COUNTY-LEVEL CROP AREA ESTIMATES FROM LANDSAT: EVALUATION OF THE BATTESE-FULLER ESTIMATOR

Gail Walker and Richard Sigman, U.S. Department of Agriculture

I. INTRODUCTION

Annually in late May and early June the Statistical Reporting Service (SRS) of the U.S. Department of Agriculture conducts the nationwide June Enumerative Survey (JES). From the data collected in the JES, state and national estimates of the amount of land planted to various crops are calculated, as well as estimates of intended crop utilization, farm grain storage, livestock inventories, agricultural labor, and farm economic data.

Crop-area and production estimates for individual counties are also an integral part of the SRS estimates program. Such estimates are used by the Agricultural Stabilization and Conservation Service and by the Federal Crop Insurance Corporation. Published county estimates are used by agri-business concerns in making decisions on marketing of farm products and in transportation scheduling of agricultural commodities.

SRS calculates county estimates by subdividing the official state estimate into crop reporting districts (collections of contiguous counties) and then further subdividing into counties. Several types of indicator data are used in subdividing the state estimate. These include:

1. JES expansions at a district level,
2. Non-probability mail surveys, and
3. State farm census data.

The resulting estimates are at least partially subjective and as a result variance estimates for individual counties are not calculable using this method.

In recent years, a number of states have discontinued their state farm census. This has prompted research by SRS into alternative methods of calculating county estimates. Ford (1981), for example, evaluates direct, synthetic, and composite estimators for crop and livestock items utilizing a probability mail survey in North Carolina.

For county crop-area estimates, a number of researchers have proposed the auxiliary use of data from the LANDSAT earth-resources satellite. The model-based estimators proposed by Huddleston and Ray (1976) and by Battese and Fuller (1981) are discussed later in this paper. Cardenas, Blanchard, and Craig (1978) have proposed a LANDSAT-adjusted synthetic estimator for calculating county crop-area estimates. In this paper we extend the Battese-Fuller estimator to the case of a stratified sample design and evaluate the Battese-Fuller estimator on a six-county area in eastern South Dakota.

II. DATA SOURCES

A. Ground-Survey Data. JES sample units, called segments, are selected from an area sampling frame. Segments are typically one square mile and are selected from strata defined in terms of the percent of cultivated land.

During the JES interview, all fields within the sampled segment are delineated on a non-current aerial photograph, and the crop or land use of each delineated field is recorded on a questionnaire.

B. LANDSAT Data. The basic element of LANDSAT data is the set of measurements taken by the satellite's multispectral scanner (MSS) of a 0.4 hectare area of the earth's surface. The MSS measures the amount of radiant energy reflected from the earth's surface in four different regions of the electromagnetic spectrum. The individual 0.4 hectare MSS resolution areas, referred to as pixels, are arrayed along east-west rows within the 185 kilometers wide north-to-south pass of the LANDSAT satellite. For purposes of easy data storage, the data within a swath are subdivided into overlapping square blocks, called scenes, which are 185 kilometers on a side.

III. ANALYSIS-DISTRICT LANDSAT ESTIMATOR

An analysis district is a collection of counties or portions of counties completely contained in one to three LANDSAT scenes having the same image date. In the midwestern United States, where most of the SRS LANDSAT research has been conducted, a typical analysis district contains a minimum of ten counties.

For analysis districts, SRS uses the regression estimator described by Cochran (Section 7.1.7, third edition) to obtain crop-area estimates which are more precise than the JES estimates. This procedure is described in detail in Sigman, et al (1978). Briefly, the SRS analysis-district procedure is as follows:

1. The JES data for segments in the analysis district are used to label segment LANDSAT pixels as to crop type.
2. Labeled LANDSAT pixels are used to develop discriminant functions for each crop type. (A discriminant function for "other" is also developed.)
3. The discriminant functions are used to classify the LANDSAT data in the sampled JES segments. The classification results for each segment are the auxiliary variable for the regression estimator. The survey results for each segment are the primary variable.
4. The discriminant functions are used to classify all pixels within the analysis district from which the population mean per segment of the auxiliary variable can be calculated.

The estimation procedure described above is carried out in each analysis district, and then analysis-district estimates as well as variances are combined to the state level by treating the analysis districts as post-strata. The above procedure imposes a lower bound on the size of the JES sample within the analysis district. The reasons for this are the following:

1. If the separate form of the regression estimator is used, there must be enough segments in each stratum of the analysis district to estimate the stratum regression coefficients, or
2. If the combined form of the regression estimator is used, there must be enough segments in the analysis district to estimate the combined regression coefficient.

In the mid-western United States, counties typically contain only two to four sampled JES segments and may contain no sampled segments. Thus, defining analysis districts to be individual counties and then using the above procedure is generally not feasible.

IV. LANDSAT SMALL AREA ESTIMATION

A. Huddleston-Ray Procedure. As presented above, crop acreage estimation for analysis districts is a straightforward use of a regression estimator. To provide a set of estimates for each county contained in the analysis district, Huddleston and Ray (1976) proposed that the mean calculated by classifying the entire

analysis district, $\bar{X}_{a.d.}$, be replaced by the mean calculated by classifying the full set of potential segments from a particular county, $\bar{X}_c$.

Thus, the analysis district regression estimator for the mean per segment is:

$$REG_{a.d.} = \bar{y}_{a.d.} + b_1 (\bar{X}_{a.d.} - \bar{x}_{a.d.})$$

$$= b_o + b_1 \bar{X}_{a.d.}$$

and the Huddleston-Ray county estimator is:

$$HR_c = \bar{y}_{a.d.} + b_1 (\bar{X}_c - \bar{x}_{a.d.})$$

$$= b_o + b_1 \bar{X}_c.$$

B. Battese-Fuller Model. The Battese-Fuller model for county level estimation assumes that segments grouped by county admit the same rate of change relationship (slope) as does the analysis district but that a different intercept is required. This idea is implemented by using a portion of the vertical distance from the analysis district regression line to the county sample mean. Denoting this distance by $\bar{u}_c = \bar{y}_c - b_o - b_1\bar{x}_c$, the Battese-Fuller county estimator is:

$$BF_c = b_o + b_1 \bar{X}_c + \delta_c \bar{u}_c \text{ where } 0 \leq \delta_c \leq 1.$$

This introduction is an oversimplification. Estimating county effects by $\bar{u}_c$ precludes the use of ordinary least squares in fitting the analysis district regression line and thus the choice of $\delta_c = 0$ does not coincide exactly with the Huddleston-Ray estimate.

More precisely, as originally proposed, the Battese-Fuller model assumes that for the j^{th} sampled segment from the i^{th} county we have:

$$y_{ij} = b_o + b_1 x_{ij} + u_{ij} = b_o + b_1 x_{ij} + v_i + e_{ij}$$

v_i, e_{ij} independent, normal with mean 0 and variances σ_v^2 and σ_e^2 respectively

$$\text{cov}(u_{ij}, u_{i'j'}) = \begin{cases} 0 & \text{if } i \neq i' \\ \sigma_v^2 & \text{if } i = i',\ j \neq j' \\ \sigma_v^2 + \sigma_e^2 & \text{if } i = i',\ j = j' \end{cases}$$

Thus, segments from the same county possess positively correlated residuals. The parameter σ_v^2 is both a within county covariance and a between county component of the variance of any residual. σ_e^2 is the within county variance component. This set of assumptions reduces to the standard assumptions of ordinary least squares when $\sigma_v^2 = 0$.

Assuming first that b_o and b_1 are known, the county mean residuals

$$\bar{u}_{i.} = \bar{y}_{i.} - b_o - b_1 \bar{x}_{i.} = v_i + \bar{e}_{i.}$$

are observable and give estimated county effects of

$$\hat{v}_i = \delta_i \bar{u}_{i.} \text{ where } 0 \leq \delta_i \leq 1.$$

The county mean is estimated by

$$b_o + b_1 \bar{X}_i + \delta_i \bar{u}_i.$$

with error equal to $(1 - \delta_i) v_i - \delta_i e_i$.

It follows that

$$MSE = (1 - \delta_i)^2 \sigma_v^2 + \delta_i^2 \frac{\sigma_e^2}{n_i}$$

where n_i is the size of the sample from county i. Note that, conditioned on the county effects, the average error is $(1 - \delta_i) v_i$. Squaring and averaging gives a mean squared conditional bias of:

$$MSCB = (1 - \delta_i)^2 \sigma_v^2.$$

As a function of δ_i, it is easy to see that the above expression for MSE is minimized if

$$\delta_i = \frac{\sigma_v^2}{\sigma_v^2 + \frac{\sigma_e^2}{n_i}}.$$

Denoting this quotient by γ_i, we focus our attention on the three specific estimates obtained from:

a. $\delta_i = 0$
- estimate lies on analysis district regression line
- MSE = MSCB

b. $\delta_i = 1$
- MSCB = 0

c. $\delta_i = \gamma_i$
- minimum MSE is obtained
- $\frac{MSCB}{MSE} = 1 - \gamma_i$

Note that estimates for unsampled counties may be obtained by choosing $\delta = 0$.

As discussed in the Battese-Fuller paper, a best linear unbiased estimate $\hat{b}$ for an unknown b is obtainable by an appropriate transformation of the data. A fitting of constants procedure handles estimation of the variance components. Formulas for the MSE and MSCB when b is estimated are given in Battese and Fuller (1981) and in Walker and Sigman (1982). The same choice of $\delta_i = \gamma_i$ minimizes the MSE when b is estimated.

C. Stratification. Like the regression procedure used at the analysis district level, the Battese-Fuller model is applicable within individual strata. The procedures set forth by Battese and Fuller and presented above suffice for estimating b_o, b_1, σ_v^2, σ_e^2 in each stratum. However, the presence of a county main effect across strata introduces a cross strata covariance and requires revisions in both the MSE formula and the choice of an optimal set of multipliers for the mean residuals.

At Fuller's suggestion, the authors developed the following extension of the model presented in the last section. For the j^{th} segment from county i and stratum h, assume that

$$y_{hij} = b_h^0 + b_h^1 x_{hij} + v_{hi} + e_{hij}$$

with variance - covariance structure

$$\text{cov}(u_{hij}, u_{h'i'j'}) = \begin{cases} 0 & \text{if } i \neq i' \\ \sigma_{v_h}^2 & \text{if } i = i',\ h = h',\ j \neq j' \\ \sigma_{v_h}^2 + \sigma_{e_h}^2 & \text{if } i = i',\ h = h',\ j = j' \\ \sigma_{v_{hh'}} & \text{if } i = i',\ h \neq h' \end{cases}$$

Under these assumptions one must estimate a vector of county effects denoted $v^i = (v_{1i}, \ldots, v_{si})'$ where s is the number of strata. Each component is estimated using the vector of mean residuals $\bar{u}^i = (\bar{u}_{1i}, \ldots, \bar{u}_{si})'$ where

$$\bar{u}_{hi} = \frac{1}{n_{hi}} \sum_{j=1}^{n_{hi}} u_{hij}$$

thereby requiring an s by s coefficient matrix. That is;

$$\hat{\mu}_{hi} = b_h^0 + b_h^1 \bar{X}_{hi} + \sum_{k=1}^{s} c_{kh}^i \bar{u}_{ki}$$

estimates the average amount of the crop per segment for the part of county i that falls into stratum h. The mean for the county is then the appropriate weighted sum over strata.

To put this in a convenient notation, let

$$BX^i = \begin{pmatrix} 1 & \bar{X}_i & . & . & 0 & 0 \\ . & . & & & . & . \\ . & . & & & . & . \\ . & . & & & . & . \\ 0 & 0 & . & . & 1 & \bar{X}_{si} \end{pmatrix}$$

and similarly for LX^i using $\bar{x}_{hi}$. Also, set

$$B = (b_1^0, b_1^1 \ldots, b_s^0, b_s^1)'$$

and

$$w^i = \left(\frac{N_{1i}}{N_{\cdot i}}, \ldots, \frac{N_{si}}{N_{\cdot i}}\right)$$

where N_{hi} = total number segments in county i and stratum h and $N_{\cdot i} = \sum_h N_{hi}$.

For known b values, the vector of estimated means for county i is

$$\hat{\underset{\sim}{\mu}}_i = BX^i \cdot B + C^{i\prime} \underline{\bar{u}}^i$$

and the final county mean is estimated by

$$\hat{\mu}_i = w^i \hat{\underset{\sim}{\mu}}_i .$$

Introducing the s by s matrices

$$H = E(v^i v^{i\prime}) = \begin{pmatrix} \sigma^2_{v_1} & . & . & . & \sigma_{v_1 s} \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ \sigma_{v_1 s} & . & . & . & \sigma^2_{v_s} \end{pmatrix}$$

and

$$SE^i = \begin{pmatrix} \frac{\sigma^2_{e_1}}{n_{1i}} & . & . & . & 0 \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ 0 & . & . & . & \frac{\sigma^2_{e_s}}{n_{si}} \end{pmatrix}$$

we have $A^i = E(u^i u^{i\prime}) = H + SE^i$.

Then

$$MSE(\mu_i) = w^i E((v^i - C^{i\prime}u^i)(v^{i\prime} - u^{i\prime}C^i))w^{i\prime}$$
$$= w^i(H - 2HC^i + C^{i\prime}A^iC^i)\, w^{i\prime}$$

and

$$MSCB = w^i(H - 2HC^i + C^{i\prime}HC^i)\, w^{i\prime}.$$

Applying a minimization criterion to each component of v^i results in

$$C^i = (A^i)^{-1}H$$

which reduces to

$$C^i = \begin{pmatrix} \gamma_{1i} & & & 0 \\ & . & & \\ & & . & \\ 0 & & & \gamma_{si} \end{pmatrix}$$

if $\sigma_{v_{hk}} = 0$ for all $h \neq k$.

The coefficient matrices for which we carried out the estimation procedure are the following:

a. $C^i =$ 0
- regression line used in each stratum
- MSE=MSCB

b. $C^i =$ I
- MSCB = 0

c. $C^i = \Gamma^i = \begin{pmatrix} \gamma_{1i} & & 0 \\ & . & \\ & & . \\ 0 & & \gamma_{si} \end{pmatrix}$

- minimizes MSE if $\sigma_{v_{hk}} = 0$

d. $C^i = (A^i)^{-1}H$
- minimizes MSE in general

The estimates obtained using these matrices will be denoted BFREG, BFONE, BFGAM and BFOPT, respectively, in section V.C. The Huddleston-Ray estimate discussed in section IV. A. will be denoted HR.

Formulas for the mean square error and mean square conditional bias when b is estimated are given in Walker and Sigman (1982).

V. EVALUATION OF BATTESE-FULLER ESTIMATOR

A. Description of Data Set. An empirical evaluation of the Battese-Fuller estimator was performed over a six-county area in eastern South Dakota. The major feature of this data set which made it advantageous for use in a county-estimation study was that it contained a large number of segments within a relatively small area. Specifically, there were enough segments to calculate a within-county regression estimate for each county against which to compare other county estimators. This amounts to treating each county like an analysis district. Also, there were enough segments in the data set to simulate repeated selection of samples smaller in size then the full data set. A negative feature of the data set, however, is that the quarter-section (160 acres) segment size is smaller than normal JES segments.

Table 1 shows the sample size broken down by county and stratum.

Table 1: Sample Allocation by County and SRS Stratum

	Stratum			
County	11	12	20	Total
Codington	8	14	5	27
Spinks	21	24	2	47
Beadle	13	26	3	42
Clark	15	14	7	36
Kingsbury	7	21	2	30
Hamlin	10	8	0	18
	74	107	19	200

For purposes of simulating repeated samples, eight samples of size 75 were developed by dividing the 200 segments into 8 mutually exclusive sets and then forming samples from groups of three sets. Calculation of

discriminant functions, classification of LANDSAT data, and calculation of Battese-Fuller county estimates were performed for each sample of 75 and for the full sample. A lengthier description of the data set appears in Walker and Sigman (1982).

B. Validity of Model Assumptions. To determine whether or not the assumptions of the Battese-Fuller estimator are valid, ordinary least-squares LANDSAT regressions were performed within strata 11 and 12 for each of the six South Dakota counties. The following statistics of comparison were calculated:

$\hat{b}_{hi}$ = regression intercept for stratum h, county i

S^2_{hi} = error mean sum of squares for stratum h, county i

$\hat{b}^1_{hi}$ = regression slope for stratum h, county i

If the unstratified Battese-Fuller model assumptions are true, then the calculated comparison statistics satisfy the following properties:

1. Each $\hat{b}_{hi}$ is an unbiased estimate of $b_0 + v_i$.
2. Each S^2_{hi} is an estimate of σ^2_e.
3. Each $\hat{b}^2_{hi}$ is an unbiased estimate of b_1.

If, on the other hand, the stratified Battese-Fuller model assumptions are correct, the comparison statistics will exhibit the following behavior:

4. $\hat{b}_{hi}$ unbiasedly estimates $b^0_h + v_{hi}$.
5. S^2_{hi} estimates $\sigma^2_{e_h}$ for each county in stratum h.
6. $\hat{b}^1_{hi}$ unbiasedly estimates b^1_h within statum h.

The above statements and alternatives to them can be concisely expressed by using the regression-hypothesis notation of McLaughlin (1975). McLaughlin considers the triplet of parameter vectors

(intercepts, residual variances, slopes)

for a set of regressions. A hypothesis concerning the triplet is denoted by a three-letter word. The component letters correspond in position to the triplet parameter vectors, and each letter is either E for homogeneity (equality) or V for heterogeneity (variability).

For the case of regressions performed within each stratum of each county, we extend the notation as follows:

E = Homogeneity across both strata and counties

Ec = Vs = Homogeneity across counties within each stratum. Heterogeneity across stata.

Es = Vc = Homogeneity across strata within each county. Heterogeneity across counties.

V = Heterogeneity across both counties and strata.

Thus, statements 1 through 3 above, are the hypothesis VcEE and statements 4 through 6 the hypothesis VEcEc. These models can be tested using the procedure described in (McLaughin, 1975).

Though the Battese-Fuller estimator does not require that the form of the probability distributions of the regression errors be known, testing of the postulated model assumptions does. We assume that the regression errors have Gaussian distributions.

Walker and Sigman (1982) contains the model test results. Only model VVE_c for corn cannot be readily rejected (p = .21). This model for corn assumes that regression slopes are homogeneous across counties within each strata but that intercepts and error variances are heterogeneous. For sunflowers, flax, and oats there is significant heterogeneity of regression slopes across counties.

Though the likelihood ratio tests reject VVE_c for all crops except corn, further study indicated that departures from the model (homogeneous slopes across counties within each stratum; heterogeneous intercepts and residual variances) are not overly large for oats and sunflowers, but model departures are pronounced for flax. Furthermore, the heterogeneity of regression slopes is more evident for low R^2 values, where R^2 is the coefficient of determination between classification results and ground truth.

Models which assume the homogeneity of error variance across counties were readily rejected. Flax, oats, and sunflowers exhibit high heteroscedacity, whereas for corn the departure from homogeneous error variances is moderate.

In summary, the model tests performed do not support either the unstratified or the stratified assumptions for the Battese-Fuller estimator. For corn, and corn only, the heterogeneity of stratum regression slopes over counties was not significant, but this was accompanied by heterogeneity of residual variances. Sunflowers and oats failed model tests for homogeniety of stratum regression slopes, but the observed departures from homogeniety were not overly large.

C. Results. The fitting of constants procedure discussed in Battese-Fuller (1981) was used to obtain estimates of the variance components $\sigma^2_{v_h}$ and $\sigma^2_{e_h}$ in each stratum and an F test of the hypothesis H_0: $\sigma^2_{v_h} = 0$ was carried out. The between county variance component $\sigma^2_{v_h}$ has a large variance; a situation that would be eased if the number of counties in the region was greater. The sample sizes in stratum 20 were too small to provide viable estimates of $\sigma^2_{v_{20}}$, so ordinary least squares regression was used in that stratum.

The most convincing evidence of a nonzero county effect was found for corn in both strata and for oats in stratum 12.

Correlations of residuals within and across strata were found from the estimated variance components. Except for corn, low within strata correlations resulted because $\sigma^2_{v_h}$ was small relative to $\sigma^2_{e_h}$. See Walker and Sigman (1982) for details.

It seemed appropriate to assume that $\sigma^2_{v_{11,12}} = 0$ for all crops except corn. Moreover, the procedures described herein do not guarantee that the estimated matrix $H = E(v^i v^{i'})$ will be positive definite and, indeed, four of the eight groups posed this problem.

For all crops and all groups estimation was carried out using $\sigma^2_{v_{11,12}} = 0$. For the set of all 200 segments and half of the eight smaller groups, we also obtained estimates for corn using a nondiagonal H. This provides information on the effect of ignoring the cross strata correlation.

Values of the optimal scale factor γ_{hi} appear in table 2 indicating that we were able to make a sizeable adjustment away from the regression line when estimating corn. Note that flax and sunflowers usually require the use of a regression line estimate in at least one stratum.

Table 2: Optimal Scale Factor γ_{hi}
C200 = result using all 200 segments
Med. = median for eight groups of 75 segments each

County	Stratum	Corn		Oats		Flax		Sunflower	
		C200	Med.	C200	Med.	C200	Med.	C200	Med.
Codington	11	.80	.59	.19	.07	.24	.18	0	.07
	12	.85	.61	.54	.47	0	0	.04	.24
Spink	11	.91	.79	.38	.22	.45	.38	0	.14
	12	.91	.74	.67	.42	0	0	.07	.32
Beadle	11	.86	.68	.28	.18	.34	.31	0	.13
	12	.92	.77	.68	.59	0	0	.08	.38
Clark	11	.88	.72	.31	.22	.37	.25	0	.09
	23	.85	.62	.54	.52	0	0	.04	.24
Kingsbury	11	.77	.52	.17	.09	.22	.15	0	.05
	12	.90	.67	.64	.52	0	0	.06	.27
Hamlin	11	.83	.63	.23	.13	.28	.20	0	.05
	12	.77	.42	.40	.27	0	0	.02	.10

An initial assesment of the Battese-Fuller estimates was made by calculating relative root mean square errors. It is desirable to have these below 20%. Part 1 of Table 3 shows that corn estimates satisfy this requirement with few exceptions when we assume $\sigma_{v11,12} = 0$. Part 2 of Table 3 indicates that these relative root mean square errors go up a few percentage points when the cross strata correlation is used.

Table 3 - Part 1: Relative Root Mean Square Error
Assuming Zero Cross Strata Correlation
(Relative RMSE = (RMSE/Estimate) · 100%)
Abbreviations are as defined in section IV C.

Crop	County	using 200 segments			BFGAM 8 groups Median
		BFREG	BFONE	BFGAM	
Corn	Codington	27	20	17	18
	Spink	77	12	12	19
	Beadle	81	12	12	18.5
	Clark	32	24	21	19
	Kingsbury	21	8	7	9.5
	Hamlin	15	10	9	11.5
Oats	Codington	29	20	15	15
	Spink	43	36	66	41
	Beadle	60	25	198	33.5
	Clark	23	33	20	21
	Kingsbury	28	48	17	29.5
	Hamlin	15	17	11	17.5
Flax	Codington	15	21	21	15
	Spink	107	53	6	51
	Beadle	197	237	61	76
	Clark	22	22	16	19
	Kingsbury	16	21	308	15
	Hamlin	12	14	106	14
Sunflower	Codington	21	60	21	33
	Spink	6	10	6	13
	Beadle	64	76	61	75
	Clark	17	23	16	21
	Kingsbury	388	147	308	126
	Hamlin	106	210	106	91

Table 3 - Part 2: Relative Root Mean Square Error
Using an Estimated Nonzero
Cross Strata Correlation

Crop	County	using 200 segments			BFGAM 8 groups Median
		BFREG	BFONE	BFGAM	
Corn	Codington	35	20	18	19
	Spink	99	12	12	24
	Beadle	101	12	12	18
	Clark	40	24	22	20
	Kingsbury	24	8	7	10
	Hamlin	19	10	9	13

For oats and flax the comparison values are poor with regard to relative root mean square error. Nonetheless, the Battese-Fuller estimation procedure using $C^i = \Gamma^i$ gave acceptable results across the eight groups for half the county oat estimates and four of the six county flax estimates. The most concentrated crop, sunflowers, is well estimated only in the one county that accounts for the bulk of the production.

Because corn presented the best relative RMSE's using the Battese-Fuller formulas as well as the best comparison values some further study was done with this crop. RMSE's found from the Battese-Fuller formulas were compared against an interval estimate of the RMSE based on the 8 estimates obtained from the groups of 75 segments each. This empirical RMSE was calculated by taking the square root of the observed variance of the 8 estimates and adding the following interval estimate of the squared bias:

$$\left[\begin{array}{c}\text{average}\\ \text{of 8}\\ \text{estimates}\end{array} - \left(\begin{array}{c}\text{comparison}\\ \text{value}\end{array} + \begin{array}{c}\text{standard}\\ \text{deveation of}\\ \text{comparison}\\ \text{value}\end{array}\right)\right] \quad 2.$$

Using the estimated RMSE from column 5 together with the observed variance of the 8 estimates, the portion of MSE which is not attributable to bias was calculated.

Although it is difficult to determine the bias, these calculations indicate that:

1. bias is not a negligible portion of the RMSE for any of the estimators considered.
2. for 5 of the 6 counties, the Huddleston-Ray and the attese-Fuller estimate which uses C=0 both contain substantially more bias than do the Battese-Fuller estimates which use C= Γ and C=I.

Furthermore, it was discovered, that the closest agreement between formula based RMSE's and empirically estimated ones occurred for the Battese-Fuller estimate which uses C=I. For this estimate only one county displayed an empirical RMSE that was larger than the median of the 8 formula values. This happened for 4 counties using C= Γ and for 5 counties using C=0. Thus, the formula RMSE's for the optimal Battese-Fuller estimate appear to underestimate the actual RMSE.

An absolute average relative bias was calculated according to the formula:

$$\frac{\text{average of the 8 estimates-comparison value}}{\text{comparison value}} \cdot 100\%.$$

A plot of the results for corn showed that the larger relative biases were associated either with the regression line estimators or with the two smallest producing counties . This pattern was less pronounced for oats but the comparison values used for this crop have larger standard deviations. For flax and sunflowers

the only acceptably small biases occur in the largest of the producing counties. These results are, perhaps, accounted for by the large coefficients of variation for the comparison values.

Consider finally the importance of the cross strata portion of the correlation for the residuals. This was successfully estimated for corn using all 200 segments and using four of the eight smaller groups. To assess the percent change in the optimal estimates we calculated:

$$\frac{\text{estimate using } C = \bar{A}^{1}H - \text{estimate using } C = \Gamma}{\text{estimate using } C = \Gamma} \cdot 100\%$$

and similarly for the root mean square error. Most of these quantities fell between 2 and 6%.

All of the results described in this section appear in greater detail in Walker and Sigman (1982).

VI. CONCLUSIONS

The analysis done thus far on the six county region in South Dakota supports the following conclusions:

1. Models without strata-specific parameter values do not appear to be correct.
2. The assumption of homoscedatic errors across counties within each stratum and county does not appear to be valid.
3. Heterogeneity of regression slopes across counties may be explained by low values of r^2 (coefficient of determination between classification results and ground truth). Large r^2 values appear to indicate near homogeneity of these slopes.
4. The presence of a nonzero county effect appears to be both crop and strata specific. It may be an increasing function of crop proportion.
5. RMSE's calculated according to the Battese-Fuller model were smallest for the coefficient matrices $C = \Gamma$ and $C = A^{-1}H$ as predicted by the theory.
6. The optimal Battese-Fuller estimate gives relative RMSE's (from the equations of Setion IV) below the desired 20% level for corn and in certain counties also for oats, flax and sunflowers. Thus, for this study, low relative RMSE's were associated with the largest crop proportion and the strongest county effect.
7. Empirically estimated RMSE's for corn are larger than formula derived values; the discrepancy being greatest for $C = 0$ and least for $C = I$.
8. A major portion of the empirical RMSE (for corn) is attributable to bias but, as predicted by the theory, bias is less when using $C = \Gamma$ or $C = I$ than when using $C = 0$.
9. Bias appears to be a decreasing function of crop proportion.
10. Battese-Fuller interval estimation based on the choice of $C = I$ fit the comparison values better than those using $C = 0$ and $C = \Gamma$.
11. The cross strata correlation of residuals appears to be weaker than that within strata.
12. Ignoring the cross strata correlation gives an optimal estimate whose RMSE is underestimated in most cases by 2-6%.

REFERENCES

Battese, G.E., W.A. Fuller. 1981. Prediction of County Crop Areas Using Survey and Satellite Data. Survey Section Proceedings, 1981 American Statistical Association Annual Meeting, Detroit, Michigan.

Cardenas, M., M. Blanchard, M.Craig. 1978. On the Development of Small Area Estimators Using LANDSAT Data as Auxiliary Information. USDA, ESCS, Washington, D.C.

Cochran, W.G. 1977. Sampling Techniques, Third edition. John Wiley and Sons, Inc., New York.

Ford, B.L. 1981. The Development of County Estimates in North Carolina, USDA, SRS, Washington, D.C.

Hanuschak, G., R. Allen, W. Wigton. 1982. Integration of LANDSAT Data into the Crop Estimation Program of USDA's Statistical Reporting Service: 1972-1982. Proceedings, 1982 Machine Processing of Remotely Sensed Data Symposium, West Lafayette, Indiana.

Huddleston, H.F., R. Ray, 1976. A New Approach To Small Area Crop-Acreage Estimation. Annual Meeting of the American Agricultural Economics Association, State College, Pennsylvania.

McLaughlin, D.M. 1975. A Test for Homogeneity of Regression Without Homogeneity of Variance. Educational and Psychological Measurement volume 35, pp 79-86.

Sigman, R., G. Hanuschak, M. Craig, P. Cook, M. Cardenas. 1978. The Use of Regression Estimation with LANDSAT and Probability Ground Sample Data. Survey Section Proceedings, 1978 American Statistical Association Annual Meeting, San Diego, California.

Walker, G., R. Sigman. 1982. The Use of LANDSAT for County Estimates of Crop Areas: Evaluation of the Huddleston-Ray and the Battese-Fuller Estimators, USDA, SRS, Washington, D.C.