(2)	(8)	(k)	(2)	(8)	(k)	(2)	(8)	(k)	(2)	(8)	(2)	(8)
1 (2.3%)	-(S,R,G)	.15	.19	.26	2.3	2.2	2.1	1.1	1.15	1.3	.13	.09	.85	.93
	-R	.08	.11	.15	2.2	2.2	2.1	1.1	1.1	1.15	.09	.04	.93	.96
	-(4,7,R)	.08	.12	.15	2.3	2.3	2.3	1.1	1.1	1.1	.08	.03	.93	.97
3C (6.1%)	-(6,7)	.40	.40	.40	5.0	5.0	5.0	1.6	1.6	1.6	0	0	1.00	1.00
	-(R,G)	.40	.40	.40	5.0	5.0	5.0	1.6	1.6	1.6	0	0	1.00	1.00
4 (3.7%)	-5	.07	.11	.14	3.8	3.7	3.6	0.9	0.9	0.9	.08	.03	.93	.97
	-D	.06	.11	.15	4.0	3.9	3.8	0.9	0.9	0.9	.10	.04	.91	.96

1/ See footnote, Table 8.

2/ (2) = 2 plots/field, (8) = 8 plots/field
(k) = prediction for known field yield

3/ Calculated with MSEIF for 8 plots/field

4/ f_{field} = (estimated) fraction of $MSE(\hat{P}_{reg})$ attributable to estimation of field-level yield

5/ max R^2 = (estimated) maximum R^2; i.e. R^2 when there exists a perfect linear relationship between x_i and known y_i

Table 11 - Comparison of 2-plot, 8-plot, and predicted known-field-yield results for soybeans

analysis district ($CV(\hat{P}_{DE})$)	LANDSAT1/ variables	$R^2$2/			estimated3/ relative root-MSE (%)			RE3/			f_{field}4/		max $R^2$5/	
		(2)	(8)	(k)	(2)	(8)	(k)	(2)	(8)	(k)	(2)	(8)	(2)	(8)
1 (4.8%)	-(4,5)	.41	.66	.76	3.8	2.9	2.4	1.6	2.7	3.8	.59	.29	.65	.90
	-(4,7)	.39	.65	.76	4.0	3.0	2.5	1.5	2.7	3.9	.61	.31	.63	.89
	D	.38	.64	.73	3.9	3.0	2.6	1.6	2.7	3.6	.56	.25	.65	.91
2A (4.4%)	-(6,G)	.32	.52	.73	3.8	3.2	2.4	1.3	1.9	3.4	.60	.44	.51	.79
	-V	.27	.47	.67	3.8	3.2	2.5	1.4	1.9	3.1	.55	.38	.60	.80
	-4	.33	.46	.62	3.7	3.3	2.8	1.5	1.9	2.7	.43	.23	.71	.84
	-T	.25	.46	.67	2.6	3.3	2.6	1.2	1.9	3.1	.56	.39	.58	.79
3C (6.5%)	+4	.31	.42	.53	5.8	5.3	4.8	1.3	1.5	1.9	.32	.19	.78	.89
	+(S,G)	.22	.46	.64	6.5	5.4	4.4	1.0	1.5	2.3	.54	.33	.58	.82
	+(S,T)	.25	.45	.60	6.3	5.4	4.6	1.1	1.5	2.1	.47	.27	.65	.85
	+(7,V)	.24	.45	.60	6.3	5.4	4.6	1.1	1.5	2.1	.47	.27	.66	.85
	+(P,V)	.24	.44	.60	6.4	5.5	4.6	1.1	1.5	2.1	.47	.29	.66	.84

1/, 2/, 3/, 4/, 5/ See footnotes, Table 10.

State-level Analysis

There are at least two ways to obtain an estimate of the mean yield for the state with the biased-regression estimator. The first is to simply apply the biased-regression estimator to all fields with both yield and LANDSAT data. An alternative method would be a union of the separate analysis district regression estimates with RE greater than one and the direct expansion estimates for those analysis districts with RE less than one or cloud cover. The first method is presented here while the second is deferred for later consideration except to say that results appear to be similar.

There are many alternative levels at which to regress y on x. While the most natural is probably the field level, aggregation of yield and LANDSAT provides inference to larger areas. As discussed earlier, we have three field-level yield estimates -- OY estimate (two units), research estimate (eight units) and farmer reported yield. LANDSAT means can be calculated on any aggregate level using two or more pixels. There are two main factors which determine whether the biased-regression estimator is an improvement over the direct expansion estimator -- the magnitude of R^2 and the size of the bias, $\underline{D}$.

We would most like to gain in relative efficiency by using the OY estimate since this does not require yield data collection above what is normally obtained in the regular operating program. Also, use of the OY estimate avoids questions related to the credibility of the farmer reported yield. However, the OY estimate is generally imprecise at the field level. If we regress OY on the LANDSAT variable(s) at the field level, the imprecision of the yield estimate might destroy any correlation which could have been measured had we known the "true" yield. Of course, if the LANDSAT data do not relate well to the radiance directly above the crop canopy, this also effects the R^2 measured by the regression. Since it is not necessary to regress y on x at the field level, several fields can be aggregated so that more precise yield estimates can be made. One consideration is to retain a sufficient number of aggregated data points to avoid possible bias from a small sample size. To obtain more precise yield estimates at an aggregate level, fields were blocked together geographically by grouping adjacent counties within analysis districts. We tried to end up with approximately 25 aggregated data points. For soybeans, each block contained a group of counties with roughly 4 OY samples. Since we had no control over the distribution of samples within counties, some blocks had as few as two samples or as many as eight. The total number of soybean blocks was 24. In the case of corn, where we had more samples to work with, blocks contained roughly 5 samples each but varied from one to seven. The block with one sample occurred because this was the only sample in the entire analysis district. The total number of corn blocks was 29.

The basis for area blocking is that the true yield is thought to be somewhat less variable for small areas of land than over large areas which may contain widely differing weather conditions, soils and other factors related to yield. If area blocks do contain fairly homogeneous yields, four or five OY samples might provide a more precise block level estimate than individual field-level estimates. In the process of trying to keep roughly the same number of samples in each block, the number of counties per block ranges from one to seven. Unfortunately, the larger blocks start to lose any advantage of being small, homogeneous areas.

There are a multitude of different biased-regression estimators which can be calculated to find the one which maximizes the relative efficiency. We have eleven LANDSAT variables -- four channel means and seven indices. Each LANDSAT variable has been calculated from all pixels classified to the field or from just those pixels which are completely interior to the field boundaries. Each LANDSAT variable has also been calculated with uncorrected and corrected pixels (refer to Appendix C for the haze correction algorithm). This gives four sets of eleven LANDSAT variables. Because the LANDSAT variables are highly correlated, we likely will obtain the best results with either one or two regressor variables. The set of dependent variables consists of our three yield estimates. The regressions can be run with field-level data or the area-blocked data.

To reduce the number of regression models to be considered, we recall that R^2 and $\underline{D}$ are the two principle factors in maximizing the relative efficiency. An indicator combining R^2 and the associated bias was used to select models with the best chance of success. The Statistical Analysis System (2) has a procedure called RSQUARE which calculates the R^2 values for all the possible combinations of dependent and regressor variables it is given. If only R^2 values are needed, it is much more efficient than obtaining all the other associated regression parameters. We ran the RSQUARE procedure to determine which one and two-variable regressions had the highest R^2 values for all previously mentioned yield and LANDSAT variables with and without blocking. Bias was estimated for each LANDSAT variable from the JES segment data (details in Appendix B). The relative bias was calculated by dividing each element in $\underline{\hat{D}}$ by the corresponding $\underline{\hat{\bar{\bar{X}}}}$. The index used to select the regressions with the greatest potential for maximizing the relative efficiency is the following:

$$I = \frac{100}{R^2} \sum_{i=1}^{p} \frac{(\hat{D}_i / \hat{\bar{\bar{X}}}_i)}{2}$$

where i = 1, ..., p regressor variables

Since we are searching for regressions with high R^2 and low relative bias, the lowest values of I were selected and the other factors needed to calculate the relative efficiencies were estimated. When blocking was used, the regressions were weighted to account for differences in the number of observations in the blocks.

Table 12 shows the relative bias in percent for each LANDSAT variable. It can be seen that the biases span a wide range all the way from .06% to 10.12%. The LANDSAT variables from interior pixels generally, but not always, had smaller biases than the corresponding variables calculated from all the pixels. Also, with the exception of the interior soybean variables, the corrected LANDSAT variables generally had larger bias than the corresponding uncorrected variables. The signs associated with the biases have been included in Table 12. This is important because in the two-variable regressions the biases might offset one another so that even though they are comparatively large individually, the relative efficiency might not be adversely affected.

Tables 13 and 14 summarize the relative efficiencies greater than one for each yield estimate. These are the highest RE's which were found by using the previously mentioned index but others may have been found if all possible combinations had been considered. However, it is unlikely that

Table 12 - State-level relative biases expressed in percent

LANDSAT Variable[1/]	Uncorrected		Corrected	
	All	Interior	All	Interior
	Soybeans			
4	+1.80	+.58	+.64	-.25
5	+3.03	+.22	+1.47	-.85
6	-.84	+1.65	-1.15	+1.09
7	-1.42	+1.43	-1.64	+.96
P	-2.26	+1.64	-2.43	+1.39
S	-.19	+1.11	-.63	+.38
D	-4.17	+2.11	-4.59	+2.51
V	-3.68	+1.24	-2.96	+3.02
T	-.90	+.27	-.62	+.64
R	-4.10	+1.27	-2.30	+2.19
G	-2.11	+1.92	-2.25	+1.66
	Corn			
4	+1.88	+.25	+2.22	+1.18
5	+3.43	+.43	+5.32	+3.19
6	-.34	-1.54	+.73	-.37
7	-.79	-1.27	-.06	-.51
P	-1.87	-1.69	-1.85	-1.72
S	+.60	-.72	+1.96	+.86
D	-4.82	-2.81	-8.17	-5.86
V	-5.30	-2.15	-10.12	-6.90
T	-1.18	-.51	-1.77	-1.22
R	-2.73	-.82	-4.97	-3.73
G	-1.88	-2.19	-1.41	-1.88

[1/] See footnote Table 6 for variable definitions.

Table 13 - Relative efficiencies for soybean regression estimators

Blocking	Independent Variable(s)1/	R^2	Relative Efficiency
	Dependent Variable - OY Estimate		
No	IC4*, ICS	.0565*	1.0323
No	IT	.0391	1.0241
No	S	.0334	1.0221
No	I5, IT*	.0582*	1.0134
Yes	IT, IR	.1217	1.0124
No	ICT*, ICR	.0530*	1.0100
No	ICT*	.0474*	1.0059
	Dependent Variable - Research Yield Estimate		
No	IT*	.1324*	1.1216
No	IC4*, ICS*	.1526*	1.1166
Yes	I4, IS	.1885*	1.1096
No	S*	.1080*	1.1041
No	I5*, IT*	.1904*	1.0911
Yes	IC4, ICS	.1289	1.0791
Yes	IT	.1172	1.0629
Yes	IC4	.0831	1.0384
Yes	IC4, ICT	.1167	1.0285
No	S*, T*	.1880*	1.0211
No	I7*	.1810*	1.0146
Yes	I5	.0541	1.0110
	Dependent Variable - Farmer Yield Estimate		
Yes	IC4*	.2025*	1.1785
Yes	IC4, IC5	.2104*	1.1582
No	IC4*,IC5	.1740*	1.1574
No	IC4*	.1489*	1.1488
No	IC4*, IC5	.1508*	1.1411
Yes	C4*	.2047*	1.0258

1/ See footnote Table 6 for variable definitions. In addition, a prefix "I" indicates interior pixels only and a prefix "C" indicates haze correction.

* Statistical significance at the 5% level.

Table 14 - Relative efficiencies for corn regression estimators

Blocking	Independent Variable(s)[1/]	R^2	Relative Efficiency
	Dependent Variable - OY Estimate		
Yes	6	.0590	1.0230
	Dependent Variable - Research Yield Estimate		
Yes	I4*, IS	.1891*	1.0633
No	I4*	.0526*	1.0443
Yes	I4, I7	.1771*	1.0433
Yes	IT	.1258	1.0381
Yes	I4	.0716	1.0366
No	I4, IT	.0746*	1.0074
No	C7	.0132	1.0060
	Dependent Variatle - Farmer Yield Estimate		
Yes	6	.1129	1.0838
Yes	I4, I6*	.1891*	1.0170

[1/] See footnote Table 13.

* Statistical significance at 5% level.

there are any greater relative efficiencies than the highest of those presented. The column headed "Blocking" indicates whether or not the data were blocked by geographical areas. The independent variables are in some cases prefixed with an "I", a "C" or both. The "I" indicates interior pixels only were used and the "C" indicates that the XSTAR haze correction was applied. The presence of an asterisk (*) indicates that either the coefficient of the given variable or the R^2 value is significantly different from zero at the .05 level.

Blocking the data was of some value for corn but not for soybeans. Not surprisingly, use of interior pixels generally gave the highest relative efficiencies. The use of corrected LANDSAT data was of some value for soybeans but not for corn. As would be expected, the highest RE for the research yield estimate was greater than the highest RE for the OY estimate.

Interestingly, the highest relative efficiency for each crop occurred with the farmer reported yield. One would expect LANDSAT data to be more strongly related to gross yield than net yield since the amount of harvest loss may be completely unrelated to the plant characteristics indicated by solar reflectance. If harvest loss was fairly consistent from field to field then whether net or gross yield was used would not really matter. However, our somewhat limited harvest loss data were roughly two to three times more variable than the corresponding OY data. The comparatively high RE for farmer yield is of limited value because the accuracy of the direct farmer yield estimate cannot be measured.

It should also be pointed out from Tables 13 and 14 that, with two exceptions, at most only one coefficient in the two-variable regressions was significantly different from zero. This is largely due to strong linear dependencies among the LANDSAT variables which tend to inflate the variance of the coefficient estimates. This is why we did not investigate multiple regressions with more than two variables.

Compared to relative efficiencies obtained in LANDSAT-acreage studies, our best are low. The question which needs to be asked then is how high of a RE is needed to obtain some benefit? Clearly, if LANDSAT data is being obtained solely for improving yield estimates, the cost could not be justified with the results we have demonstrated. However, if the LANDSAT data are available for yield estimation at no cost because it has already been obtained for improving acreage estimates then any gain in yield estimation efficiency is of use. Actually, we may not really need to improve the precision of our state level yield estimates in major producing states. If the direct expansion estimate has only 2% error then improvement might be considered a waste of resources.

Rather than talking about improving the direct expansion estimate, we would like to demonstrate how much might be saved in reduced sample size by using the regression estimator and keeping the precision the same. Table 15 shows the sample sizes needed to obtain a 2% CV at the state level with various relative efficiencies. The RE of 1.00 shows the needed sample size when only the direct expansion estimator is used.

Table 15 - Comparison of sample sizes needed for a 2% state level coefficient of variation with various relative efficiencies and estimators

Relative		Soybean Sample		Corn Sample	
Efficiency	Estimator:	OY	Research	OY	Research
1.00		192	107	131	93
1.01		190	105	130	92
1.02		188	105	129	91
1.05		183	102	125	89
1.10		174	97	119	85
1.15		167	93	114	81
1.20		160	89	109	78
1.25		153	85	105	74
1.30		148	82	101	72
1.35		142	79	97	69
1.40		137	76	94	66
1.45		132	74	90	64
1.50		128	71	87	62

Table 16 - Estimated cost savings using highest PE's from Tables 13 and 14

Total cost per sample (2 plots) for entire survey	$135.00
Total cost per sample (8 plots) for entire survey	215.00
Total cost per sample (2 plots) for final pre-harvest visit only	37.50
Total cost per sample (8 plots) for final pre-harvest visit only	117.50

	Soybeans[1]				Corn[1]			
	Method I		Method II		Method I		Method II	
Estimator	Cost	%	Cost	%	Cost	%	Cost	%
Direct Expansion	$25,900	100	$25,900	100	$17,700	100	$17,700	100
OY (2 plots)	25,100	97	25,700	99	17,300	98	17,550	99
Research (8 plots)	20,400	79	29,900	115	18,700	106	23,000	130

[1] Method I assumes that samples in excess of indicated sample size are eliminated for entire survey period. A corresponding number of post-harvest interviews and harvest-loss samples are eliminated. Method II assumes that only the final pre-harvest visit is eliminated for samples in excess of indicated sample size. The number of post-harvest interviews and harvest-loss samples remains unchanged.

The sample sizes were calculated from the yield-LANDSAT data set which was used for the preceding state-level analysis. It should be pointed out that these sample sizes pertain to gross yield estimates. However, the required sample size for an equally precise net yield estimate would be roughly the same. This is so, at least for this data set, because the negative covariance observed between gross yield and harvest loss approximately offsets the harvest loss variance.

The actual initial sample sizes were 170 for soybeans and 240 for corn. Therefore, in Iowa in 1978 the soybean sample was not large enough to obtain a 2% CV while the corn sample was larger than it needed to be for a 2% CV. For sake of example, it can be inferred from Table 15 that the sample size could be reduced 23% with either the soybean or corn OY regression estimator if the relative efficiency was 1.3. If eight units were laid out in each field as was done to obtain the research data then a relative efficiency of 1.3 would allow a 57% reduction is sample size for soybeans and 45% for corn.

Table 16 gives some cost figures associated with the highest relative efficiencies from Tables 13 and 14 using two different methods. These cost figures are rough estimates based on 1978 Iowa data showing the length of time taken for various field enumerator activities and other cost data from the Data Collection Branch. The figures attempt to account for costs associated with field enumeration, travel to and from the sample fields, laboratory processing and editing. Costs related to supplies, forms, manuals, equipment and computer processing are not included. The figures are intended to be used for demonstration purposes only.

Two different methods were used to calculate the survey costs for the three estimators. The first method assumes that the sample size for the entire survey period can be reduced to the level indicated by the relative efficiency of the regression estimator. This means that less data would be collected to make early season yield forecasts. While the LANDSAT data possibly could be of some benefit for yield forecasting, we have only addressed the estimation phase when yield data are available just before harvest. It is probably unreasonable to assume that the yield forecasting program could be successfully run with a greatly reduced sample size. The second method assumes that just the final pre-harvest visit is eliminated for those samples in excess of the indicated sample size.

It is evident from Table 16 that the modest gains in relative efficiency demonstrated with the biased-regression estimator produce small savings. It may well be, however, that other alternative regression estimators proposed in Appendix D may work better.

REFERENCES

1. Ahern, F. J., D. G. Goodenough, S. C. Jain, V. R. Rao, and G. Rochon. 1977. "Use of Clear Lakes as Standard Reflectors for Atmospheric Measurements." Proceedings of Eleventh International Symposium on Remote Sensing of the Environment. I:731-755.

2. Barr, A. J., J. H. Goodnight, and J. P. Sall. 1979. "SAS User's Guide - 1979 Edition." SAS Institute Inc., Raleigh, NC.

3. Colwell, J. E., D. P. Rice, and R. P. Nalepka. 1977. "Wheat Yield Forecasts Using LANDSAT Data," Proceedings, Eleventh International Symposium on Remote Sensing of Environment.

4. Cook, P. W. 1980. "Satellite Provides Assist in Improving USDA Acreage Estimates". The Sunflower, 6(7). pp. 28-30.

5. Craig, M. E. and C. E. Miller. 1980. Area Estimates by LANDSAT: Arizona 1979. Economics and Statistics Service, U. S. Dept. of Agriculture, Washington, D. C.

6. Craig, M. E., R. S. Sigman, and M. Cardenas. 1978. Areas Estimates by LANDSAT: Kansas 1976 Winter Wheat. Economics and Statistics Service, U. S. Dept. of Agriculture, Washington, D. C.

7. Gleason, C. P., R. A. Starbuck, R. S. Sigman, G. A. Hanuschak, M. E. Craig, P. W. Cook, and R. D. Allen. 1977. *The Auxiliary Use of LANDSAT Data in Estimating Crop Acreages: Results of the 1975 Illinois Crop-Acreage Experiment*. Economics and Statistics Service, U. S. Dept. of Agriculture, Washington, D. C.

8. Hanuschak, G. A., R. S. Sigman, M. E. Craig, M. Ozga, R. Luebbe, P. W. Cook, D. Kleweno, and C. E. Miller. 1979. *Obtaining Timely Crop Area Estimates Using Ground-Gathered and LANDSAT Data*, Economics and Statistics Service, U. S. Dept. of Agriculture, Washington, D. C., Technical Bulletin No. 1609.

9. Kauth, F. J. and D. S. Thomas. 1976. "The Tasselled Cap - A Graphic Description of the Spectral-Temporal Development of Agricultural Crops as seen by LANDSAT." *Proceedings of 1976 Symposium on Machine Processing of Remotely-Sensed Data*, Purdue University, West Lafayette, Indiana.

10. Lambeck, P. F. 1977. *Signature Extension Preprocessing for LANDSAT MSS Data*, ERIM 122700-32-F. Environmental Research Institute of Michigan, Ann Arbor, Michigan.

11. Lambeck, P. F. 1977. *Revised Implementation of the XSTAR Haze Correction Algorithm and Associated Pre-Processing Steps for LANDSAT Data*. Environmental Research Institute of Michigan Memo Number IS-PFL-1916, November 1, 1977.

12. Larsen, G. A. 1979. Report on 1978 LANDSAT, Yield and Production Study as it Pertains to the Objective Yield Program. Economics and Statistics Service, U. S. Dept. of Agriculture, Washington, D. C. Unpublished manuscript.

13. Mergerson, J. W. Crop-Area Estimates Using Ground-Gathered and LANDSAT Data, a Multitemporal Approach: Missouri 1979, Economics and Statistics Service, U. S. Dept. of Agriculture, Washington, D. C. In press.

14. Richardson, A. J., D. E. Escobar, H. W. Cansman, and J. H. Everitt. 1980. "Comparison of LANDSAT-2 and Field Spectrometer Reflectance Signatures of South Texas Rangeland Plant Communities." 1980 Proceedings of Machine Processing of Remotely-Sensed Data Symposium, Purdue University, West Lafayette, Indiana. In press.

15. Sigman, R. S., C. P. Gleason, G. A. Hanuschak, and R. A. Starbuck. 1977. "Stratified Acreage Estimates in the Illinois Crop-Acreage Experiment." Proceedings of the 1977 Symposium on Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana.

16. Tucker, C. J. 1977. Use of Near Infrared/Red Radiance Ratios for Estimating Vegetation Biomass and Physiological Status. NASA X-Document, Goddard Space Flight Center, Greenbelt, Maryland.

17. Wigton, W. H. and H. F. Huddleston. 1978. "A Land Use Information System Based on Statistical Inference," *Proceedings, Twelfth International Symposium on Remote Sensing of Environment*, Manilla, Philippines.

18. Wiegand, C. L., A. J. Richardson, and E. T. Kanemasu. 1979. "Leaf Area Index Estimates for Wheat from LANDSAT and their Implications for Evapotranspiration and Crop Modeling." *Agronomy Journal*, 7(12), pp. 336-342.

Figure 1

LANDSAT Scene Locations

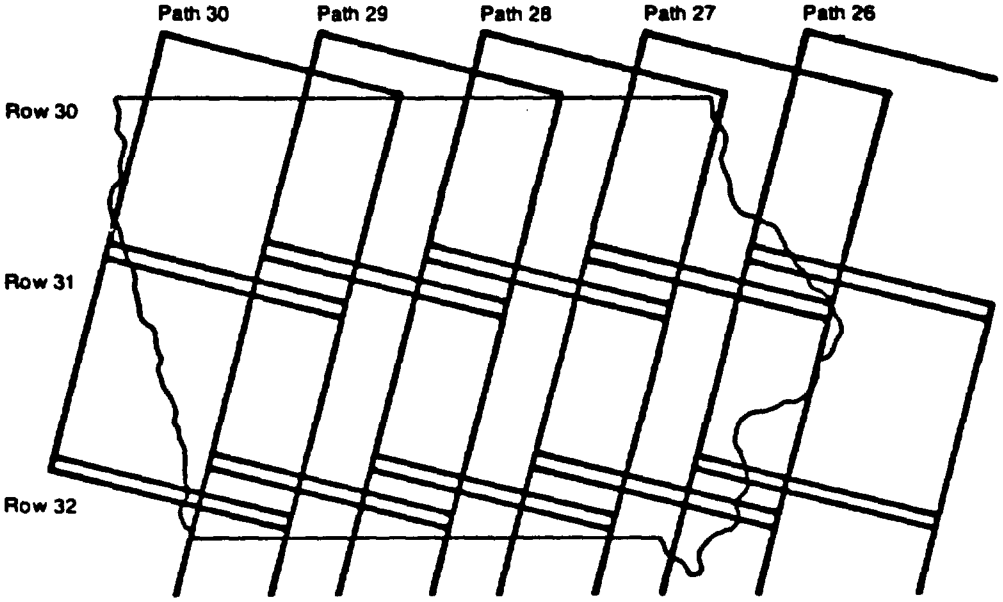


Figure 2

Iowa Analysis Districts

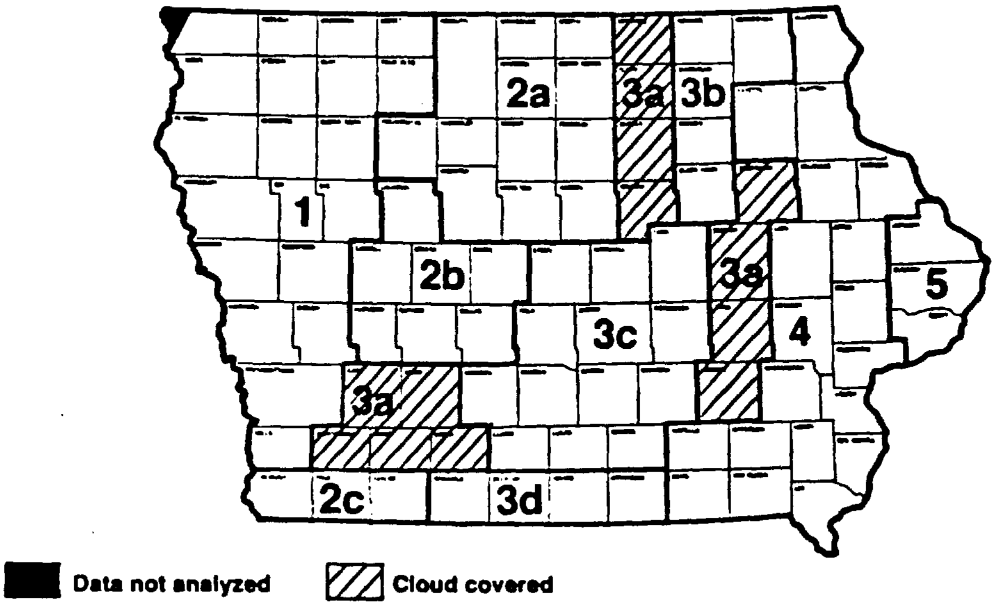
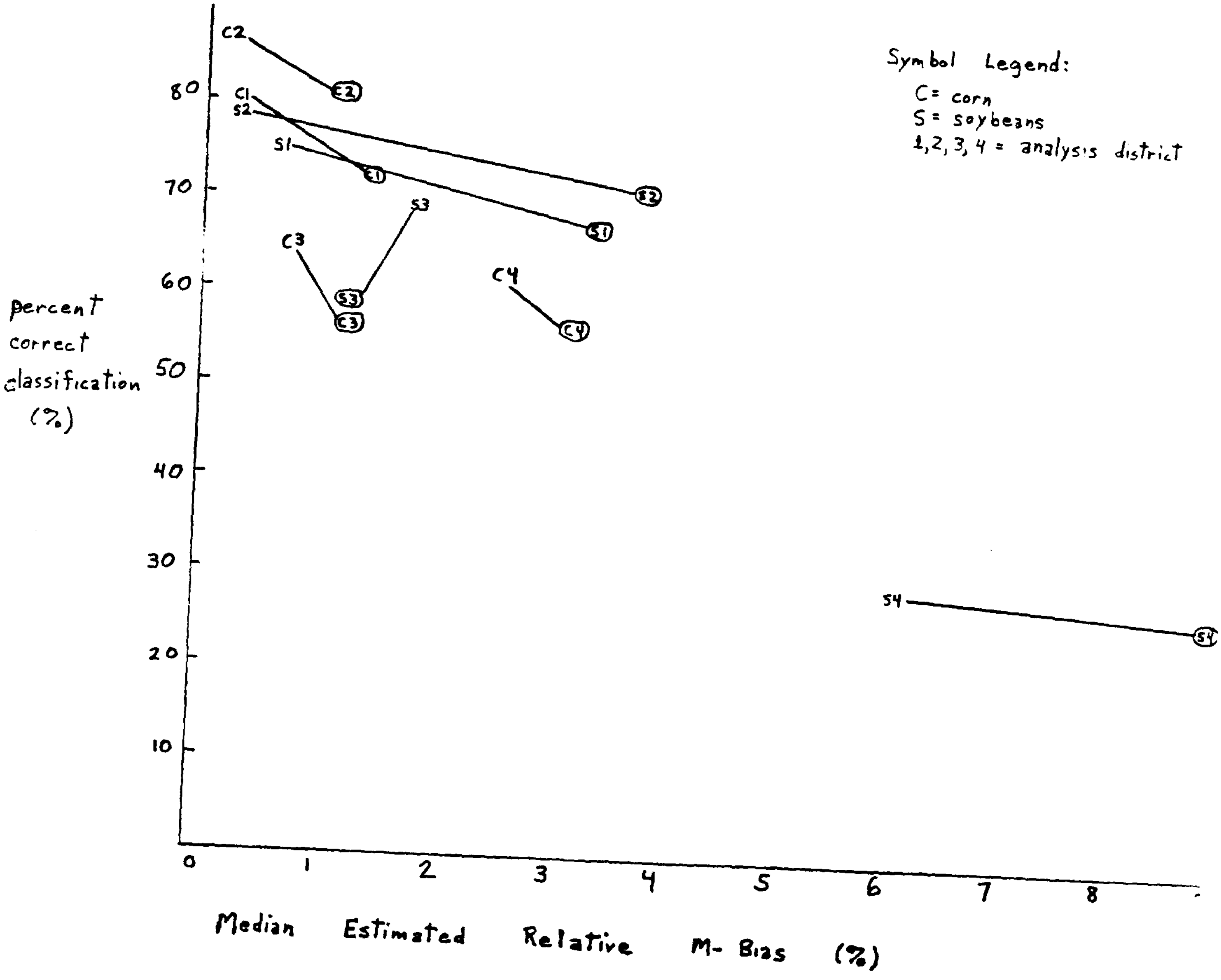


Figure 3. Percent-Correct Classification vs. M- Bias



Appendix A - PPS Estimators

Direct-Expansion Estimator

In objective yield procedures for corn and soybeans, the fields containing gross yield plots are selected by approximate PPS (probability proportional to size) sampling of fields from the JES sample of segments. Since an equal number of plots are laid out in each sampled field (with multiple selections of the same field being considered different sampled fields), the sample of plots is a self-weighted sample from the population of all permissible plot locations. For data which is available only at field level, however, such as in this case LANDSAT data, the PPS sampling probabilities must be considered in estimators involving field totals.

In such a situation, it is advantageous to think of the selection of objective yield fields as cluster sampling such that fields are sample units and acres inside fields are population elements. Population-level yield is then simply population-mean-per-element of production, which we denote $\bar{\bar{P}}$. A key result from the theory of PPS sampling is that the sample mean of

$$\frac{\text{sample-unit datum}}{\text{sample-unit size}}$$

is an unbiased estimator of the population mean per element. Thus, an unbiased direct-expansion estimator of $\bar{\bar{P}}$ is the sample mean of

$$\frac{\text{production of field i}}{\text{size of field i}} = \text{yield of field i.}$$

In other words, the PPS weighting required in the direct-expansion estimator of $\bar{\bar{P}}$ transforms the estimator into an unweighted sample mean of field yield.

As is the case in objective-yield procedures, PPS-selected sample units are often subsampled by independent simple random sampling in each sample unit. Sample-unit data is then not available but can be unbiasedly estimated from subsample data. A second major result from PPS sampling theory is that in this situation single-stage PPS sampling formulas for various unbiased estimators and associated estimated variances remain valid when estimated sample-unit data is used instead of known sample-unit data. The direct-expansion estimator for $\bar{\bar{P}}$ is an example of such an estimator.

Regression and Regression-like Estimators

The availability of field-level LANDSAT variables permits the calculation of the difference estimator $\hat{\bar{\bar{P}}}_{diff}(\underline{B}_o)$ which is the sample mean of

y_i^* = LANDSAT-adjusted field yield of sampled field

$= y_i + \underline{B}_o'(\underline{\bar{\bar{X}}} - \underline{x}_i^*)$

where

y_i = yield of sampled field i

$\underline{B}_o$ = an arbitrary vector of constants

$\underline{x}_i^*$ = vector of means per pixel for sampled field i of LANDSAT measurements and indices

and $\underline{\bar{\bar{X}}}$ = vector of population means per pixel of LANDSAT measurements and indices.

The variance of $\hat{\bar{\bar{P}}}_{diff}(\underline{B}_o)$ is minimized by

$$\underline{B}_o = \underline{B} = (\underline{Z}\ A\ \underline{Z})^{-1}\ \underline{Z}\ \underline{A}\ \underline{Y}$$

where

$$\underline{Z} = (\underline{x}_1^* - \underline{\bar{\bar{X}}}\ \mid\ \underline{x}_2^* - \underline{\bar{\bar{X}}}\ \mid\ \cdots\ \mid\ \underline{x}_N^* - \underline{\bar{\bar{X}}})$$

with N = number of fields in the population

$\underline{A}$ = diagonal matrix with i^th diagonal entry equal to the size of field i, i=1, 2, ..., N

and $\underline{Y}$ = population column vector of field yields.

When the asymptotically unbiased estimate of $\underline{B}$,

$$\hat{\underline{B}} = (\underline{z}\ \underline{z}^t)^{-1}\ \underline{z}\ \underline{y},$$

where z and y are sample analogues of Z and Y, respectively, is substituted into $\hat{\bar{\bar{P}}}_{diff}$, a regression estimator is obtained. As in the case of the direct-expansion estimator, when sample units are subsampled the y_i's are replaced with estimated $\hat{y}_i$'s determined from subsample data.

Related to the regression estimator is an estimator we call a biased regression estimator. It has the same form as the regression estimator except that $\underline{\bar{\bar{X}}}$ is not known and is thus replaced by a proxy value, denoted $\underline{\bar{\bar{W}}}$.

Appendix B - Estimation of $D = \bar{\bar{X}} - \bar{\bar{W}}$

$\underline{D} = \bar{\bar{\underline{X}}} - \bar{\bar{\underline{W}}}$ is the difference of two population-means-per-elements corresponding to two different populations. The population for $\bar{\bar{\underline{X}}}$ is

U_1 = (pixels associated with fields planted to the crop of interest)

or, alternatively,

U_2 = (field-interior pixels associated with fields planted to the crop of interest),

U' = (pixels classified to the crop of interest).

$\underline{D}$ can be estimated from JES segment data, plus corresponding raw and classified LANDSAT data, by estimating $\bar{\bar{\underline{Y}}}$ and $\bar{\bar{\underline{W}}}$ separately.

$\bar{\bar{\underline{X}}}$ is estimated as follows:

Let

$\underline{v}_{hsj}$ = LANDSAT variable vector for j^{th} pixel in segment s of stratum h,

I_{hsj} = an indicator which is 1 if the pixel for $\underline{v}_{hsj}$ is in U_1 (U_2) and is 0 if not,

$\underline{x}_{hs} = \sum_j I_{hsj}\underline{v}_{hsj}$,

$m_{hs} = \sum_j I_{hsj}$,

and E_h = expansion factor for stratum h.

Then

$$\hat{\bar{\bar{\underline{Y}}}} = (\sum_h E_h \sum_s \underline{x}_{hs})/(\sum_h E_h \sum_s m_{hs}).$$

In a similar fashion, $\underline{W}$ is estimated as follows:

Let

I'_{hsj} = an indicator which is 1 if the pixel for $\underline{v}_{hsj}$ is in U' and is 0 if not,

$$\underline{w}_{hs} = \sum_{j} I'_{hsj}\underline{v}_{hsj}$$

and

$$m'_{hs} = \sum_{j} I'_{hsj}.$$

Then

$$\underline{\hat{\bar{\bar{W}}}} = (\sum_{h} E_h \sum_{s} w_{hs})/(\sum_{h} E_h \sum_{s} m'_{hs}).$$

Thus

$$\underline{\hat{D}} = \underline{\hat{\bar{\bar{X}}}} - \underline{\hat{\bar{\bar{W}}}}.$$

A rather lengthy expression for the approximate variance of $\underline{\hat{D}}$ can be obtained without difficulty by using the approximation

$$\begin{aligned}\mathrm{Cov}(a/b,\ c/d) = {} & (EbEd)^{-1}\mathrm{Cov}(a,c) - (EbE^2d)^{-1}Ec\,\mathrm{Cov}(a,d) \\ & - (EdE^2b)^{-1}Ea\,\mathrm{Cov}(c,b) + (E^2bE^2d)^{-1}EaEc\,\mathrm{Cov}(b,d)\end{aligned}$$

Appendix C - Haze Correction

This is a general review of the factors that make the reflectance measured by LANDSAT different from the reflectance at the earth's surface. Much of the information to follow is based on a discussion of atmospheric radiative transfer theory by Richardson, et al. (14). Specific equations in the Richardson paper were taken from previous work by various authors. Our purpose here is not to present the actual theory so the reader is referred to the reference section of the Richardson paper for sources of the theory.

The radiance detected by LANDSAT is a function of reflectance at the earth's surface, total incident solar irradiance, atmospheric transmittance and path radiance. The first three factors mentioned are multiplicative while the fourth is additive. The reflectance at the earth's surface is, of course, what we would like to obtain since it has the most direct relationship to the plant canopy (or whatever else is on the earth's surface).

Simplistically speaking, the total incident solar irradiance refers to the amount of the electromagnetic radiation from the sun in the particular wavelengths of interest that actually reaches the earth's surface. The atmospheric transmittance affects both incident solar irradiance (sun to surface) and reflected solar radiance (surface to satellite). The path radiance arises from atmospheric scattering and absorption of the incident irradiance and reflected radiance.

The total incident solar irradiance is composed of direct and diffuse components. The direct component is a function of solar irradiance at the top of the atmosphere. This varies according to the earth-sun distance, atmospheric transmittance between the sun and the earth, and the solar zenith angle. The diffuse component is the part of the total incident solar irradiance which does not directly penetrate the atmosphere but does reach the earth by bouncing off of various atmospheric scatterers. The portion of incident irradiance which does not reach the earth is included in the path radiance.

The atmospheric transmittance is a function of the optical thickness of the atmosphere and the zenith angle -- sun angle if irradiance is being considered or sensor angle if radiance. Since the multispectral scanner on LANDSAT moves laterally over an arc of about 11.6°, the sensor or view angle varies from -5.8° to +5.8° of nadir. In many applications, the sensor zenith angle is assumed to be zero. The optical thickness of the atmosphere is a measure of scattering and absorption due to gaseous molecules, aerosol particulates and water molecules. Water absorption is generally assumed to be negligible in LANDSAT bands 4, 5 and 6 but not in band 7. Scattering is more severe for the shorter wavelengths.

Needless to say, the atmospheric effect on solar reflectance is a complex interaction of many factors. Some of these can be measured directly and others cannot. The problem of adjusting for atmospheric effect has fallen under the general heading of haze correction. There are many ways to approach the problem but perhaps the best way in terms of potential accuracy is to measure the solar reflectance at the plant canopy with a spectroradiometer and adjust the corresponding LANDSAT data accordingly. To work effectively, however, many ground measurements have to be taken to adequately estimate mean reflectance for any area large enough to obtain corresponding LANDSAT data. Since registration errors are considered to be in the neighborhood of plus or minus one half pixel, it would be difficult to match ground data with individual pixels. Perhaps one could match ground data with 10 to 15 pixel areas and presumably reduce bias due to registration error but a large number of radiometer measurements would be needed to estimate mean canopy reflectance with good precision. Haze correction by using ground data can work well but is very expensive for large land areas.

Another, less expensive method of haze correction was developed by Ahern et al. (1), and tested in the previously mentioned paper by Richardson et al. Ahern's method makes use of LANDSAT measured radiance of clear water bodies to infer the various parameters needed to apply radiative transfer equations to other surfaces such as plant canopies. Clear water bodies are used because the interaction of solar irradiance with water is sufficiently simplified to permit the estimation of path radiance which cannot be measured directly. With the estimation of the other radiative transfer factors mentioned earlier, an adjustment can be found to change the radiance at the top of the atmosphere as measured by LANDSAT to the radiance at the earth's surface. Of course, implicit in this method is the assumption that the composition of the atmosphere over the lake is the same as that for the other areas of interest.

Ahern's method has the advantage that expensive ground radiometric data are not required. Several apparent disadvantages are that clear water bodies may not be available or near the agricultural areas of interest. Also, Ahern's method assumes that the water bodies are, in fact, clear and calm. Suspended particles change the transmittance properties of the water and, hence, the radiance of the water. Rough water surfaces reflect differently due to increased surface area, glint and shadow. The depth of the water body makes a difference because in clear, shallow water the bottom would reflect according to its appearance. Despite all the possible complications, Richardson et al. reported that LANDSAT data adjusted by Ahern's method did not significantly differ from corresponding ground-based spectroradiometer measurements for four prominent south Texas rangeland plants.

A third method of haze correction which does not employ any of the previously mentioned techniques is the XSTAR Haze Correction Algorithm as proposed by Lambeck (10, 11). The basis for the XSTAR correction was described in a paper by Kauth and Thomas (9). We shall present a condensed description of the underlying logic.

The basis of the method is an attempt to view the life span of a crop in 4-dimensional LANDSAT signal space. While it is difficult to view anything in four dimensions, Kauth has conceived an image of a "tasselled cap" to aid in discussion. The tasselled cap was envisioned by examining 2-dimensional plots of the LANDSAT bands for an agricultural scene which had a wide range of cover types and soil backgrounds during mid-June. The 2-dimensional plots revealed a triangular shape when bands 4 or 5 were plotted against bands 6 or 7. The plots were essentially linear with band 4 versus 5 or band 6 versus 7. In 4-dimensions, the picture would resemble a flattened triangular structure. The triangle shape can be explained with a canopy model using wavelengths at the midpoints of bands 5 and 6. Figure C1 has been reproduced from the paper by Kauth and Thomas.

In Figure C1, points 1A and 1B represent the extremes of soil color -- dark to light. Intermediate shades would fall approximately on a line connecting the two points. This line (called the line of soils) forms the base of the triangle seen when viewing a general agricultural scene.

As the crop (in this case wheat) grows out of the soil, a triangular shape develops to the left of the line of soils. This shape arises from the fact that soil reflects more irradiance than green vegetation in the .65μm wavelength and, except where the soil is very light, just the opposite is true in the .75μm wavelength. The apex of the triangle corresponds to full green canopy cover and is independent of soil backgound so long as canopy cover is full or near full. As the crop yellows and leaves start to senesce, the line of reflectance appears to fall back toward the line of soils. This is true but the line is actually in another plane.

The 4-dimensional triangular shape is refined somewhat by the tasselled cap image. The base of the cap corresponds to the plane of soils. A plane is used rather than a line because in analyzing the principal spectral components of a broad range of soil reflectances, the first principal component was about 7 times larger than the second, 12 times larger than the third and 23 times larger than the fourth. The first two principal components contain most of the spectral information. The top of the cap is the area of green vegetation. The tassells correspond to yellow vegetation of various shades and fall back to the plane of the soils.

Since all the LANDSAT bands contain varying degrees of information about the soil and green and yellow vegetation, Kauth proposed a transformation of the LANDSAT data to separate soil and vegetation effects. This transformation is as follows:

$$\underline{u} = R^T \underline{x} + r$$

where $\underline{x}$ is the LANDSAT signal vector,

$\underline{u}$ is the transformed vector,

r is an arbitrary constant to prevent negative values, and

R is a rotation matrix in which the columns are unit vectors all orthogonal to each other.

The columns of R are as follows:

R_1 is a unit vector pointing along the major axis of the plane of soils,

R_2 is orthogonal to R_1, and pointing toward the green area at the apex of the triangle,

R_3 is orthogonal to R_1 and R_2 and points to a point in the yellow region, and

R_4 is orthogonal to R_1, R_2, and R_3.

The projection of a data point onto R_1, is a feature called "brightness." The projection onto R_2 is called "green stuff." The projection onto R_3 is "yellow stuff." The fourth feature arising from the somewhat vague vector R_4 has an equally vague name, "non-such." If the R matrix is chosen properly, plots of the transformed bands will show orthogonality and magnitude. The plots in the Kauth paper revealed orthogonality among transformed bands and magnitude as measured by the range of values was greatest for the first two bands, much less for the third and practically nothing contained in the fourth. These results were consistent with the particular agricultural scene being viewed.

Among other things, Kauth suggested that the transformed data could be used to diagnose the presence of certain external effects such as haze, water vapor, sun angle and view angle. In the case of haze, the transformed bands would suffer a negative shift in the "yellow stuff" direction accompanied by a positive shift in brightness and a negative shift in greenness. Haze also causes a general loss of contrast. These shifts can be large relative to the magnitude of the reflectances without haze. If one knows generally what to expect in a particular agricultural scene, external effects can be diagnosed and adjusted for.

The XSTAR haze correction adjusts for differences in sun angle; diagnoses pixels which are garbled, clouds, water or cloud shadows; and adjusts for haze. Since the XSTAR correction is comparatively easy to use and does not require LANDSAT data for clear water bodies or ground reflectance measurements, we chose to use it in our study. A method along the lines Ahern suggested appears promising and is deferred for future research. Since our yield-LANDSAT data set is thought to be clean as far as garbled pixels, clouds, water or cloud shadows are concerned, we used the XSTAR correction to adjust for sun angle and differences in haze level.

The details of the XSTAR correction algorithm are contained in a memo from P. F. Lambeck of ERIM (11). They are presented here in a somewhat condensed form. The first step is to calibrate the LANDSAT data to the same type of data the algorithm was developed from. Our data came from both LANDSAT's II and III while the algorithm came from LANDSAT II LACIE segment data.

Let X_{1_i} = LANDSAT II digital count in band i

Let X_{2_i} = LANDSAT III digital count in band i

Then, $X_i' = A_{1i}X_{1i} + B_{1i}$

$$\text{where } A_1 = \begin{pmatrix} 1.275 \\ 1.141 \\ 1.098 \\ .948 \end{pmatrix} \text{ and } B_1 = \begin{pmatrix} -1.445 \\ -2.712 \\ -2.950 \\ .446 \end{pmatrix}$$

Also, $X_i' = A_{2i}X_{2i} + P_{2i}$

$$\text{where } A_2 = \begin{pmatrix} 1.1371 \\ 1.1725 \\ 1.2470 \\ 1.1260 \end{pmatrix} \text{ and } B_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Set $X_i = X_i'$

The second step is to correct for differences in sun angle. The sun angle is specified with each LANDSAT acquisition.

Let θ = Solar zenith angle

X_i = digital count following step 1

Then $X_i' = \dfrac{\cos\theta_0}{\cos\theta} X_i$

where $\theta_0 = 39°$

Set $X_i = X_i'$

Step 3 is to calculate the scene diagnostic signal value for each acquisition.

Let X_{ijk} = digital count following step 2 in band i, acquisition j and pixel k

$$\text{Then, } \hat{X}_{ij} = \frac{\sum_{k=1}^{n} X_{ijk}}{n}$$

where n = number of pixels in i^{th} band and j^{th} acquisition.

Step 4 is to determine the amount of change in optical thickness (γ) from a reference haze condition. This is calculated by rotating the $\hat{X}_{ij}$ from step 3 and comparing the magnitude of the value in the "yellow stuff" direction to the reference yellow value. As mentioned earlier, increasing haze causes the yellow value to move in a negative direction. By comparing the calculated yellow value to the reference yellow value, a change in optical thickness ( ) from the reference thickness can be calculated. Since the optical thickness is exponentially related to the atmospheric transmittance fraction, a change in optical thickness of zero would produce a change in atmospheric transmittance of one. The change in atmospheric transmittance is multiplicative so that a value of one would produce no change in the LANDSAT signal value. If γ is negative, the calculated change in atmospheric transmittance is between zero and one and the LANDSAT signal value is increased. In other words, if the haze level in the LANDSAT data is less than the reference haze level as measured by the relative magnitude of the yellow value, the LANDSAT signal values are increased. The opposite is true if the haze level in the LANDSAT data is greater than the reference haze level. This sort of a standardization of haze level is intended to decrease the effect of differential amounts of haze over different LANDSAT acquisitions. The procedure used to calculate γ is as follows:

$$\text{Let } R_i = \begin{pmatrix} -.89952 \\ .42830 \\ .07592 \\ -.04080 \end{pmatrix} = \text{3rd column of rotation matrix}$$

$$\alpha_i = \begin{pmatrix} 1.2680 \\ 1.0445 \\ .9142 \\ .7734 \end{pmatrix} = \text{values related to reference optical thickness}$$

$$X_i^* = \begin{pmatrix} 61.9 \\ 66.2 \\ 83.2 \\ 33.9 \end{pmatrix} = \text{very hazy point in LANDSAT signal space called point of all haze}$$

$$Y^* = -11.2082 \quad = \text{reference yellow value}$$

$\hat{X}_{ij}$ = scene diagnostic signal value from step 3.

$$\text{then, } \gamma_j = -\frac{b_j}{a_j}\left(1-(1-\frac{2a_jc_j}{b_j^2}^{1/2}\right)$$

$$\text{where } a_j = \sum_{i=1}^{4} \alpha_i^2 \; (\hat{X}_{ij} - X_i^*) \; R_i$$

$$b_j = \sum_{i=1}^{4} \alpha_i \; (\hat{X}_{ij} - X_i^*) \; R_i$$

$$c_j = (\sum_{i=1}^{4} \hat{X}_{ij} \; R_i) - Y^*$$

The fifth and final step is to apply the XSTAR correction to the LANDSAT digital counts following step 2.

Let X_{ij} = LANDSAT digital count following step 2 for i^{th} band and j^{th} acquisition.

γ_j = change in optical thickness from step 4 for acquisition j.

α_i and X_i^* are as defined in step 4.

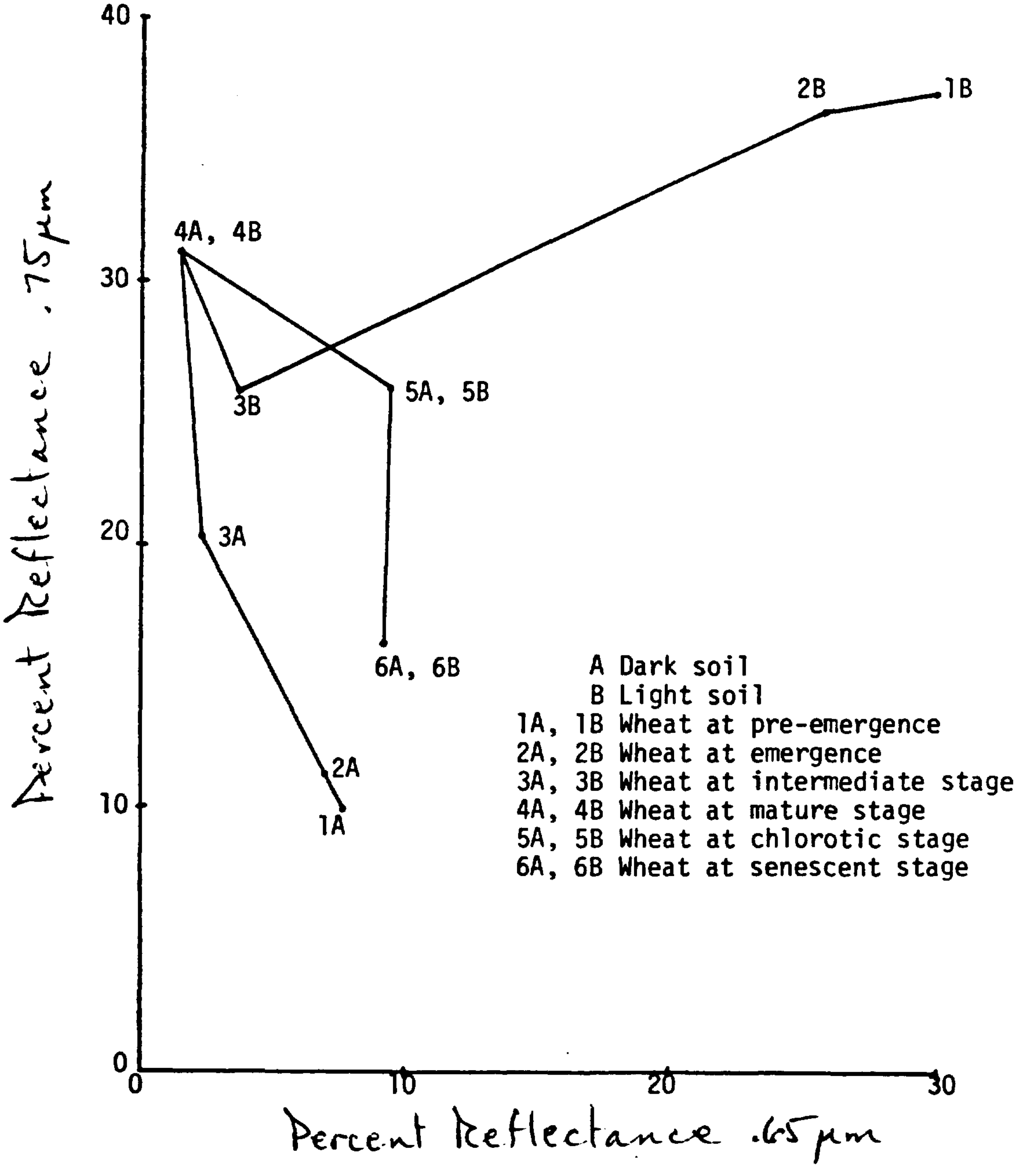
Then, $X_{ij}' = \mathrm{Exp}\ (\alpha_i \gamma_j)\ (X_{ij} - X_i^*) + X_i^*$

Set $X_{ij} = X_{ij}'$

The corrected LANDSAT data which were obtained by the above procedure are larger than the corresponding uncorrected data. The correction just for the differential haze levels in steps 3 through 5 generally increased the raw digital counts but in one LANDSAT pass where haze was known to be a problem, the correction caused a negative shift.

While the XSTAR haze correction algorithm is comparatively convenient to use, the basis of the method is the magnitude of the calculated yellow vector relative to a reference. The rotation matrix to transform the LANDSAT signal values was obtained from an agricultural scene containing many different crops and soil types. As mentioned earlier, the rotation matrix did produce orthogonality among the transformed channels in the particular data set in the Kauth paper. However, in our data set, only two crops were involved and since all the imagery was obtained between early August and early September, we generally expect to have full canopy cover and green vegetation. This data space is therefore only a small subset of the data space used to obtain the rotation matrix. This raises questions as to whether the rotation matrix in the XSTAR correction is still applicable. We plotted the transformed bands against one another and they did not *appear* to be orthogonal in most cases. However, since we are only looking at a small portion of the possible range of LANDSAT response, it is impossible to tell whether this is causing the appearance of nonorthogonality in itself or whether the rotation matrix is incorrect.

Figure C1 - Phenology for wheat based on canopy model.



Appendix D - Alternative Regression Estimators

A further note should be added concerning the potential benefits of using LANDSAT data. As pointed out earlier, the regression estimator considered in this report is biased to the extent of classification error. An unbiased estimator would likely have higher relative efficiency because the MSE would not be inflated by a squared bias term. One such unbiased regression estimator is

$$\hat{\bar{\bar{P}}}_{reg} = \hat{\bar{\bar{P}}}_{DE} + \underline{\hat{B}}'(\underline{\bar{\bar{W}}} - \underline{\bar{x}}^* - \underline{\hat{D}})$$

where the bias, $\underline{\hat{D}}$ is as estimated in Appendix B. Evaluation of this regression estimator remains for future research. Regression yield estimates were not actually calculated in this report because the "wall to wall" set of LANDSAT data classified to the crop of interest ($\underline{\bar{\bar{W}}}$) has not yet been obtained. This task also remains for future research contingent upon the value of the results demonstrated thus far.

Another alternative unbiased regression estimator which avoids classification error completely is similar to the biased estimator except $\underline{\bar{\bar{W}}}$ is replaced by $\underline{\bar{\bar{X}}}$. That is, rather than having a large sample consisting of all LANDSAT data classified to the crop of interest, the large sample would be just those fields within the JES segments known to be in the crop. This is a considerably smaller "large" sample size than with $\underline{\bar{\bar{W}}}$ but is still large relative to the small sample size consisting of OY fields. While this regression estimator suffers no bias due to classification error, it may have a different problem. In using a double sampling regression estimation approach, we are assuming that the large sample is a random sample from the population and the small sample is a random subset of the large sample. The biased regression estimator is alright in this regard since the large sample is the whole population and the OY field can be correctly considered a random sample from the population. However, in the presently proposed unbiased regression estimator, the OY fields are clustered within the JES segments. Strictly speaking then, the small sample is not a random subsample of the JES fields. This may not be a problem if it can be shown that the correlation among field-level LANDSAT means within JES segments is not larger than the between segment correlation. If the segments are considered as clusters of OY fields then when the intracluster correlation is positive a one-way analysis of variance can be used to test for the presence of a clustering effect. If the within sum of squares divided by the total sum of squares is smaller than .5 then it would probably be safe to use the proposed regression estimator. This analysis remains for future research.

Appendix E - Rankings of Biased Regression Estimators

Table E1 - Biased-regression estimators for corn with field yields estimated from 8 plots per field, ranked by relative efficiency, ten estimators per analysis district.

analysis district	LANDSAT variables[1]	R^2	estimated relative bias (%)	VIF	C.V.(%)	MSEIF	estimated relative root-MSE(%)	RE[2]
1 (1)	-(S,R,G)	.19	0.04	1.003	2.2	1.003	2.2	1.15
(2)	-R	.11	0.18	1.001	2.2	1.007	2.2	1.1±.02
(3)	-(2,3,R)	.20	-0.67	1.026	2.2	1.127	2.3	1.1
(4)	-(2,R,G)	.27	-1.40	1.059	2.2	1.520	2.6	0.85
(5)	-(V,T)	.23	-1.48	1.098	2.2	1.602	2.7	0.8±5.5
(6)	-(T,R)	.22	-1.85	1.148	2.3	1.915	3.0	0.6±0.7
(7)	-(V,R)	.20	-1.86	1.160	2.3	1.932	3.0	0.6±0.75
(8)	-(S,R)	.29	-2.97	1.204	2.3	3.359	3.8	0.4
(9)	-(P,S,R)	.29	-2.99	1.207	2.3	3.393	3.9	0.4
(10)	-(S,D,R)	.29	-2.99	1.207	2.3	3.393	3.9	0.4
3 (1)	-(6,7)	.41	0.5	1.005	4.9	1.015	5.0	1.6±(.35-.39)
(2)	-(R,G)	.40	0.5	1.002	5.0	1.011	5.0	1.6±.21
(3)	-(6,R)	.37	-.05	1.002	5.2	1.002	5.2	1.5±.01
(4)	-(6,P)	.37	.02	1.002	5.1	1.002	5.2	1.5±.01
(5)	-(V,G)	.37	.17	1.001	5.1	1.002	5.2	1.5±.07
(6)	-(D,G)	.37	.28	1.001	5.2	1.004	5.2	1.5±(.14-.16)
(7)	-(T,G)	.36	.28	1.001	5.2	1.004	5.2	1.4±.10
(8)	-(6,V)	.35	-.18	1.001	5.3	1.003	5.3	1.4±.04
(9)	-(6,T)	.34	-.12	1.001	5.3	1.002	5.3	1.4±.03
(10)	-(6,D)	.34	-.20	1.001	5.3	1.003	5.3	1.4±.04
4 (1)	-5	.11	1.36	1.034	3.5	1.1848	5.1	0.9±.09
(2)	-D	.11	1.68	1.052	3.5	1.2833	5.3	0.9±.08
(3)	-T	.13	1.90	1.057	3.5	1.3578	5.4	0.8±.11
(4)	-R	.12	1.88	1.058	3.5	1.3511	5.4	0.8±.08
(5)	-V	.12	1.93	1.058	3.5	1.3684	5.4	0.8±.08
(6)	+7	.11	1.99	1.076	3.6	1.3969	5.6	0.8±.08
(7)	-(6,7)	.14	2.68	1.086	3.5	1.6975	6.0	0.7±(.26-.29)
(8)	+P	.13	2.66	1.104	3.5	1.6996	6.0	0.7±.08
(9)	+5	.13	2.72	1.113	3.5	1.7305	6.1	0.7±.09
(10)	+D	.15	3.02	1.120	3.5	1.9044	6.3	0.6±.07

[1] See footnote, Table 8.
[2] See footnote, Table 8.

Table E2 - Biased-regression estimators for soybeans with field yields estimated from 8 plots per field, ranked by relative efficiency, ten estimators per analysis district.

analysis district	LANDSAT variables[1]	R^2	estimated relative bias (%)	VIF	C.V.(%)	MSEIF	estimated relative root-MSE(%)	RE[2]
1 (1)	-(4,S)	.66	0.007	1.012	2.9	1.012	2.9	2.7±.04
(2)	-(4,7)	.65	-0.09	1.012	2.9	1.013	3.0	2.7±.12
(3)	-D	.64	0.07	1.00002	3.0	1.0006	3.0	2.7±.12
(4)	-(5,7)	.64	-0.1	1.015	3.0	1.017	3.0	2.5±.18
(5)	-(5,S)	.64	-0.1	1.015	3.0	1.017	3.1	2.5±.18
(6)	-(7,S)	.64	-0.1	1.015	3.0	1.017	3.1	2.5±(.62-.80)
(7)	-P	.62	0.42	1.0005	3.1	1.020	3.1	2.4±.54
(8)	-(4,6)	.63	-0.55	1.011	3.1	1.043	3.1	2.4±.65
(9)	-R	.63	0.66	1.001	3.1	1.048	3.1	2.4±.85
(10)	-G	.61	0.1	1.00006	3.1	1.002	3.2	2.3±.16
2 (1)	-(6,G)	.52	0.7	1.019	3.2	1.073	3.2	1.9±(1.4-1.7)
(2)	-V	.48	-0.1	1.0001	3.2	1.001	3.2	1.9±.05
(3)	-4	.48	0.2	1.0002	3.3	1.004	3.3	1.9±.05
(4)	-T	.48	-0.2	1.0002	3.3	1.005	3.3	1.9±.05
(5)	-(7,S)	.47	-0.4	1.0006	3.3	1.015	3.3	1.8±.35
(6)	-(P,D)	.47	-0.4	1.0006	3.3	1.016	3.3	1.8±(.6-.7)
(7)	-(S,D)	.47	-0.4	1.0006	3.3	1.016	3.3	1.8±.2
(8)	-(P,S)	.47	-0.4	1.0006	3.3	1.016	3.3	1.8±.3
(9)	-(7,D)	.47	-0.4	1.0006	3.3	1.016	3.3	1.8±.4
(10)	-(7,P)	.47	-0.4	1.0006	3.3	1.016	3.3	1.8±.9
3 (1)	+4	.42	0.6	1.002	5.3	1.015	5.3	1.5±.11
(2)	+(S,G)	.40	0.6	1.010	5.4	1.023	5.4	1.5±.21
(3)	+(S,T)	.38	-0.4	1.003	5.4	1.009	5.4	1.5±.10
(4)	+(7,V)	.38	0.07	1.005	5.4	1.005	5.4	1.5±.02
(5)	+(P,V)	.38	0.09	1.005	5.5	1.005	5.5	1.5±.04
(6)	+(7,T)	.40	-1.0	1.010	5.3	1.047	5.4	1.4±.5
(7)	+(S,V)	.37	0.05	1.005	5.5	1.005	5.5	1.4±.01
(8)	+(7,S)	.36	0.02	1.005	5.5	1.005	5.5	1.4±.02
(9)	+(P,S)	.36	0.02	1.005	5.5	1.005	5.5	1.4±.01
(10)	+(7,P)	.36	0.01	1.005	5.5	1.005	5.5	1.4±.01

[1] See footnote Table 8.

[2] See footnote Table 8.

Appendix F - Prediction of Known Field Yield R^2's

From the two sets of regressions with dependent variables of field yield estimated from two and eight plots per field, respectively, the R^2 which would result if field yields were known can be predicted.

Assume

$$y_{ij} = \text{yield of plot } j \text{ in field } i$$

$$= \text{regression} + u_{ij}$$

and $\quad u_{ij} = v_i + e_{ij}$

where $\quad Ev_i = Ee_{ij} = 0$

$$\begin{aligned} Cov(v_i v_r) &= \sigma_v^2 \quad \text{if } i=r \\ &= 0 \quad \text{if } i \neq r \end{aligned}$$

and $\quad Cov(e_{ij} e_{rs}) = \sigma_e^2$ if $i=r$, $j=s$

$\qquad = 0$ otherwise.

Hence

$$Eu_{ij} = 0$$

and
$$\begin{aligned} E(u_{ij} u_{rs}) &= \sigma_v^2 + \sigma_e^2 \quad &\text{if } i=r,\ j=s \\ &= \sigma_v^2 \quad &\text{if } i=r,\ j\neq s \\ &= 0 \quad &\text{if } i\neq r. \end{aligned}$$

Then the following analysis of variance (AOV) tables result from regressions with different sets of dependent variables:

AOV for Field Yield Estimated from k Plots per Field

Source	df	E(MS)
Regression	p	(explained variation)/p
Residual	n-p-1	$\sigma_v^2 + \sigma_e^2/k$
Total	n-1	

AOV for Known Field Yields

Source	df	E(MS)
Regression	p	(explained variation)/p
Residual	n-p-1	σ_v^2
Total	n-1	

Thus

$$E(MS^{(8)}_{resid}) = \sigma_v^2 + \sigma_e^2/8$$

and $E(MS^{(2)}_{resid}) = \sigma_v^2 + \sigma_e^2/2.$

Hence

$$\sigma_v^2 = (4\ E(MS^{(8)}_{resid}) - E(MS^{(2)}_{resid}))/3$$

and $(n-p-1)\,\sigma_v^2 = (4E(SS^{(8)}_{resid}) - E(SS^{(2)}_{resid}))/3.$

The R^2 which would result if field yields were known is thus predicted by

$$\frac{\text{explained variation}}{\text{explained variation} + (n-p-1)\,\sigma_v^2}$$

$$= \frac{SS^{(8)}_{reg}}{SS^{(8)}_{reg} + (4\ SS^{(8)}_{resid} - SS^{(2)}_{resid})/3}$$