I. INTRODUCTION

The USDA's National Agricultural Statistics Service (NASS) has long been involved with the application of earth resources satellite data to crop area estimation. In certain regions of the United States, estimates of area planted or harvested in specific crops have been generated using combined satellite and ground survey data. These figures are used as inputs to NASS's setting of official state level crop area estimates.

The satellite data used for this application are usually from Landsat or French SPOT satellites. A multispectral satellite scene consists of many pixels, each having a vector of scaled energy reflectance values in several bands of the electromagnetic spectrum. SPOT's spatial resolution (pixel length) of 20 meters is finer than TM's 30 meters, but TM has seven bands vs SPOT's three. NASS often uses multitemporal data, consisting of two scenes from different dates over the same area.

Crop area estimates at the state or regional (large domain) level are computed by applying regression equations to population level classified pixel counts within area frame land use strata, then summing over strata. The use of regression between ground survey and satellite data often results in much lower variance than direct expansion estimation using survey data alone.

In 1972, the first Landsat satellite was launched, carrying the Multispectral Scanner (MSS) with 80 meter resolution. From 1972 to 1979, NASS developed the basic methodology for satellite based crop area estimation and tested it in several states (Hanuschak et. al., 1982). The first large scale operational remote sensing program was the Domestic Crops and Land Covers (DCLC) project (1980-87), involving eight states in the central United States (Allen and Hanuschak, 1988). Landsat MSS data were used to generate annual acreage indications for corn, soybeans, cotton, rice, sorghum and winter wheat. The DCLC estimates had lower sampling errors than survey based estimates and were usually closer to the official estimates issued by NASS's Agricultural Statistics Board.

The DCLC program was discontinued in 1987 in order to perform research on new sensors and implement advanced computing technology. During 1988-90, research projects in four states found that use of the Landsat TM sensor led to more efficient crop estimates than either the Landsat MSS or French SPOT sensors. The largest improvements of TM based estimates over survey based estimates were found for cotton and rice in Arkansas (Allen, 1990b). NASS returned to operational remote sensing in 1991 with a project in the Mississippi River Delta region, discussed in Section III. In 1994, Landsat TM data were used for crop classifications and mapping in a pilot study involving the Crow and Northern Cheyenne Reservations in Montana (Graham and Hanuschak, 1994). In addition to crop area estimation, NASS uses earth resources satellite data in the construction of area sampling frames (Bush and House, 1993).

Section II discusses the methodology for satellite based crop area estimation, including the separate regression estimator. Section IV introduces four alternative satellite based estimators and discusses their properties. Section V compares estimators using empirical data.

II. METHODOLOGY

This section summarizes the procedures NASS follows to process satellite and ground survey data for crop area estimation. More detailed descriptions of the methodology are provided by Allen (1990a) and Graham (1993).

The PEDITOR software system, developed at NASS, is used for most data processing (Ozga et. al., 1992). PEDITOR is installed on a MicroVax 3500 computer and IBM compatible personal computers.

NASS conducts the June Agricultural Survey (JAS) annually in almost every state. The area frame portion of the survey uses a stratification of each state's area based on land use (Bush and House, 1993). The sample units are land areas called segments, usually one square mile. Each year, about 20 percent of segments are rotated into or out of a state's area sample. During the survey, enumerators interview the land operators in each sampled segment, recording the cover (crop/land use), size and boundaries of every field. The field boundaries within segments are drawn onto aerial photographs.

Field boundaries are transferred from segment photos to digital form. For segments that remain in the sample from one year to the next but have field boundary changes, the previous year's digitized files are updated to reflect the changes. The satellite scenes are registered to a map base in latitude/longitude coordinates, allowing JAS enumerated fields to be matched with their corresponding satellite pixels.

Remote sensing analysts divide a state into analysis districts, within which separate analyses are done. An analysis district is an area covered
by one or more satellite scenes having the same overpass date, or an area for which usable satellite coverage is not available. Crop area estimates are generated at the district level and later summed to obtain state level estimates.

For a given analysis district, the pixels representing specific ground cover types are gathered into separate files and clustered using a modified ISODATA algorithm to generate multivariate discriminant functions known as cover signatures (Bellow and Ozga, 1991). All pixels in the analysis district are then categorized to cover types using maximum likelihood classification. The subset of classified pixels representing the sample segments is used for classification accuracy assessment and regression.

Analysts use a first-order regression model to relate classified pixel counts to the ground survey data on a per stratum basis. Regression is performed only in strata having sufficient sample sizes to obtain a valid relationship. The regression equations are applied to stratum level classified pixel counts to obtain stratum level crop area estimates. Survey based direct expansion estimates are substituted for regression estimates in strata where regression is not done. The estimates are summed over strata to get analysis district level estimates, then over districts to obtain state level estimates. Variance estimates at the district and state levels are also computed.

For convenience, the regression strata will be labeled \( h=1, \ldots, H_r \) and the non-regression strata \( h=H_r+1, \ldots, H \), where \( H_r \) is the number of regression strata and \( H \) is the total number of strata in the analysis district. The formula for the separate regression estimator (SRGE) of crop acreage in the regression strata is:

\[
\hat{\gamma}(\text{SRG}) = \sum_{h=1}^{H_r} N_h \left( \bar{y}_h - \bar{x}_h \right)
\]

where:
- \( N_h \) = number of population units in stratum \( h \)
- \( \bar{y}_h \) = mean reported crop acreage per sample segment in stratum \( h \)
- \( \bar{x}_h \) = mean pixels per sample segment classified to crop in stratum \( h \)
- \( \hat{b}_h = \sum_{i=1}^{n_h} \frac{(x_{hi} - \bar{x}_h)(y_{hi} - \bar{y}_h)}{n_h - 1} \) (in place of \( s_{xyh} \))
- \( n_h \) = number of sample segments in stratum \( h \)
- \( \bar{x}_h \) = mean pixels per population unit classified to crop in stratum \( h \)

The formula for the direct expansion estimator (DE) in the non-regression strata is:

\[
\hat{\gamma}(\text{DE}) = \sum_{h=H_r+1}^{H} N_h \bar{y}_h
\]

The variance estimator of SRGE is:

\[
v(\hat{\gamma}(\text{SRG})) = \sum_{h=1}^{H_r} \frac{N_h(N_h-n_h)(n_h-1)}{n_h(n_h-2)} \left[ s_{ywh} - b_h s_{xwh} \right]
\]

where:
- \( s_{ywh} = \frac{1}{n_h-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2 \)
- \( s_{xwh} = \frac{1}{n_h-1} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)(y_{hi} - \bar{y}_h) \)
- \( \bar{y}_h \) = reported crop acreage in stratum \( h \), sample segment \( i \)
- \( x_{hi} \) = number of pixels classified to crop in stratum \( h \), sample segment \( i \)

The variance estimator of DE in the non-regression strata is:

\[
v(\hat{\gamma}(\text{DE})) = \sum_{h=H_r+1}^{H} \frac{N_h(N_h-n_h)}{n_h} s_{ywh}^2
\]

The composite state level estimate is the sum of the separate regression and direct expansion terms:

\[
\hat{\gamma}(\text{CMP}) = \hat{\gamma}(\text{SRG}) + \hat{\gamma}(\text{DE})
\]

Similarly, the variance estimate of the composite state level estimator is the sum of the variance estimates of the two components.

Cochran (1977) remarks that the ratio of bias to standard error of the SRGE may become appreciable. Since stratum level regression estimates of means can have biases of order \( 1/n_h \) and the biases may be in the same direction in all strata, the bias of the overall estimate of total could be of order \( N_h/n_h \). However, the risk of that large a bias is small if the relation between the two variables is fairly linear.

Chhikara et al. (1988) identified a problem known as "overfitting". The use of the same area frame segments to develop both the crop signatures and the regression relationships can contribute additional bias to the estimates.

The methodology for state and regional level estimation has been adapted for county level (small domain) estimation, using a Battese-Fuller random effects model (Bellow, 1993).

III. MISSISSIPPI DELTA PROJECT

NASS's remote sensing effort in the Mississippi Delta region began in 1991 as an operational crop area estimation program. The goal was to provide timely state and county level acreage estimates of major crops to the Agricultural Statistics Board and NASS State Statistical Offices involved. The states in the
The two main crops estimated in the Delta project are cotton and rice. Table 1 gives the 1991-93 Landsat composite estimates, JAS direct expansion estimates and NASS official estimates of both crops. The relative efficiency (RE), defined as the ratio of the variance of the JAS direct expansion estimate to that of the Landsat composite estimate, is also shown.

In 1993, two separate analyses were done. The first analysis used unitemporal satellite data from the spring and provided inputs in time for NASS's August Crop Production Report. The second analysis, similar to previous years, used multitemporal satellite data from spring and summer to produce end-of-year estimates. This analysis also used follow-up ground survey information not available in time for the early season estimation. Both types of estimate are given in Table 1 for comparison.

Table 1 shows that the Landsat separate regression estimate was always below the JAS direct expansion estimate for both cotton and rice. The Landsat estimate was below the final NASS estimate in five of seven cases for cotton and four of seven cases for rice, and was closer than DE to the final NASS estimate in three of seven cases for cotton and four of seven cases for rice. Relative efficiencies ranged from 2.2 to 21.0 for cotton and 1.5 to 5.5 for rice. In the 1993 Arkansas analyses, the late season RE was more than twice the early season RE for both crops.

From the above results and previous DCLC findings, the following observations can be made. Satellite based estimation can achieve dramatic reductions in variance over the traditional survey based methods. The degree of reduction varies with crop. However, in the Delta project the Landsat based estimate was closer than the JAS direct expansion estimate to the final NASS estimate less than half the time, as opposed to 60 percent for the DCLC project over eight years. The Landsat estimate tended to fall below the corresponding direct expansion estimate.

IV. ALTERNATIVE SATELLITE BASED ESTIMATORS

In this section, four alternative large domain crop area estimators are described. These estimators are based on the same pixel classification used to compute the separate regression estimator. Three of the four estimators use the overall (across-strata) count of pixels classified to a crop. The remaining estimator requires the individual stratum level pixel counts for its computation. The rationale for introducing these estimators is to compare their bias and variance properties with those of SRGE.

A. Raw Pixel Count Estimator

\[ \hat{y}_{RPC} = \lambda X \]

where:

\( \lambda \) = conversion factor (area units per pixel)

\( X \) = number of pixels classified to crop of interest in analysis district

The raw pixel count estimator (RPC) is a direct count of pixels classified to the crop of interest, converted to area units. Since it represents a complete enumeration of classified pixels in an area, RPC does not have sampling error. However, there is a theoretical bias due to classification error, which can be approximated by:

\[ B(RPC) = Y \left[ \alpha_c - \alpha_o \right] \left[ 1 - \alpha_c \right] \]

where:

\( Y \) = true area planted to crop of interest in analysis district

\( \alpha_o \) = probability that a pixel belonging to crop of interest is not classified to that crop (omission error)

\( \alpha_c \) = probability that a pixel classified to crop of interest does not belong to that crop (commission error)

Thus the bias is positive or negative depending upon whether the pixel level probability of commission error is greater than or less than that of omission error. The denominator term indicates that the bias is especially sensitive to commission error, so RPC can severely overestimate the true crop acreage if \( \alpha_c \) is high.

B. Separate Ratio Estimator

\[ \hat{y}_{SR} = \sum_{h=1}^{H} \left( \hat{y}_h / \hat{x}_h \right) x_h \]

Variance Estimator

\[ \text{Var}\left( \hat{y}_{SR} \right) = \sum_{h=1}^{H} \left[ n_h (N_h - n_h) / n_h \right] \left[ s_{yh}^2 + \hat{r}_h^2 s_{xh}^2 - 2 \hat{r}_h s_{xyh} \right] \]

where:

\[ s_{xh}^2 = \frac{1}{(n_h - 1)} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2 \]
Chhikara et al. (1986) studied ratio estimators for crop area estimation at the individual stratum level. Ratios can be computed in each stratum having a positive number of sample segments and for which a positive number of pixels were classified to the crop of interest. Direct expansion is used in other strata to form an overall composite estimate. If ratios are computed in all H strata, then the following statement can be made about the bias of SRE:

\[ B[\hat{Y}(SR)] = o[H^{1/2}G(\hat{Y}(SR))CV(\hat{x}_{st})] \]

where \( o(.) \) means "on the order of", \( G(.) \) denotes the true standard deviation and \( CV(.) \) denotes the true coefficient of variation. Cochran (1977) does not recommend SRE unless the sample size in each stratum is large enough that the variance estimate is valid and the cumulative bias is negligible.

C. Combined Ratio Estimator

\[ \hat{Y}^{(CR)} = \frac{\hat{Y}_{st}}{\hat{x}_{st}} \] \( \times \)

where:

\[ N = \text{total number of population units} \]

\[ \hat{x}_{st} = \frac{1}{N} \sum_{h=1}^{H} N_h \hat{x}_h \]

\[ \hat{y}_{st} = \frac{1}{N} \sum_{h=1}^{H} N_h \hat{y}_h. \]

Variance Estimator -

\[ v(\hat{Y}^{(CR)}) = \frac{H}{H-1} \sum_{h=1}^{H} \left\{ \frac{(N_h - N_h^*)/n_h}{N_h} \right\} \left[ s_{yh}^2 + R^2 s_{xh}^2 - 2 R s_{xyh} \right] \]

The combined ratio estimator (CRE) represents an adjustment of RPCE to compensate for bias. The adjustment factor is the ratio of expanded reported acreage to expanded classified pixel count for the crop of interest. The combining of data from all strata eliminates the need to use direct expansion in weak strata, as was the case with SRE. The bias of CRE has the following upper bound (Cochran, 1977):

\[ B[\hat{Y}^{(CR)}] \leq G(\hat{Y}^{(CR)})CV(\hat{x}_{st}) \]

Thus the bias is negligible relative to the standard error if the CV of the weighted pixel mean is less than 0.1. CRE is much less prone to bias than SRE.

D. Combined Regression Estimator

\[ \hat{Y}^{(CRG)} = N \left[ \hat{y}_{st} + b(\hat{x} - \hat{x}_{st}) \right] \]

where:

\[ b = \left( \sum_{h=1}^{H} A_h s_{xyh} / \left( \sum_{h=1}^{H} A_h s_{xh}^2 \right) \right) \]

\[ A_h = \frac{N_h (N_h - N_h)^*}{n_h} \]

\[ \hat{x} = \frac{X}{N} \]

Variance Estimator -

\[ v(\hat{Y}^{(CRG)}) = \frac{H}{H-1} \sum_{h=1}^{H} \left\{ \frac{(N_h - N_h^*)/n_h}{N_h} \right\} \left[ s_{yh}^2 + R^2 s_{xh}^2 - 2 R s_{xyh} \right] \]

The combined regression estimator (CRGE) is analogous to the combined ratio estimator in that information from all strata is combined. This estimator requires that sample segment sizes be the same in all strata having a positive number of segments in the analysis district. Cochran (1977) observes that CRGE is less prone to bias than SRGE when sample sizes are small within individual strata. Furthermore, the variance of SRGE has a larger contribution from sampling errors in the regression coefficients. The variance of CRGE is inflated if the population regression coefficients differ from stratum to stratum. CRGE is preferred if the regressions are linear with slopes roughly the same in all strata.

V. COMPARISON USING EMPIRICAL DATA

An empirical comparison of the five satellite based estimators (SRE, CRE, SRGE, CRGE, RPCE) and the survey based direct expansion estimator (DE) was done using 1991 data from Mississippi and 1993 data from Arkansas. The Mississippi district extends the length of the state along the Mississippi River, containing all or part of 33 counties. The Arkansas district contains 12 counties in the east central part of the state. The following discussion is intended to illustrate estimator performance using the two data sets; more general hypotheses or conclusions beyond the context of the study should not be inferred.

As a benchmark for evaluating the estimators, prorated "official" estimates of crop acreage in the two districts were calculated. Official county estimates are issued by NASS's State Statistical Offices. The official crop area estimates for counties entirely contained in the analysis district were summed, then added to the sum of scaled official estimates for counties partially inside the district. The scaling was done by multiplying the full county estimate by the ratio of number of population units in the included portion of the county to number of population units in the whole county.

The results are given in Tables 2 through 5. The agricultural land use strata are defined based on percent cultivation, given in parentheses in the "stratum" column. RPCE was either the highest
or lowest of the five satellite based estimates in all cases, and was much higher than the others for cotton in Mississippi and rice in Arkansas. This observation is not surprising in light of the discussion of RPCE’s bias in Section IV. The four satellite based estimates other than RPCE always fell below the “official” estimates. DE was much higher than SRGE, CRGE, SRE, CRE in both districts for rice, while fairly close to them for cotton. The two combined-type estimates (CRE, CRGE) were higher and closer to “official” than both separate-type estimates (SRE, SRGE) for both crops in Mississippi and rice in Arkansas. In the same three cases, the variances of those four estimators did not differ appreciably.

These empirical results and the theoretical properties given in Section IV suggest that the three alternative estimators most competitive with SRGE are SRE, CRE and CRGE. While these estimators exhibit similarities, a given estimator may be preferred under certain conditions based upon its unique attributes. In particular, the two combined-type estimators have more favorable bias properties than SRGE. Future research will compare the estimators for other crops in different regions.

VI. SUMMARY

This paper described the history and status of large domain satellite based crop area estimation at NASS, and compared several estimators for this application. Since the onset of satellite data research in 1972, NASS has developed and refined the methodology through a series of research and operational programs. The procedures have been consistently updated to take advantage of improving remote sensing and computing technology. The SRGE shows significantly reduced variance when compared with the survey based direct expansion estimator. However, operational satellite based estimates from the DCLC and Delta projects have generally fallen below direct expansion estimates. Four alternative satellite based estimators were introduced and their properties discussed. An empirical study evaluated estimator performance using cotton and rice data from the Mississippi Delta area. Three of the alternative estimators (SRE, CRE, CRGE) are competitive with SRGE. There will be further research on these estimators.

REFERENCES


Table 1: Mississippi Delta Crop Area Estimates (1000 Acres)

<table>
<thead>
<tr>
<th>State</th>
<th>Year</th>
<th>JAS DE</th>
<th>Landsat Reg.</th>
<th>Final NASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arkansas</td>
<td>1991</td>
<td>1256.0</td>
<td>1104.4</td>
<td>1000.0</td>
</tr>
<tr>
<td></td>
<td>1992</td>
<td>1003.0</td>
<td>870.1</td>
<td>1000.0</td>
</tr>
<tr>
<td></td>
<td>1993(E)</td>
<td>1091.5</td>
<td>895.0</td>
<td>980.0</td>
</tr>
<tr>
<td></td>
<td>1993(L)</td>
<td>1094.0</td>
<td>805.2</td>
<td>980.0</td>
</tr>
<tr>
<td>Louisiana</td>
<td>1992</td>
<td>898.5</td>
<td>894.4</td>
<td>880.0</td>
</tr>
<tr>
<td>Mississippi</td>
<td>1991</td>
<td>1277.1</td>
<td>1175.4</td>
<td>1245.0</td>
</tr>
<tr>
<td></td>
<td>1992</td>
<td>1352.8</td>
<td>1210.4</td>
<td>1350.0</td>
</tr>
</tbody>
</table>

E - early season; L - late season

Table 2: Estimated Cotton in Mississippi Research District (1000 Acres)

<table>
<thead>
<tr>
<th>Stratum</th>
<th>SRGE</th>
<th>CRGE</th>
<th>SRE</th>
<th>CRE</th>
<th>RPCE</th>
<th>DE</th>
<th>&quot;Official&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (&gt;75%)</td>
<td>552.2</td>
<td>28.7</td>
<td>553.5</td>
<td>27.3</td>
<td>--</td>
<td>--</td>
<td>568.7</td>
</tr>
<tr>
<td>B (51-75%)</td>
<td>195.6</td>
<td>9.9</td>
<td>192.8</td>
<td>10.3</td>
<td>--</td>
<td>--</td>
<td>159.1</td>
</tr>
<tr>
<td>C (15-50%)</td>
<td>213.5</td>
<td>14.0</td>
<td>212.1</td>
<td>15.1</td>
<td>--</td>
<td>--</td>
<td>199.6</td>
</tr>
<tr>
<td>D (&lt;15%)</td>
<td>54.0</td>
<td>10.7</td>
<td>52.5</td>
<td>10.6</td>
<td>--</td>
<td>--</td>
<td>115.7</td>
</tr>
<tr>
<td>Total</td>
<td>1015.3</td>
<td>33.5</td>
<td>1031.0</td>
<td>34.7</td>
<td>1293.5</td>
<td>1046.1</td>
<td>1280.0</td>
</tr>
</tbody>
</table>

Table 3: Estimated Rice in Mississippi Research District (1000 Acres)

<table>
<thead>
<tr>
<th>Stratum</th>
<th>SRGE</th>
<th>CRGE</th>
<th>SRE</th>
<th>CRE</th>
<th>RPCE</th>
<th>DE</th>
<th>&quot;Official&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (&gt;75%)</td>
<td>171.5</td>
<td>9.3</td>
<td>168.1</td>
<td>19.2</td>
<td>--</td>
<td>--</td>
<td>270.6</td>
</tr>
<tr>
<td>B (51-75%)</td>
<td>31.3</td>
<td>2.9</td>
<td>32.3</td>
<td>3.6</td>
<td>--</td>
<td>--</td>
<td>43.8</td>
</tr>
<tr>
<td>Total</td>
<td>202.8</td>
<td>19.5</td>
<td>202.7</td>
<td>19.3</td>
<td>212.5</td>
<td>19.3</td>
<td>314.4</td>
</tr>
</tbody>
</table>

Table 4: Estimated Cotton in Arkansas Research District (1000 Acres)

<table>
<thead>
<tr>
<th>Stratum</th>
<th>SRGE</th>
<th>CRGE</th>
<th>SRE</th>
<th>CRE</th>
<th>RPCE</th>
<th>DE</th>
<th>&quot;Official&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (&gt;75%)</td>
<td>126.7</td>
<td>6.2</td>
<td>127.2</td>
<td>6.4</td>
<td>--</td>
<td>--</td>
<td>143.3</td>
</tr>
<tr>
<td>B (25-75%)</td>
<td>23.4</td>
<td>0.5</td>
<td>21.9</td>
<td>1.7</td>
<td>--</td>
<td>--</td>
<td>12.3</td>
</tr>
<tr>
<td>C (&lt;25%)</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>--</td>
<td>--</td>
<td>0.7</td>
</tr>
<tr>
<td>Total</td>
<td>150.7</td>
<td>6.3</td>
<td>148.9</td>
<td>9.0</td>
<td>148.9</td>
<td>9.0</td>
<td>156.1</td>
</tr>
</tbody>
</table>

* - Direct expansion value used for SRGE and SRE

Table 5: Estimated Rice in Arkansas Research District (1000 Acres)

<table>
<thead>
<tr>
<th>Stratum</th>
<th>SRGE</th>
<th>CRGE</th>
<th>SRE</th>
<th>CRE</th>
<th>RPCE</th>
<th>DE</th>
<th>&quot;Official&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (&gt;75%)</td>
<td>431.9</td>
<td>11.6</td>
<td>438.2</td>
<td>11.8</td>
<td>--</td>
<td>--</td>
<td>566.1</td>
</tr>
<tr>
<td>B (25-75%)</td>
<td>59.1</td>
<td>5.0</td>
<td>63.8</td>
<td>5.6</td>
<td>--</td>
<td>--</td>
<td>106.8</td>
</tr>
<tr>
<td>Total</td>
<td>491.0</td>
<td>12.6</td>
<td>521.3</td>
<td>13.1</td>
<td>583.0</td>
<td>13.1</td>
<td>672.9</td>
</tr>
</tbody>
</table>

199