United States
Department of
Agriculture

Statistical
Reporting
Service

Research
Division

September 1982

The Use of LANDSAT for County Estimates of Crop Areas

Evaluation of the Huddleston-Ray and Battese-Fuller Estimators

Gail Walker
Richard Sigman

The Use of LANDSAT for County Estimates of Crop Areas: Evaluation of the Huddleston-Ray and the Battese-Fuller Estimators

by

Gail Walker
Mathematics Department
Colby College
Waterville, Maine

and

Richard Sigman
Statistical Reporting Service
U.S. Department of Agriculture
Washington, DC

This paper was prepared for presentation at the 1982 joint statistical meetings of the American Statistical Association. The first author is a Fellow under the Statistical Survey Institute Program of the American Statistical Association.

Use of LANDSAT for County Estimates of Crop Areas: Evaluation of the Huddleston-Ray and Battese-Fuller Estimators. Gail Walker and Richard Sigman, Research Division, Statistical Reporting Service, U.S. Department of Agriculture, Washington D.C. 20250, SRS Staff Report No. AGES 820909, September 1982.

ABSTRACT

The purpose of this report is to develop and compare estimators which use LANDSAT data to estimate crop areas at the county level. This report extends the Battese-Fuller estimator to a stratified sample design and evaluates the Huddleston-Ray estimator and variations of the Battese-Fuller estimator on a six-county area in South Dakota. For SRS LANDSAT studies, the authors recommend replacing the Huddleston-Ray estimator with one of the favorably evaluated estimators in the Battese-Fuller family.

```
*****************************************************
*  This paper was reproduced for limited distribution to the     *
*  research community outside the U.S. Department of Agricul-*
*  ture.                                                                                    *
*****************************************************
```

TABLE OF CONTENTS

SUMMARY

This report addresses the problem of using LANDSAT data to obtain crop estimates at the county level.

First, the theory behind the presently employed Huddleston-Ray estimator and the family of estimators proposed in 1981 by Battese and Fuller is outlined. A necessary extension of the Battese-Fuller estimation model to a stratified sample design is then developed. Both types of estimators are studied over a six county region in eastern South Dakota.

There was a modest lack of fit of the Battese-Fuller model for the study data set, with larger model departure corresponding to low correlation between LANDSAT classification results and ground survey observations. A key feature of the Battese-Fuller model is a county effect parameter and this effect was found to be highly significant for corn, the largest of the four crops considered in the study. Furthermore, this effect manifested itself within several strata but was negligible across strata. The empirical work done for this study nonetheless indicates robustness of the Battese-Fuller estimators against departure from certain model assumptions.

Two members of the Battese-Fuller family of estimators satisfied the criterion for small relative root mean square error; that is, the percentage of the estimate attributable to root mean square error was less than 20%. These are the estimators which, under the assumed model, minimize mean square error and bias respectively. On the other hand, the Battese-Fuller estimate closest to the Huddleston-Ray estimate was far less satisfactory, failing to meet the desired ceilings for bias and mean square error. Therefore, for SRS LANDSAT studies, the authors recommend replacing the Huddleston-Ray estimator with one of the favorably evaluated estimators in the Battese-Fuller family.

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. Segment sizes are typically one square mile. In the JES there are two levels of stratification. The first-level strata are individual states. Secondary strata are areas of land within a state which have similar land use. Defined in terms of the percent of land under cultivation, these secondary strata are determined by visual interpretation of aerial photography and satellite imagery. Stratum definitions in the state of South Dakota, for example, are the following:

o Stratum 11: 75% + cultivated

o Stratum 12: 50% - 75% cultivated

o Stratum 20: 15% - 49% cultivated

o Strata 31, 32, 33: urban and residential

o Stratum 40: rangeland

o Stratum 61: proposed water

o Stratum 62: water

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. Currently, a given point on the earth's surface is imaged once every eighteen days. Satellite passes which are adjacent on the earth's surface are at least one day apart with respect to their dates of imagery.

III. ANALYSIS-DISTRICT LANDSAT ESTIMATOR

Since 1972, SRS has been using LANDSAT data to improve crop-area estimates for unions of multi-county areas called analysis districts. These efforts have been research studies but since 1978 have provided timely end-of-year estimates to the SRS Crop Reporting Board. Hanuschak, Allen, and Wigton (1982) chronicle the 1972 to 1982 results from these studies.

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 areas 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:

$$\begin{aligned} REG_{a.d.} &= \bar{y}_{a.d.} + b_1 (\bar{X}_{a.d.} - \bar{x}_{a.d.}) \\ &= b_0 + b_1 \bar{X}_{a.d.} \end{aligned}$$

and the Huddleston-Ray county estimator is:

$$\begin{aligned} HR_c &= \bar{y}_{a.d.} + b_1 (\bar{X}_c - \bar{x}_{a.d.}) \\ &= b_0 + b_1 \bar{X}_c. \end{aligned}$$

The problems with this procedure are:

1. it is unclear how to calculate an accurate variance for the county estimate so obtained, and
2. the use of the difference

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

   lumps together a difference attributable to sampling error within the county and a difference that measures the inherent distinction between a given county and the analysis district.

B. Battese-Fuller Model.

In addressing the above issues 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 $\bar{u}_c = \bar{y}_c - b_0 - b_1\bar{x}_c$, the Battese-Fuller county estimator is:

$$BF_c = b_0 + b_1 \bar{X}_c + \delta_c \bar{u}_c \text{ where } 0 \leq \delta_c \leq 1.$$

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 jth sampled segment from the ith county we have:

$$y_{ij} = b_0 + b_1 x_{ij} + u_{ij} = b_0 + b_1 x_{ij} + v_i + e_{ij}$$

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_0 and b_1 are known, the county mean residuals

$$\bar{u}_{i.} = \bar{y}_{i.} - b_0 - b_1 \bar{x}_{i.} = v_i + e_{i.}$$

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. $\delta_i = \gamma_i$

- o minimum MSE is obtained
- o $\frac{MSCB}{MSE} = 1 - \gamma_i$

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. One then has:

$$
\begin{aligned}
MSE = \quad & (1-\delta_i)^2\sigma_v^2 + \delta_i^2 \frac{\sigma_e^2}{n_i} \\
& +2\,(\delta_i - \gamma_i)\ (\bar{X}_i - \delta_i \bar{x}_i)\, V(\hat{b})\, \bar{x}_i'. \\
& + (\bar{X}_i - \delta_i \bar{x}_i)\, V(\hat{b})\, (\bar{X}_i - \delta_i \bar{x}_i)'
\end{aligned}
$$

and

$$
\begin{aligned}
MSCB = {} & (1-\delta_i)^2\sigma_v^2 \\
& -2(1-\delta_i)\gamma_i\, (\bar{X}_i - \delta_i \bar{x}_i)\, V(\hat{b})\, \bar{x}_i' \\
& +(\bar{X}_i - \delta_i \bar{x}_i)\, V(\hat{b}) \sum_{j=1}^{t} \bar{x}_j \bar{x}_j \frac{\gamma_j}{\sigma_v^2} - V(\hat{b})\, (\bar{X}_i - \delta_i \bar{x}_i)'
\end{aligned}
$$

where $\bar{x}_i$ and $\bar{X}_i$ are vectors $(1, \bar{x}_i)$ and $(1, \bar{X}_i)$. The same choice of $\delta_i = \gamma_i$ minimizes the MSE.

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

$$\text{cov}(u_{hij}, u_{h'i'j'}) = \begin{cases} 0 & \text{if } i \neq i' \\ \sigma^2_{vh} & \text{if } i=i' \;\; h=h' \;\; j \neq j' \\ \sigma^2_{vh} + \sigma^2_{eh} & \text{if } i=i' \;\; h=h' \;\; j=j' \\ \sigma_{vhh'} & \text{if } i=i' \;\; h \neq h' \end{cases}$$

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;

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

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

$$BX^i = \begin{pmatrix} 1 & \bar{X}_1 & . & . & . & 0 & 0 \\ . & . & & & & . & . \\ . & . & & & & . & . \\ . & . & & & & . & . \\ . & . & & & & . & . \\ 0 & 0 & . & . & . & 1 & \bar{X}_{si} \end{pmatrix}$$

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

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

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

$$\hat{\underset{\sim}{\mu}}_i = BX^i \cdot B + C^{i'} D^i$$

and the final county mean is estimated by

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

Introducing the s by s matrices

$$H = E(v^i v^{i'}) = \begin{pmatrix} \sigma^2_{v_1} & \cdot & \cdot & \cdot & \sigma_{v_{1s}} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \sigma_{v_{1s}} & \cdot & \cdot & & \sigma^2_{v_s} \end{pmatrix}$$

and

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

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

Then $MSE\,(\mu^i) = w^i\,E(\,(v^i - C^{i'}u^i)\,(v^{i'} - u^{i'}C^i))\,w^{i'}$

$= w^i(H - 2HC^i + C^{i'}A^iC^i)\;w^{i'}$

and

$MSCB \;\; = w^i(H - 2HC^i + C^{i'}HC^i)\;w^{i'}$

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

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

which reduces to

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

if $v_{hk} = 0$ 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. $C^i = \Gamma^i = \begin{pmatrix} \gamma_{1i} & & & 0 \\ & \cdot & & \\ & & \cdot & \\ & & & \cdot \\ 0 & & & \gamma_{si} \end{pmatrix}$

- 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.

In order to display formulas for the mean square error and mean square conditional bias when b is estimated, we introduce the 2s by 2s matrices

$$VB = \begin{pmatrix} V(\hat{b}^1) & & & 0 \\ & \cdot & & \\ & & \cdot & \\ & & & \cdot \\ 0 & & & V(\hat{b}^s) \end{pmatrix}$$

$$VCB = E\,((\hat{B} - B)\,(\hat{B} - B)')$$

$$CS = \sum_{j=1}^{t} LX^{j'}(SV+SE^{j})^{-1}H\,(SV+SE^{j})^{-1}LX^{j} \text{ where } SV = \begin{pmatrix} \sigma^2_{v_1} & & & 0 \\ & \cdot & & \\ & & \cdot & \\ & & & \cdot \\ 0 & & & \sigma^2_{v_s} \end{pmatrix}$$

It then follows that the estimates

$$\hat{\mu}\underset{\sim}{i} = BX^{i}\hat{B} + C^{i'}D^{i}$$

and

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

give

$$\begin{aligned} MSE\,(\mu^{i}) = w^{i}\;E(\,& v^{i} - C^{i'}u^{i} - (BX^{i} - C^{i'}LX^{i})(\hat{B} - B) \\ & v^{i} - C^{i'}u^{i} - (BX^{i} - C^{i'}LX^{i})(\hat{B} - B)\;')\;w^{i'} \end{aligned}$$

$$\begin{aligned} = w^{i}\;\{ & H - 2HC^{i} + C^{i'}A^{i}C^{i} \\ & -2(BX^{i} - C^{i'}LX^{i})\,VB\,LX^{i}\;\;\Gamma^{i} - C^{i} + (\Gamma^{i} - (SV+SE^{i})^{-1}H)(C^{i} - I) \\ & +(BX^{i} - C^{i'}LX^{i})\,VCB\,(BX^{i} - C^{i}LX^{i})'\}\,w^{i'} \end{aligned}$$

and

$$\begin{aligned} MSCB\,(\mu^{i}) \;\; = w^{i}\;\{ & H-2HC^{i}+C^{i'}HC^{i} \\ & - 2\,(BX^{i}-C^{i'}LX^{i})\,VB\,LX^{i'}\;\;(SV+SE^{i})^{-1}H\,(I-C_{i}) \\ & + (BX^{i}-C^{i'}LX^{i})\,VB\,CS\,VB\,(BX^{i}-C^{i'}LX^{i})\,\}w^{i'}. \end{aligned}$$

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. A LANDSAT and ground-truth data set was available for this area as a result of a joint study by SRS and the Remote Sensing Institute (RSI), located in Brookings, South Dakota. The original SRS-RSI use of this data set was in determining the affect of soil type on LANDSAT data characteristics.

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.

The location of the data set is a six-county area in eastern South Dakota making up approximately 40 percent of a LANDSAT scene. Two of the counties have small fractions of their areas (4% and 7%, respectively) lying outside the LANDSAT scene.

The sample design of the original soil study consisted of ten strata defined in terms of soil characteristics. Sample selection was by proportional allocation with increased sampling in small strata. Ground data collection was performed by RSI and was by observation only. No interviews with farm operators were conducted.

For the county-estimator study, however, generalization of conclusions to the JES was desired. Consequently, for the county-estimation work some segments were randomly discarded to restore proportional allocation, and the segments then reassigned to the SRS land-use strata. The resulting sample size was 200 segments. As can be seen in Table 1, there was a sufficient number of segments to calculate within-county regression estimates for all counties in strata 11 and 12 and for most counties in strata 20.

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 from the 200 segments 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 (also called training groups) and for the full sample.

This data set was also used in another county estimation study comparing the Huddleston-Ray and the Cardenas estimators. (Amis, et. al., 1982) The present study uses the LANDSAT classification results from this earlier work.

The LANDSAT data used in the county-estimation study came from two image dates — July 20, 1979, and August 25, 1979. Thus, the MSS measurement vector was eight dimensional — four measurements from July 20 and four measurements from August 25. The early image date was before the 1979 start of harvesting for oats and flax.

Although the six-county analysis district estimates were not of interest in this study, they were calculated in order to compare results with other SRS LANDSAT studies. The stratum variance of an analysis district regression estimate is proportional to $1-R^2$, where R^2 is the coefficient of determination between the LANDSAT classification results and the ground truth.

For the full sample of 200 segments, the R^2 values were as follows:

	Stratum		
	11	12	20
Corn	.78	.76	.33
Sunflower	.92	.86	—
Flax	.66	.26	.46
Oats	.37	.23	.23

The relative efficiency of an analysis-district regression estimator is the variance of the direct-expansion estimator divided by the variance of the regression estimator. Using all 200 segments the analysis-district relative efficiencies were 3.9, 8.8, 1.7, and 1.1 for corn, sunflowers, oats and flax, respectively. For the eight samples of size 75 the analysis-district relative efficiencies for corn ranged from 3.2 to 7.9 with a median of 4.4.

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 and county i

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

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

If the <u>unstratified</u> 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 <u>stratified</u> 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. Each $\hat{b}^1_{hi}$ 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 example, VEE denotes homogeneity of residual variance and slopes, but heterogeneity of intercepts.

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. Additional hypotheses of interest are the unstratified models

EEE: Homogeneity of intercepts, residual variances, and slopes across both county and strata

VcVcE: Homogeneity of slopes across county and strata. Heterogeneity of intercepts and residual variances,

and the corresponding stratified models

EcEcEc: Homogeneity of intercepts, residual variances, and intercepts across counties within each strata. Heterogeneity across strata.

VVEc: Homogeneity of slopes across counties within each stratum, but heterogeneity of intercepts and residual variances.

Such models can be tested by the following general procedure (McLaughin, 1975):

1. Calculate maximum likelihood estimates for the triplet under both hypothesis VVV and the restricted hypothesis which is of interest.

2. Use the calculated estimates to evaluate the likelihood ratio, L, that has hypothesis VVV corresponding to the denominator and the restricted hypothesis corresponding to the numerator.

3. Reject the restricted hypothesis if $G^2 = -2 \text{ Log } (L)$ is large.

The critical region of the test is calculated from the asymptotic distribution of G^2 under the restricted hypothesis. This distribution is a chi-square distribution with degrees of freedom equal to the difference in number of parameters estimated under hypothesis VVV and the restricted hypothesis.

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.

The hypotheses of interest are listed in the third column of Table 2. For models EEE, V_cV_cE, $E_cE_cE_c$, and V_cEE, the required maximum likelihood estimates can be obtained from ordinary least squares procedures. For models V_cV_cE and VVE_c, convergence of iterated weighted least squares estimation provides the needed maximum likelihood estimates.

Table 2 lists 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. Figure 1 compares the variability of estimated regression slopes under models VVV and VVE_c as a function of R^2, the coefficient of determination between classification results and ground truth. Though the likelihood ratio tests reject VVE_c for all crops except corn, Figure 1 indicates 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. Figure 1 also shows that the heterogeneity of regression slopes is more likely for low R^2 values.

Models which assume the homogeneity of error variance across counties were readily rejected. Figure 2 compares the variability of estimated error variances under model VVV and VE_CV (homogeneous error variances across counties within each stratum; heterogeneous slopes and intercepts). 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.

Table 2: Model Tests

Crop	Type of Model	Model*	Test Statistic	Degrees of Freedom	P-value
Corn	Unstratified	EEE	123.5	33	≠
		V_CEE**	84.2	28	≠
		V_CV_CE	41.4	23	.01
	Stratified	$E_CE_CE_C$	115.4	30	≠
		VE_CE_C***	50.4	20	≠
		VVE_C	13.2	10	.21
Sunflowers	Unstratified	EEE	125.0	18	≠
		V_CEE**	124.9	15	≠
		V_CV_CE	84.3	12	≠
	Stratified	$E_CE_CE_C$	112.2	15	≠
		VE_CE_C***	108.2	10	≠
		VVE_C	48.0	5	≠
Flax	Unstratified	EEE	160.5	24	≠
		V_CEE**	159.6	20	≠
		V_CV_CE	90.0	16	≠
	Stratified	$E_CE_CE_C$	142.4	21	≠
		VE_CE_C***	133.4	16	≠
		VVE_C	37.8	8	≠
Oats	Unstratified	EEE	143.0	21	≠
		V_CEE**	130.1	16	≠
		V_CV_CE	83.1	11	≠
	Stratified	$E_CE_CE_C$	140.0	18	≠
		$V_CE_CE_C$***	120.8	12	≠
		VVE_C	37.2	6	≠

* Model notation explained on pages 19 through 21.

** Unstratified Battese-Fuller assumptions.

***Stratified Battese-Fuller assumptions.

≠ $p^{<}$ 0.01

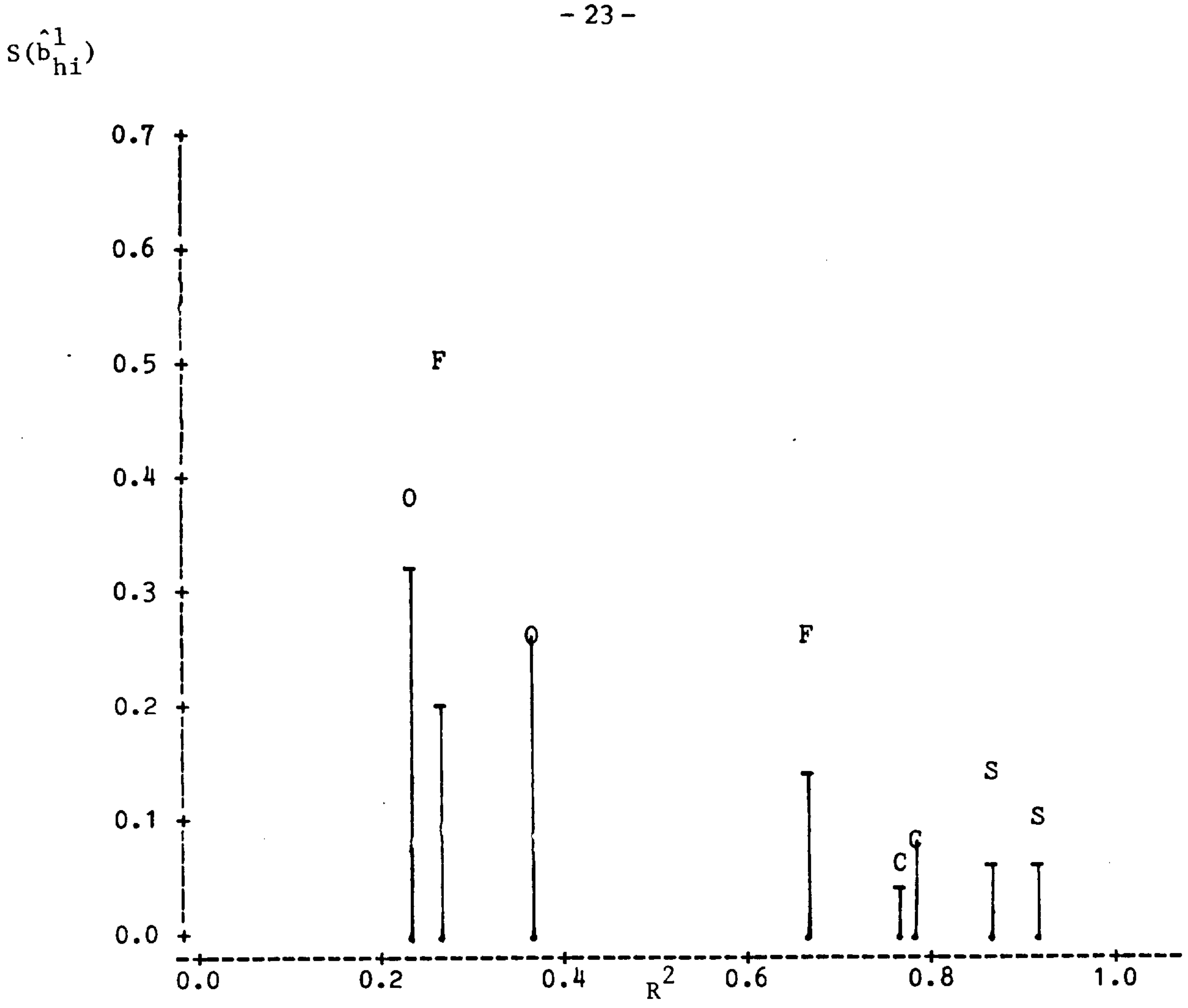


Figure 1. Variable of estimated within-county regression coefficients versus R^2. Variability measure is standard deviation (over counties) of estimated within-county stratum regression coefficients. R^2 = coefficient of determination between classification results and ground truth. Strata 11 and 12. C = corn, F = flax, O = oats, and S = sunflowers indicate observed variability (model VVV). Vertical lines indicate estimated of expected variability under model VVE_c (homogenous slopes across counties within each stratum; heterogeneous intercepts and residual variances).

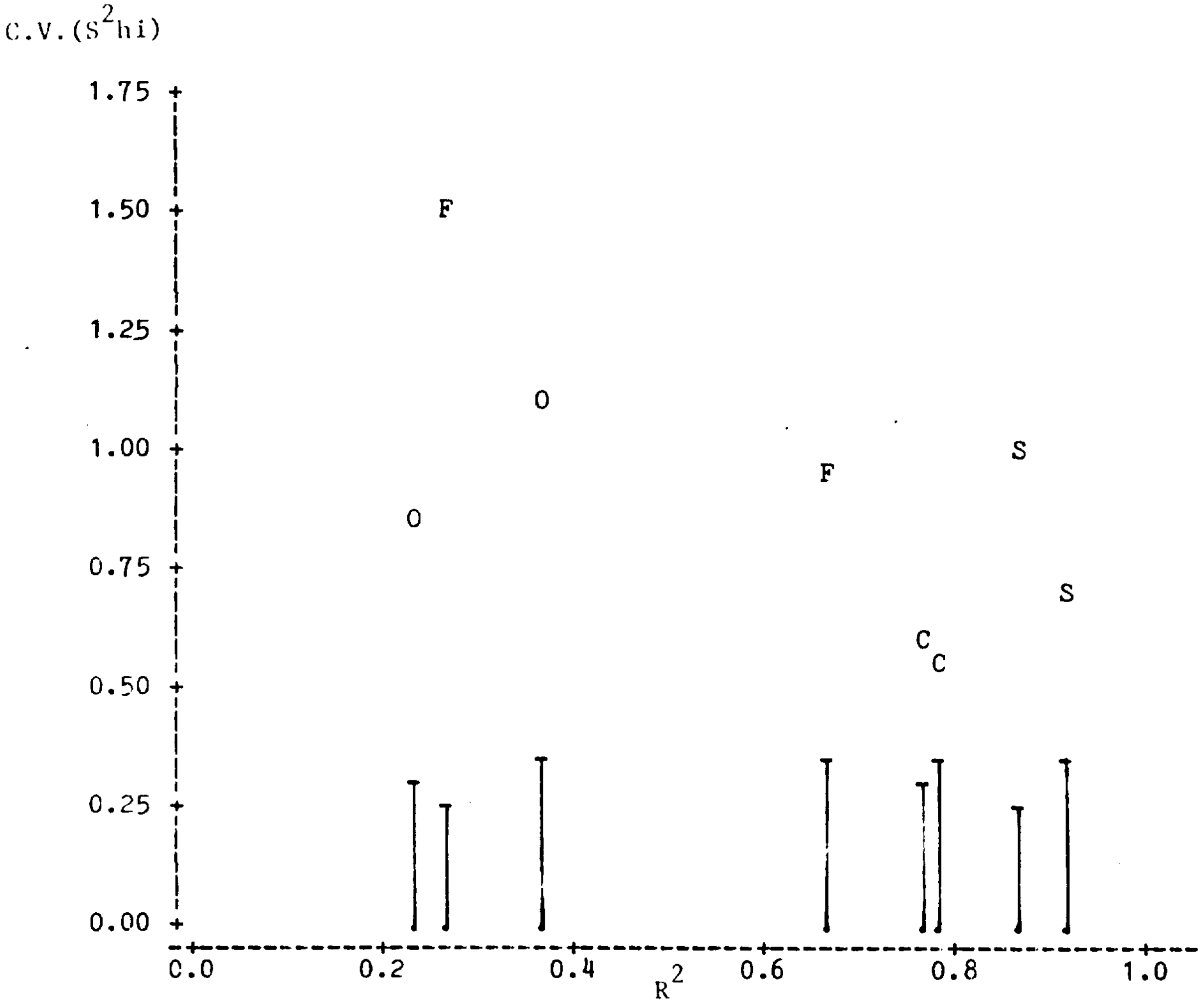


Figure 2. Variability of estimated within-county error variances versus R^2. Variability measure is coefficient of variation (over counties) of estimated within-county stratum error variances. R^2 = coefficient of determination between classification results and ground truth. Strata 11 and 12. C = corn, F = flax, O = oats, and S = sunflowers indicate observed variability (model VVV). Vertical lines indicate estimate of expected variability if error-variance parameters were homogeneous across counties within each stratum.

C. Results

1.) Comparison Values

Table 3 shows the crop estimates (in hectares) obtained for purposes of comparison by running six individual county regressions. Coefficients of variation are parenthesized to the right of the estimate and each county's percentage of total output appears in parentheses below the estimate.

Corn and oats are notably more abundant than flax and sunflowers. The distribution of corn shows three major producers (19-25%) and three minor ones (9-15%). Oats are quite evenly distributed (17-21%) among five counties with the sixth county a minor producer (8%). For flax there are only five producing counties two of which are notably larger (24-26%) than the other three (14-18%). Sunflowers present the most concentrated distribution with one county claiming three quarters of the total production.

A successful county estimator should, of course, perform well for both largeand small production whether evenly distributed or not.

Table 3: Crop estimates obtained from individual county regression and used as basis for comparisons

County	Corn (C.V.*)	Oats (C.V.)	Flax (C.V.)	Sunflowers (C.V.)
Codington (proportion of total)	16296 (16%) (9%)	19467 (32.3) (19)	11698 (42.4) (26)	3428 (62.7) (9%)
Spink (proportion of total)	40527 (10.8) (21)	17320 (22.4) (17)	6080 (36.5) (14)	30309 (8.9) (75)
Beadle (proportion of total)	36499 (11.4) (19)	19244 (32.8) (18)	0	0
Clark (proportion of total)	20099 (21.6) (10)	18268 (28.6) (17)	7704 (28.9) (17)	6095 (75.5) (15)
Kingsburg (proportion of total)	48568 (9.6) (25)	8006 (34) (8)	10756 (23.3) (24)	378 (157.4) (1)
Hamlin (proportion of total)	29517 (6.6) (15)	21605 (25.2) (21)	8109 (33.1) (18)	0
TOTAL	191506	103910	44347	40210
Analysis District estimate using 200 segments	189900 (6.2)	111323 (9.9)	45175 (23.2)	43517 (8.7)

*C.V. = coefficient of variation = $\frac{\text{standard deviation}}{\text{estimate}}$

2.) Parameter estimation

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.

Table 4 gives the results of the F test for a nonzero county effect in strata 11 and 12. The most convincing evidence of this effect is found for corn in both strata and for oats in stratum 12.

Table 4: Results from testing for a nonzero county effect ($H_0: \sigma_v^2 = 0$).

The p-value is the probability, assuming H_0, of obtaining a σ_v^2 estimate as large as the one actually observed.

* indicates the result for the set of all 200 segments.

Numerical entries are the numbers of groups of 75 segments each that had the indicated result.

	p value					
Crop	Strata	<.01	.01-.05	.05-.10	.10-.44	.44 and $\hat{\sigma}_{v_h}^2 = 0$
Corn	11	*1	4	1	2	0
	12	*2	3	2	1	0
Oats	11	0	0	0	*4	4
	12	0	*2	1	4	1
Flax	11	0	1	0	*4	3
	12	0	0	1	2	*3
Sunflower	11	0	0	1	5	*2
	12	1	0	1	3	*3

Correlations of residuals within strata were found from the variance components according to the formula

$$\rho_u^h = \frac{\sigma^2_{v_h}}{\sigma^2_{v_h} + \sigma^2_{e_h}}$$

Table 5 records these results for the full set of 200 segments together with the minimum, median and maximum for the eight groups of 75 segments each. Except for corn, low correlations resulted because $\sigma^2_{v_h}$ was small relative to $\sigma^2_{e_h}$.

The cross strata correlation was estimated from

$$\hat\rho_u^{11,12} = \frac{\hat\sigma_{v_{11,12}}}{\left[\left(\hat\sigma_{v_{11}} + \hat\sigma_{e_{11}}\right)\left(\hat\sigma_{v_{12}} + \hat\sigma_{e_{12}}\right)\right]^{0.5}}$$

where $\hat\sigma_{v_{11,12}} = \frac{1}{t-1} \sum_{i=1}^{t} \bar u_{11i}\, \bar u_{12i}$. These results also appear in Table 5.

Table 5: Estimated Correlations of Residuals (ρ)

Main entries are based on 200 segments. Parenthesized entries are the minimum, median and maximum from the eight groups of 75 segments each.

Type of Correlation	C	R	O	P
	Corn	Oats	Flax	Sunflowers
Within stratum 11	.32 (.08 .36 .63)	.029 (0 .05 .18)	.038 (0 .07 .41)	0 (0 .02 .33)
Within stratum 12	.30 (.02 .23 .55)	.077 (0 .12 .25)	0 (0 .02 .07)	.003 (0 .05 .42)
Across strata 11 and 12 (based on 4 groups)	.21 (.05 .21 .25)	.017	-.056	-.019

It seems 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 6. We know that $\gamma_{hi} \to 1$ as $n_{hi} \to \infty$ or $\sigma^2_{e_h} / \sigma^2_{v_h} \to 0$ and Table 6 shows that we were able to make a sizeable adjustment away from the regression line when estimating corn. The next largest γ_{hi} values occur in stratum 12 for oats and stratum 11 for flax. Note that flax and sunflowers usually require the use of a regression line estimate in at least one stratum.

Table 6: Optimal Scale Factors

$$\gamma_{hi} = \frac{\sigma^2_{v_h}}{\sigma^2_{v_h} + \frac{\sigma^2_{e_h}}{\bar{n}_{hi}}}$$

C200 = result using all 200 segments

Med. = median for eight groups of 75 segments each

		Corn		Oats		Flax		Sunflower	
County	Stratum	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

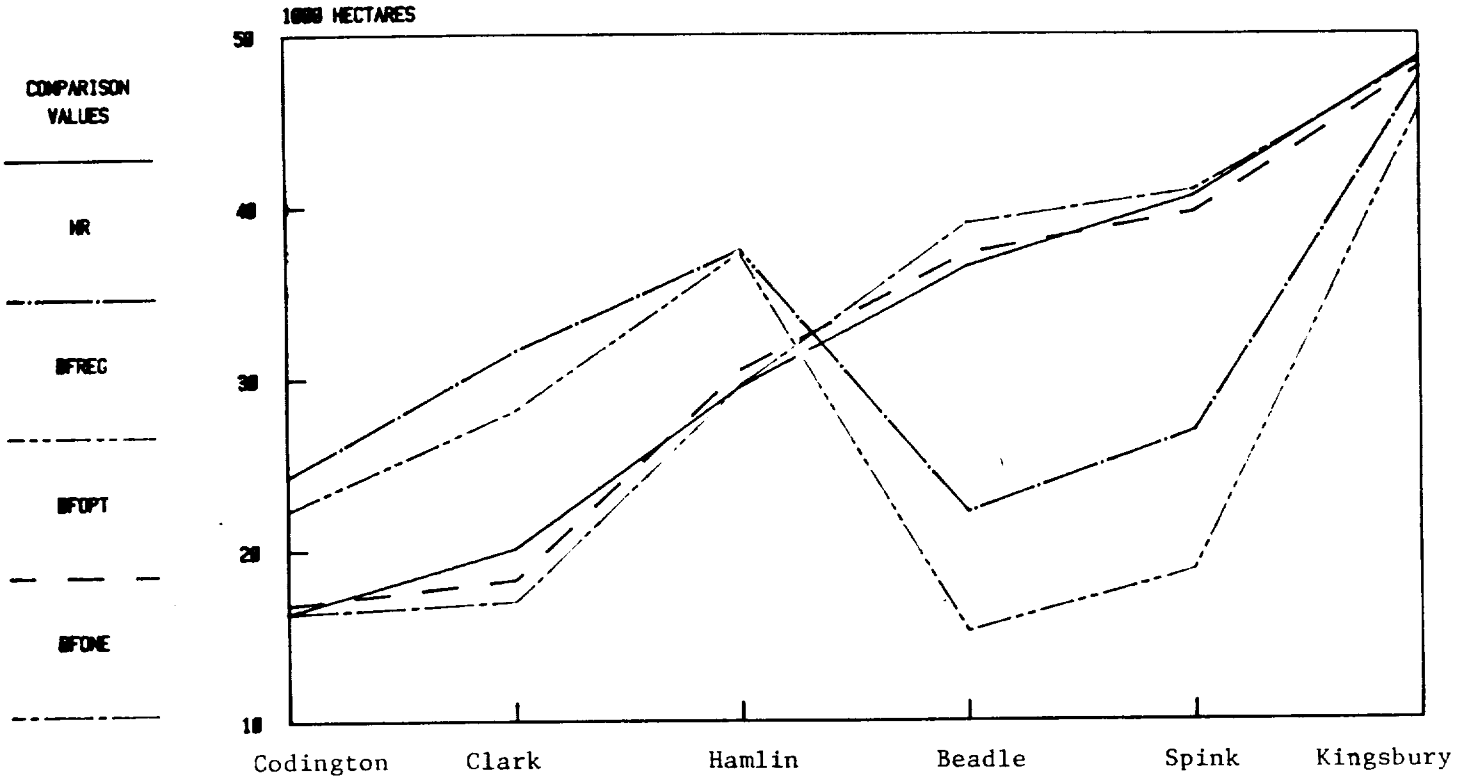
3. Estimates of county crop totals

Using the full set of 200 segments, figures 3 and 4 illustrate the three Battese-Fuller estimates against the comparison values given in table 3 and the Huddleston-Ray estimates for corn and oats respectively. (The scale is in thousands of hectares.)

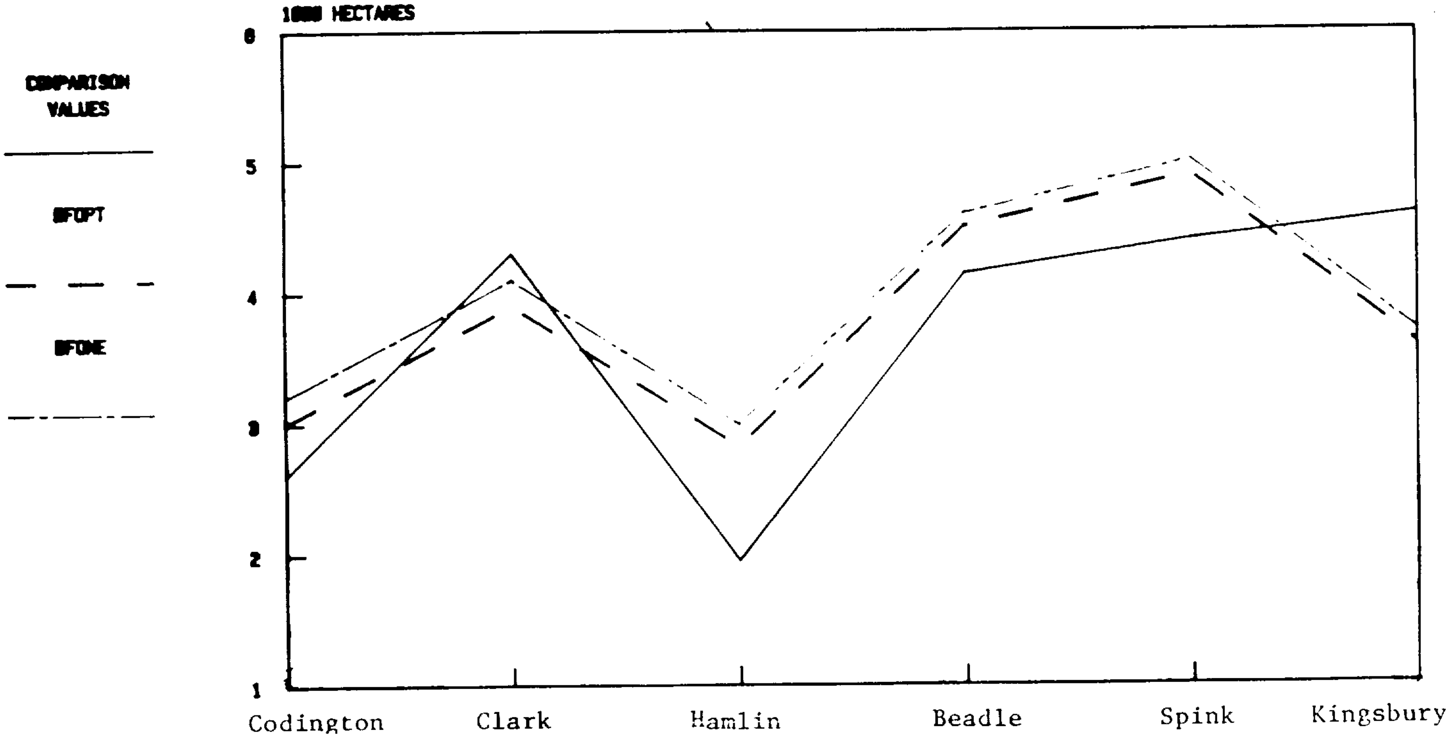
An initial assesment of the Battese-Fuller estimates was made by calculating relative root mean square errors. It is desirable to have these below 20%. Part 1 of Table 7 shows that corn estimates satisfy this requirement with few exceptions when we assume $\sigma_{v_{11,12}} = 0$. Part 2 of Table 7 indicates that these relative root mean square errors go up a few percentage points when the cross strata correlation is used.

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.

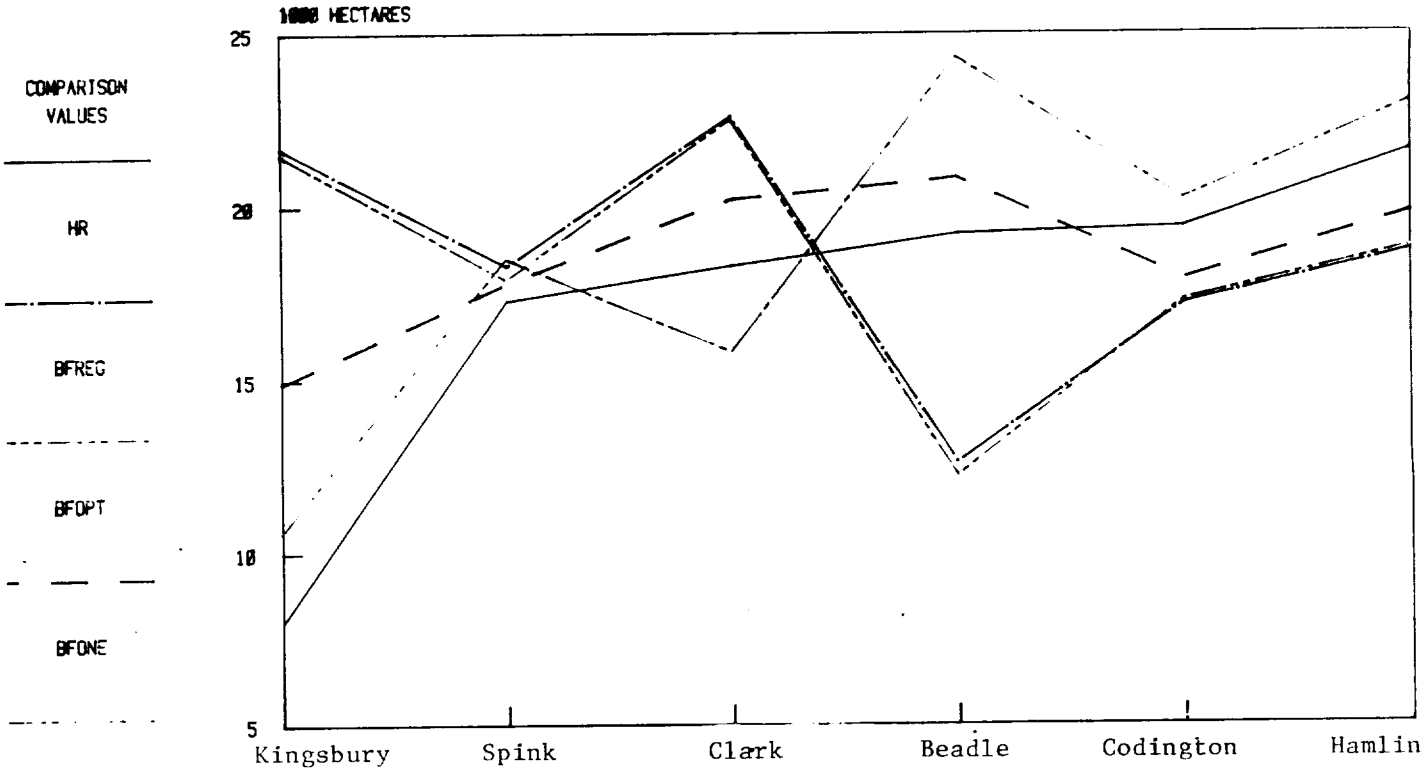
CORN ESTIMATES USING C2ØØ



RMSE FOR CORN ESTIMATES



OAT ESTIMATES USING C200



RMSE FOR OAT ESTIMATES

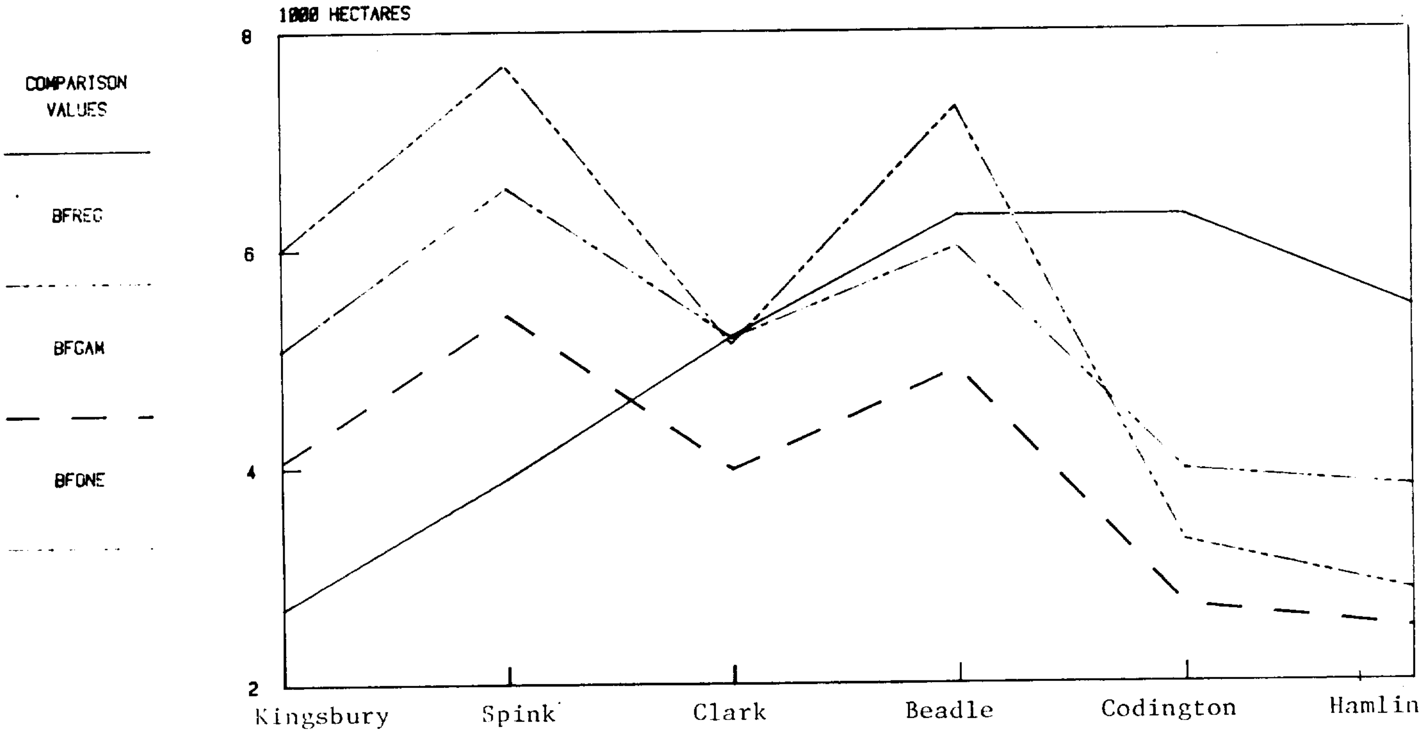


Table 7 - Part 1: Relative Root Mean Square Error
Assuming Zero Cross Strata Correlation

$$\text{Relative Root Mean Square Error} = \frac{\text{Root Mean Square Error}}{\text{Estimate}}$$

Abbreviations for types of estimates are as defined on page 14.

Crop	County	Comparison value as in table 3	BFREG using 200 segments	BFGAM using 200 segments	BFGAM 8 groups			BFONE using 200 segments
					Minimum	Median	Maximum	
Corn	Codington	16	27	17	12	18	22	20
	Spink	11	77	12	14	19	30	12
	Beadle	11	81	12	16	18.5	23	12
	Clark	22	32	21	12	19	43	24
	Kingsbury	10	21	7	8	9.5	10	8
	Hamlin	7	15	9	8	11.5	12	10
Oats	Codington	33	29	15	8	15	20	20
	Spink	22	43	66	20	41	68	36
	Beadle	33	60	198	28	33.5	62	25
	Clark	29	23	20	11	21	23	33
	Kingsbury	34	28	17	16	29.5	36	48
	Hamlin	25	15	11	12	17.5	32	17
Flax	Codington	42	15	21	8	15	28	21
	Spink	37	107	6	28	51	260	53
	Beadle	—	197	61	57	76	130	237
	Clark	29	22	16	10	19	26	22
	Kingsbury	23	16	308	9	15	25	21
	Hamlin	33	12	106	9	14	20	14
Sunflower	Codington	63	21	21	16	33	168	60
	Spink	9	6	6	5	13	26	10
	Beadle	—	64	61	33	75	175	76
	Clark	76	17	16	13	21	34	23
	Kingsbury	157	388	308	49	126	332	147
	Hamlin	—	106	106	41	91	483	210

Table 7 - Part 2: Relative Root Mean Square Error
Using an Estimated Nonzero Cross Strata Correlation

Crop	County	Comparison as in table 3	BFREG using 200 segments	BFGAM using 200 segments	BFGAM 8 groups			BFONE using 200 segments
					Minimum	Median	Maximum	
Corn	Codington	16	35	18	13	19	21	20
	Spink	11	99	12	16	24	29	12
	Beadle	11	101	12	15	18	23	12
	Clark	22	40	22	12	20	46	24
	Kingsbury	10	24	7	9	10	11	8
	Hamlin	7	19	9	11	13	14	10

Corn presented the best relative RMSE's using the Battese-Fuller formulas (table 7 - part 1) as well as the best comparison values (least coefficients of variation-table 3) and table 8 summarizes some further study done with this crop. The first four columns display RMSE's found from the Battese-Fuller formulas. (see p.15) The fifth column contains 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:

$$\left[\begin{pmatrix}\text{average of}\\ \text{8 estimates}\end{pmatrix} - \begin{pmatrix}\text{comparison} & + & \text{standard deviation of}\\ \text{value} & - & \text{comparison value}\end{pmatrix}\right]^{2}.$$

It should be noted that the 8 groups posess different county sample sizes but that the RMSE formulas of section IV refer to fixed values. Using the estimated RMSE from column 5 together with the observed variance of the 8 estimates, the portion of MSE which is <u>not</u> attributable to bias was calculated and recorded in column 6.

Although it is difficult to determine the bias, table 8 indicates 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 Battese-Fuller estimate which uses C=0 both contain substantially more bias than do the Battese-Fuller estimates which use C= Γ and C=I.

Table 8 also indicates that the closest agreement between formula based RMSE's and empirically estimated ones occurs 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.

Table 8: Formula and Empirical Root Mean Square Errors (RMSE)

The formula for the root mean square error appears on page 15.

The empirical root mean square error is computed according to the discussion on page 36.

CORN

County	Estimate (see page 14 for abbreviation key)	Formula RMSE using 200 segments	Formula RMSE 8 groups			Empirical estimate of RMSE	Empirical var. / Empirical MSE (%)
			Minimum	Median	Maximum		
Codington	HR					8909-14047	2-5
	BFREG	6032	3661	5672	8414	7569-12687	2-7
	BFGAM	2996	2814	3976	5692	5863-10697	8-26
	BFONE	3215	3768	5297	6350	6021-10081	23-63
Spink	HR					9985-18143	1-3
	BFREG	14460	8851	12641	18225	16243-24546	7-16
	BFGAM	4829	5922	6761	7774	6890-11850	33-98
	BFONE	5017	6915	8005	10443	7920-8327	67-74
Beadle	HR					9028-17231	1-4
	BFREG	12275	6946	10455	17126	14443-22566	3-6
	BFGAM	4436	4701	6400	6733	4098-8482	22-96
	BFONE	4621	6737	9154	7861	5341-8203	38-89
Clark	HR					7542-16052	2-8
	BFREG	8894	5512	8138	12423	9775-18345	1-5
	BFGAM	3874	3854	5405	7980	5449-12665	12-66
	BFONE	4076	5472	6357	8755	3727-9116	17-99
Kingsbury	HR					3418-4278	35-55
	BFREG	9444	4368	7732	13861	4613-6218	20-36
	BFGAM	3495	3883	4768	5589	4531-8298	24-80
	BFONE	3693	5169	5706	6750	4985-8281	29-79
Hamlin	HR					8487-12271	4-9
	BFREG	5740	3599	5393	7036	8443-12204	5-11
	BFGAM	2741	3212	3684	4150	4751-7644	27-70
	BFONE	2993	3680	5105	6309	4184-4694	71-89

Table 9 contains the results of calculating an absolute average relative bias 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 shows that the larger relative biases are associated either with the regression line estimators or with the two smallest producing counties (see figure 5). This pattern is less pronounced for oats (see figure 6) 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.

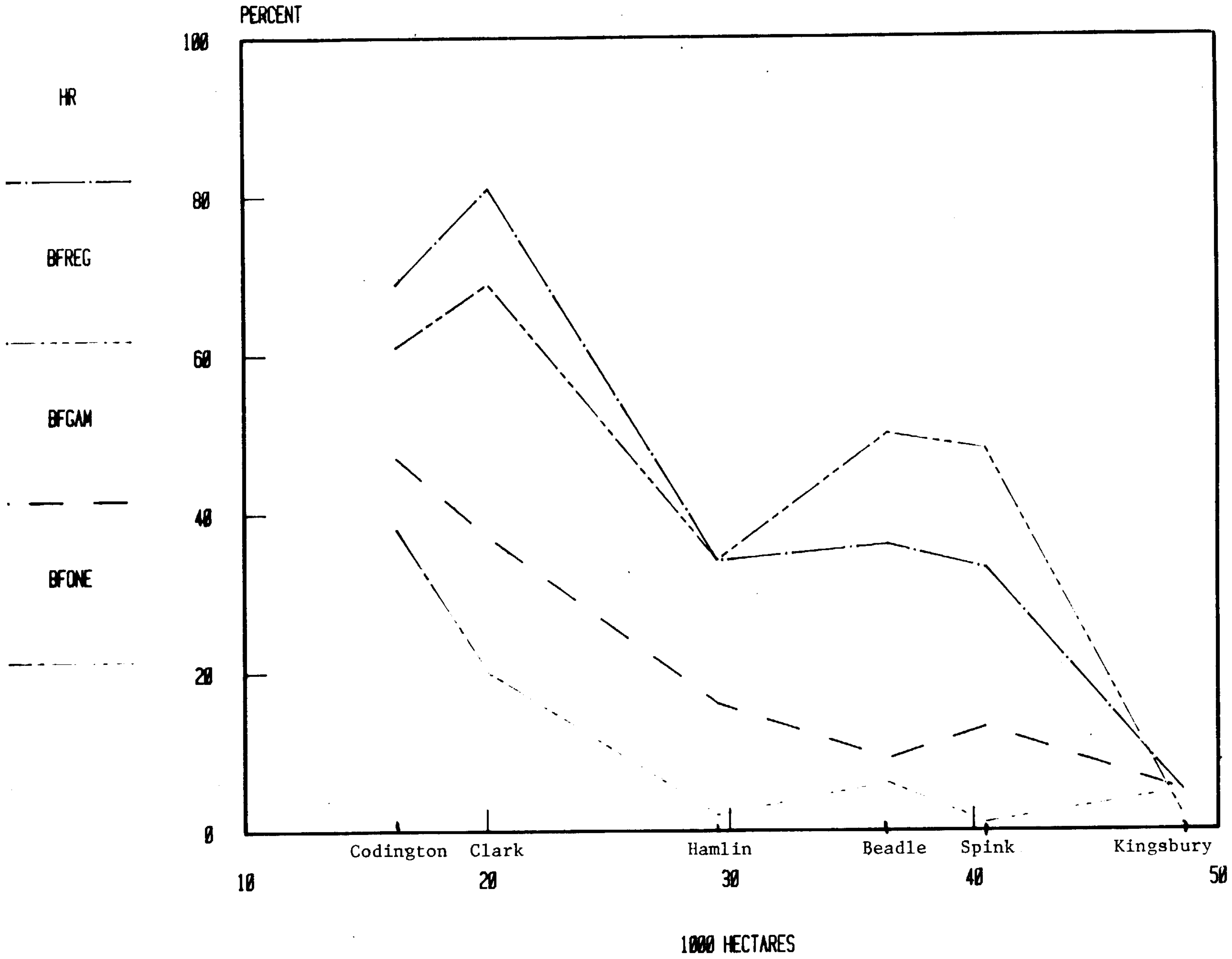
Table 9: Absolute Average Relative Bias

$$\frac{\text{average of 8 estimates - comparison value from table 3}}{\text{comparison value}}\ 100\,\%$$

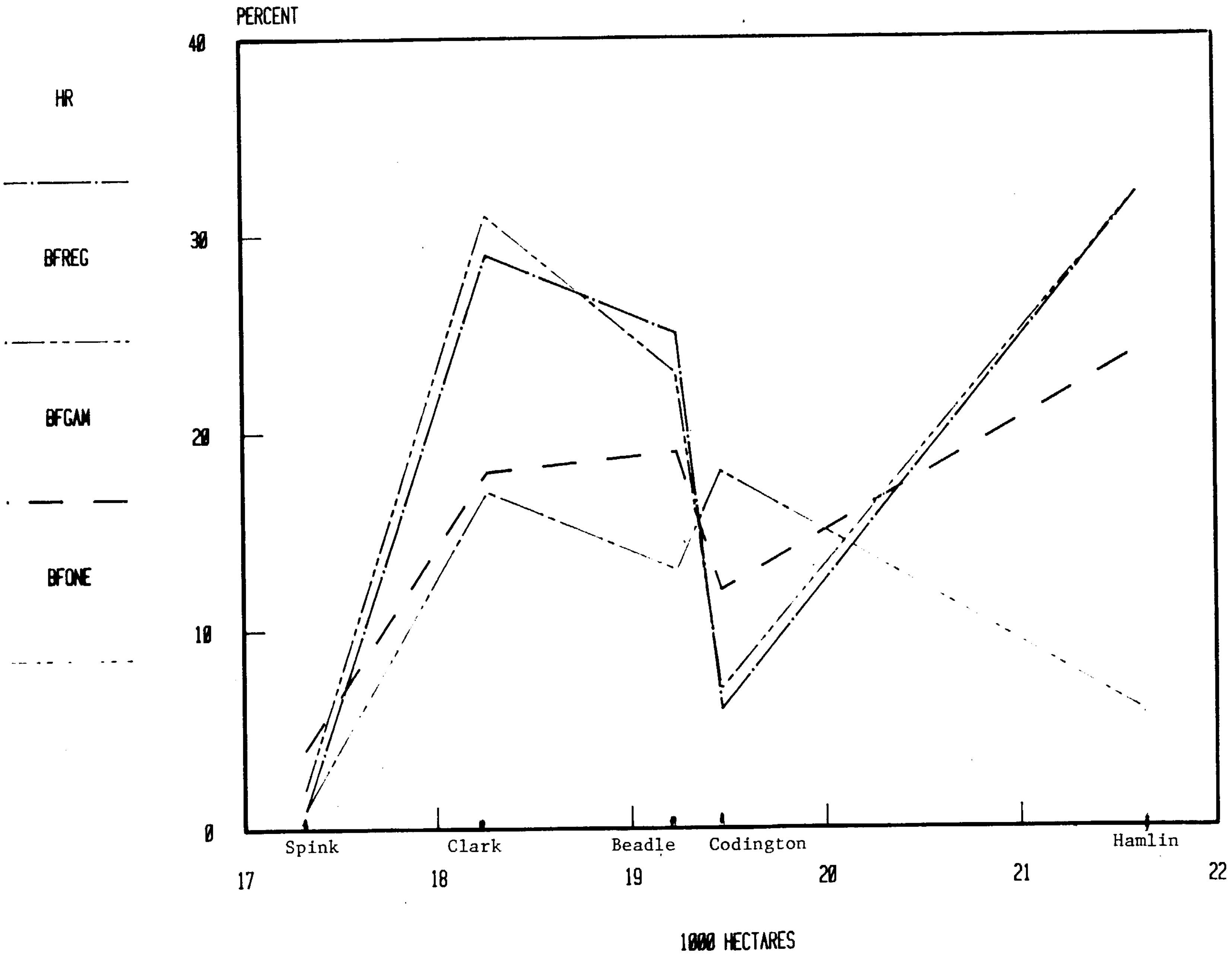
Estimator abbreviations are as defined on page 14.

CROP	COUNTY	HR	BFREG	BFGAM	BFONE
Corn	Codington	69	61	47	38
	Spink	33	48	13	1
	Beadle	36	50	9	6
	Clark	81	69	37	20
	Kingsbury	5	2	5	5
	Hamlin	34	34	16	2
	Average	43	44	31	12
Oats	Codington	6	7	12	18
	Spink	1	2	4	1
	Beadle	25	23	19	13
	Clark	29	31	18	17
	Kingsbury	—	—	—	—
	Hamlin	32	32	24	6
	Average	19	19	15	11
Flax	Codington	16	16	12	5
	Spink	12	7	19	38
	Beadle	—	—	—	—
	Clark	79	77	85	93
	Kingsbury	56	55	58	62
	Hamlin	53	53	55	56
	Average	43	42	46	51
Sunflowers	Codington	32	37	28	44
	Spink	12	14	4	1
	Beadle	—	—	—	—
	Clark	43	48	85	108
	Kingsbury	—	—	—	—
	Hamlin	—	—	—	—
	Average	29	33	39	51

ABSOLUTE AVERAGE RELATIVE BIAS - CORN



ABSOLUTE AVERAGE RELATIVE BIAS - OATS



As a measure of accuracy for the Battese-Fuller estimates, consider the quotient:

$$\frac{\text{Battese-Fuller estimate - comparison value.}}{\text{RMSE (Battese-Fuller estimate)}}$$

This quantity indicates how wide a Battese-Fuller interval estimate would have to be to contain the comparison value.

The results of carrying out this calculation are recorded in table 10 - part 1 for the crops and counties that had the best relative RMSE's as discussed in table 7. Note that $C = \Gamma$ and $C = 0$ give more instances of intervals requiring two or more RMSE's then does $C = I$.

A similar calculation was done to measure proximity of the Battese-Fuller estimate with $C = \Gamma$ to the Huddleston-Ray estimate. This appears in table 10 - part 2.

Table 10 - Part 1: $\frac{\text{BF estimate - comparison value}}{\text{RMSE(BF estimate)}}$

*indicates result for the set of 200 segments

Numerical entries are the numbers of groups of 75 segments each that had the indicated result.

Crop	Estimate (see page 14 for abreviation key)	County	Frequency			
			≤ 1	1-2	2-3	≥ 3
Corn	BFREG	Codington	*1	5	2	0
		Spink	0	*7	1	0
		Beadle	0	*6	2	0
		Clark	*1	3	3	1
		Kingsbury	*8	0	0	0
		Hamlin	1	*4	2	1
Corn	BFGAM	Codington	*1	3	4	0
		Spink	*5	2	1	0
		Beadle	*6	2	0	0
		Clark	*1	6	0	1
		Kingsbury	*6	1	1	0
		Hamlin	*3	3	1	1
Corn	BFONE	Codington	*2	4	2	0
		Spink	*5	3	0	0
		Beadle	*6	2	0	0
		Clark	*6	2	0	0
		Kingsbury	*7	1	0	0
		Hamlin	*6	2	0	0

Table 10 - Part 1—Continued

Crop	Estimate (see page 14 for abreviation key)	County	Frequency			
			≤ 1	1-2	2-3	≥ 3
Oats	BFREG	Codington Clark Hamlin	*3 *4 *0	3 3 3	1 1 3	1 0 2
Oats	BFGAM	Codington Clark Hamlin	*4 *5 *0	3 2 4	0 1 2	1 0 2
Oats	BFONE	Codington Clark Hamlin	*6 *5 *4	1 2 1	0 1 1	1 0 2
Sun-flowers	BFREG BFGAM BFONE	Spink Spink Spink	*4 *5 7	3 1 *1	0 1 0	1 1 0

Table 10 - Part 2: BFGAM Compared to Huddleston-Ray

$$\frac{\text{BFGAM-HR}}{\text{RMSE(BFGAM)}}$$

* indicates result for the set of 200 segments
Numerical entries are the numbers of groups of 75 segments each that had the indicated result.

		Frequency			
Crop	County	≤1	1-2	2-3	≥3
Corn	Codington	3	4	0*	1
	Spink	3	1	4*	0
	Beadle	0	7	0	1*
	Clark	1	5	2	0*
	Kingsbury	8*	0	0	0
	Hamlin	2	5	0*	1
Oats	Codington	6*	2	0	0
	Spink	7*	1	0	0
	Beadle	5	3*	0	0
	Clark	7*	1	0	0
	Kingsbury	6	2*	0	0
	Hamlin	6*	2	0	0
Flax	Codington	8*	0	0	0
	Spink	8*	0	0	0
	Beadle	8*	0	0	0
	Clark	8*	0	0	0
	Kingsbury	7*	1	0	0
	Hamlin	8*	0	0	0
Sunflowers	Codington	7*	1	0	0
	Spink	6*	1	1	0
	Beadle	8	0	0	0
	Clark	7*	1	0	0
	Kingsbury	6	2*	0	0
	Hamlin	8*	0	0	0

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:

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

This appears in the first column of table 11 and is followed by a similar calculation for for the RMSE.

Table 11: Effect of Assuming Zero Cross Strata Correlation. (Corn)

All entries are percentages as defined on page 46.

County	Group	Change in Estimate	Change in root mean square error	Change in root mean square error using $C=\Gamma$
Codington	200 segments	-2.9	2.6	4.3
	Range over 4 groups of 75 segments each	-2.2-2.4	3.1-8.1	4.2-11.2
Spink	200 segments	1.9	1.5	2.4
	Range over 4 groups of 75 segments each	1.1-6.6	1.3-4.5	2.3-6.8
Beadle	200 segments	2.8	1.6	2.5
	Range over 4 groups of 75 segments each	2.8-10.5	.05-4.7	2.1-7.2
Clark	200 segments	-1.6	1.7	3.0
	Range over 4 groups of 75 segments each	-1.1- -.5	.3-5.5	2.3-9.4
Kingsbury	200 segments	-.9	1.9	3.1
	Range over 4 groups of 75 segments each	-4.6- -.4	-2.9-6.2	2.9-8.7
Hamlin	200 segments	-1.7	3.2	5.0
	Range over 4 groups of 75 segments each	-4.4 - -.5	1.6-9.8	4.3-14.5

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%.

RECOMMENDATIONS

Since 1978 the SRS Remote Sensing Branch has provided LANDSAT-based crop-area estimates to the SRS Crop Reporting Board. These estimates are made for entire states and have been submitted in a timely fashion for use by the Board in making final, end-of-year estimates. Following the submission of these state-level estimates to the Board, Huddleston-Ray county estimates have been provided to the SRS State Statistical Offices (SSO's) in LANDSAT-project states. Since this effort by the Remote Sensing Branch both, for state and county level estimates, is a research activity, the authors recommend that the calculation of Huddleston-Ray estimates be discontinued and replaced with the calculation and dissemination to the SSO's of Battese-Fuller estimates. Specifically, on the basis of this study, the authors recommend the following scheme:

o Assume $\sigma_v^2 = 0$ when the test for this hypothesis cannot be rejected at the 0.25 level.

o When $\sigma_v^2 = 0$ is assumed, calculate the Battese-Fuller estimate with C = 0.

o When $\sigma_v^2 \neq 0$ is assumed, calculate the Battese-Fuller estimates with C=Γ and C=I. If both sets of estimates provide acceptably small relative root mean square errors, use the latter in order to also reduce bias.

o When a nonzero σ_v^2 is used in more than one stratum, the cross-strata covariance should be estimated.

These recommendations are supported by the following reasons:

o When $\sigma_v^2 = 0$ is assumed, the Battese-Fuller and Huddleston-Ray estimators are equivalent.

o When there is a large county effect ($\sigma_v^2 \gg 0$), this study has shown that the C = Γ and C = I Battese-Fuller estimators are superior to both the Huddleston-Ray estimator and the C = 0 Battese-Fuller estimators.

o The lack of knowledge of robustness to model violations is the same, if not less, for the Battese-Fuller estimator as for the Huddleston-Ray estimator.

o Amis, et al, (1982) have shown that the estimated variance of the Huddleston-Ray estimator (calculated with the formula in Huddleston and Ray (1976), is very conservative.

Moreover, the authors make this recommendation with the full realization that this study raises as many questions as it answers. Additional research on the behavior of Battese-Fuller estimates is definitely needed. Dissemination of these estimates to the SSO's will necessitate and contribute to additional required research.

Additional research is required to answer the following questions:

o What is the typical range of values of σ_e^2, σ_v^2, δ, and σ_{vhh}, for JES (plus LANDSAT) data? This question can be answered by analyzing retrospective data sets, such as those described by Mergerson, et al (1982).

o Are the Battese-Fuller estimators robust against nonhomogeneous error variances?

o Can the least squares estimate of the regression slope replace the Battese-Fuller procedure when error variances are nonhomogeneous?

o Can the effect of heterogeneous regression slopes on the Battese-Fuller estimates be predicted?

o Does the R^2 value establish a bound on the degree of model failure which may be present?

o Can the homogeneity tests used in this study be effectively carried out on the smaller sample sizes that are present in the JES?

ACKNOWLEDGEMENTS

The authors wish to acknowledge Ron Steele for the sample design of the original SRS/RSI soil study; Kevin Dahlstead of RSI for ground data collection; Paul Cook for scene-to-map LANDSAT registration and preliminary EDITOR analysis; Ray Seyfarth of NASA/NSTL for scene to scene LANDSAT registration; William Wigton for continual encouragement; the Remote Sensing Branch's Technical Support Group for digitization of ground data photographs and strata boundaries; Martha Amis, Virginia Martin, Bill McGuire, and Sylvia Shen of NASA/JSC for sample reallocation, signature development, and LANDSAT classifications; Cheryl Auer of the Iowa State University Statistics Laboratory for development of the estimation computer program; Professor Wayne Fuller of Iowa State University for suggestions concerning theory; and Yvonne Zamer and Kathy Reiman for manuscript word processing.

REFERENCES

Amis, M.L., M.V. Martin, W.R. McGuire, and S.S. Shen. 1982. Evaluation of Small Area Crop Estimation Techniques Using LANDSAT and Ground-Derived Data. Lockheed Engineering and Management Services Company, Inc. LEMSCO-17597. Houston, Texas.

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.

Mergerson, J.W, C.E. Miller, M. Ozga, S. Winings, P. Cook, and G. A. Hanuschak. 1982. 1981 AgRISTARS DCLC Four State Project. Proceedings, Eighth International Symposium on Machine Processing of Remotely Sensed Data, West Lafayette, Indiana.

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.