STATE LEVEL CROP AREA ESTIMATION USING SATELLITE DATA IN A REGRESSION ESTIMATOR

Mitchell L. Graham, USDA/NASS
3251 Old Lee Hwy. Rm. 305 Fairfax, Virginia 22030

KEY WORDS: regression estimator, Landsat Thematic Mapper, land cover estimate, crop acreage estimate.

ABSTRACT

The USDA's National Agricultural Statistics Service (NASS) estimates state level crop acreage in the Mississippi Delta region using area frame survey data and Landsat Thematic Mapper (TM) satellite data. Five general steps produce these acreage estimates. First, a sample of TM pixel data is clustered by land cover. Second, sampled TM pixels are assigned to a land cover class using maximum likelihood classification. Third, classified sample pixels are regressed with reported crop acreages. Fourth, TM scenes are classified. Finally, acreage is estimated with a regression estimator using classified pixel counts as ancillary information to the ground survey data. The potential benefit is mainly a reduction in variance with some adjustment of the state acreage estimates.

BACKGROUND

The Mississippi River Delta region is the most important rice producing area in the United States and is also a major cotton producing area. The region, which includes all or part of five states, accounted for 76 percent of U.S. planted rice acreage and 29 percent of U.S. planted cotton acreage in 1991. With 1.3 million planted acres of rice, Arkansas was the major Delta rice producing state accounting for 46 percent of the 1991 national total. (USDA NASS, 1992). The 1992 Arkansas rice estimate was 1.4 million planted acres; the 1993 estimate was 1.35 million planted acres (USDA NASS, 1993).

The Delta region provides an ideal setting for remote sensing based estimation techniques. NASS's current general purpose area sampling frame is not designed for crops that are localized in specific areas. This condition can lead to high state level relative sampling errors for crops such as cotton and rice. In Arkansas, nearly all the rice and cotton occur in the eastern third of the state oriented north-to-south along the Mississippi River. This geographic orientation coincides with the ground viewing orientation of polar orbiting Landsat satellites and minimizes the number of satellite scenes needed to cover Arkansas.

DATA PROCESSING

PEDITOR is used for data processing on a MicroVax 3500 computer and on IBM PC compatibles in a DOS environment. PEDITOR is a special purpose software system developed at NASS (Ozga et al., 1992) for crop area estimation. PEDITOR is mainly written in PASCAL and contains modules for image display and processing, as well as estimation. Image display and graphics modules are run on PCs, while non-graphics modules can run on a either a PC or MicroVax. Computationally intensive jobs, such as classification of multitemporal TM scenes, are processed on a Cray supercomputer (Idaho National Engineering Laboratory Supercomputing Center in Idaho Falls, Idaho).

DATA ACQUISITION

For the 1991/92 Delta Project, NASS's Remote Sensing Section (RSS) acquired ground data from the June Agricultural Survey (JAS) and Landsat data from EOSAT Corporation. Data acquisition involved the JAS, a recheck visit to JAS segments, spring TM scene selection, and summer TM scene selection.

The ground sample units were small land areas called segments, each about one square mile for strata 11, 12, 20 and 21. Segments were selected randomly from an area sampling frame stratified by land use categories ordered by percent of cultivated land. See Table 1. During the June survey, field enumerators interviewed the land managers in each segment and recorded the land cover (rice, fallow, soybeans, pasture, woods, water, etc.), size, and boundaries for every field. Uncultivated areas within a segment were also recorded. At this point, the survey data could be used to make NASS's usual preliminary crop area estimates having measurable precision, but based on ground data alone. Mid-summer, RSS rechecked segments where a farmer indicated, during the JAS, that a crop would be planted later.

Using knowledge of cropping practices, analysts selected Landsat TM scene dates to facilitate crop discrimination within the constraints imposed by cloud

cover and scene availability. TM data consists of seven spectral measurements on each of 41.6 million picture elements (pixels) arranged in a 5965 by 6967 array called a scene. When possible, spring and summer Landsat TM scenes from the same area were combined to create a single multitemporal, 14 dimensional, satellite data set. Each Landsat scene was reformatted and registered to 1:250,000 USGS maps. Then sampled segments were digitized and located within each Landsat scene. When the geographic correspondence between TM pixel data and JAS segments was established, the Landsat TM data were analyzed by land cover.

Table 1: USDA NASS Land-use Strata for Arkansas during 1991 and 1992.

Stratum #	Definition	n	N
(1991--implemented in 1974)			
11	over 80 % cultivated	144	11,723
12	51 to 80 % cultivated	48	5,697
20	15 to 50 % cultivated	84	11,673
31	agri-urban: > 20 home/mile^2	28	5,019
32	commercial: > 20 home/mile^2	4	1,371
33	resort: > 20 home/mile^2	4	532
40	less than 15 % cultivated	84	10,658
50	non-agricultural	4	889
(1992--implemented in 1992)			
11	over 75 % cultivated	195	11,673
21	25 to 75 % cultivated	40	2,718
31	agri-urban: > 100 home/mile^2	10	1,308
32	commercial: > 100 home/mile^2	5	418
42	less than 25 % cultivated	140	18,561
40	non-agricultural	5	35

Table 2: Landsat TM Scene Overpass Dates for 1991 and 1992 Arkansas Analysis Regions.

Analysis Region	Multi-temporal	Overpass Date Pass 1.	Pass 2.
1991			
Eastern	yes	4/01/91	8/23/91
Central	no	8/14/91	---
1992			
Northeast	yes	5/05/92	7/24/92
Southeast	yes	5/05/92	6/22/92
Central	yes	4/26/92	8/16/92

The TM scene acquisition dates and data quality affect the organization of both analysis and estimation. To control atmospheric and phenological factors, areas of Arkansas viewed by Landsat on different dates are analyzed and processed separately. The Landsat 5 satellite flies North to South over Arkansas in three partially overlapping passes which cover, or view, the eastern, central and eastern regions of the state on different dates. Landsat 5 repeats any given pass every 16 days with neighboring passes either seven or eleven days apart. At best, the central and eastern passes may be seven days apart. In some cases bad weather requires dividing a single path (pass) into two analysis regions that differ by 16, 32 or more days. See Table 2 for TM scene overpass dates.

SATELLITE DATA ANALYSIS

Separately, for each land cover within each analysis region, the segment Landsat data were studied for outlier pixels and then clustered using a modified ISODATA algorithm (Bellow and Ozga, 1991). Outlier pixels were identified using principal component analysis and removed from the data before clustering. The result of clustering each land cover, ç, was several separable vectors, $S_ç$, of spectral reflectance each referred to as a signature. The signatures in $S_ç$ were assumed to represent noticeable variations in the land cover. For example, in S_{rice} separate signatures were expected for unplanted fields, flooded fields, waste areas, fields in good or bad condition, and mixtures of rice and other covers.

When all land covers were clustered, the $S_ç$ were assembled into one collection of signatures, $S_{(all)}$. The separability of the land cover signatures in $S_{(all)}$ was analyzed using Swain-Fu (Swain 1972) or transformed divergence (Swain and Davis 1978) statistics. Some signatures were separable. Most signatures had a degree of separability that would allow them to still be useful for classification. The signatures with the poorest separability were removed from $S_{(all)}$, or averaged with similar signatures, producing an edited collection of signatures, $S_{(edited)}$. Each vector in $S_{(edited)}$ was still tagged with its original land cover but was also considered a separate category of surface reflectance.

Analysts used $S_{(edited)}$ as input into the discriminate function categorizing Landsat TM pixels into separate reflectance categories. There were two phases of maximum likelihood classification. First the segment pixel data were classified. Then after analysis and refinement of segment classification, whole TM scenes were classified.

Analysis of sample segment classification consisted of three parts. First classified segment pixels were tabulated by the reflectance categories in $S_{(edited)}$. Next

commission and omission error based on the original land use tags were examined using the kappa statistic (Congalton, 1991). Then segment classified pixel counts were regressed with segment land cover totals univariately for each land cover. A separate first order model was used in each applicable JAS land use stratum. If classification errors were acceptable and simple linear regression analysis revealed no problems with model assumptions nor outlier points, then the segment classified pixel counts were used to calculate the sample ancillary mean, and b_1 was used to estimate the slope in the regression estimator. Otherwise, some of the satellite data analysis steps were repeated.

When sample level analysis was complete, analysts used $S_{(edited)}$ in classifying whole Landsat scenes. After a TM scene classification, the scene pixels were tabulated within JAS land use strata by category and land cover. These counts were use in calculating the ancillary population means.

REGRESSION ESTIMATOR

Remote sensing researchers at NASS have used ancillary satellite information in a regression estimator since 1978. Analysts used the regression estimator in this manner for land cover and crop estimation projects with the National Aeronautics and Space Administration and the National Oceanic and Atmospheric Administration (Allen and Hanuschak, 1988). There is a theoretical downward bias of order 1/n with this method (Cochran, 1977).

The NASS area frame stratifies each state by percent of cultivated land (Table 1.). Let s = 1,2,...,H denote these land use strata. In each stratum there are N_s primary sampling units (PSU). NASS randomly selects n_s units (segments) from each stratum for enumeration during the JAS.

After purchasing Landsat TM scenes covering the study area, NASS creates analysis regions for the differing satellite overpass dates (Table 2.). Denote the analysis regions $\alpha = 1,2,...,k,k+1,...,A$ where k of them are covered by Landsat data and A-k of them are not.

Within each analysis region, there are H_α area frame land use strata where the regression estimator is used. If the region is covered by Landsat TM data ($\alpha \le k$), $0 \le H_\alpha \le H$. If the region is not covered by TM data ($\alpha > k$), then $H_\alpha = 0$. Denote the area frame land use strata within a covered analysis region as $h = 1,...,H_\alpha$ for strata where the regression estimator is used and as $h = H_\alpha+1,...,H$ for strata where the regression estimator is not used. If the analysis region is not covered by TM data, $h = H_\alpha+1,...,H$.

$$\text{Let } N_s = N_{.h} = \sum_{\alpha=1}^{A} N_{\alpha h} \quad , \qquad \sum_{s=1}^{H} N_s = \sum_{h=1}^{H} N_{.h} \quad ,$$

$$n_s = n_{.h} = \sum_{\alpha=1}^{A} n_{\alpha h} \quad \text{and} \quad \sum_{s=1}^{H} n_s = \sum_{h=1}^{H} n_{.h} \; .$$

The regression estimator of total acreage for a land cover in an analysis region can be expressed as

$$\hat{Y}_{ç\alpha(reg)} = \sum_{h=1}^{H_\alpha} N_{\alpha h} \left[\bar{y}_{ç\alpha h} + \hat{b}_{ç\alpha h}(\bar{X}_{ç\alpha h} - \bar{x}_{ç\alpha h})\right]$$

$$\mathrm{Var}(\hat{Y}_{ç\alpha(reg)}) = \sum_{h=1}^{H_\alpha} (N_{\alpha h}^2 - N_{\alpha h} n_{\alpha h})/n_{\alpha h}$$

$$\cdot \sum_{j=1}^{n_{\alpha h}} (y_{ç\alpha hj} - \bar{y}_{ç\alpha h})^2 (1 - \hat{R}_{ç\alpha h}^2)/(n_{\alpha h} - 2) \cdot [1 + (n_{\alpha h} - 3)^{-1}]$$

Where $b_{ç\alpha h}$ is regression coefficient b_1 for land cover ç region α and stratum h, and where

$\bar{X}_{ç\alpha h} = \sum_{i=1}^{N_{\alpha h}} X_{ç\alpha hi}/N_{\alpha h}$ and $X_{ç\alpha hi}$ is the count of full scene pixels classified to land cover ç in stratum h from the i^{th} PSU in analysis region α.

Likewise, $\bar{x}_{ç\alpha h} = \sum_{j=1}^{n_{\alpha h}} x_{ç\alpha hj}/n_{\alpha h}$ and $x_{ç\alpha hj}$ is the count of segment pixels classified to land cover ç in stratum h from the j^{th} sample unit in analysis region α.

$\hat{R}_{ç\alpha h}^2$ is the coefficient of determination between the reported acreage and classified pixel count of land cover ç for stratum h in analysis region α.

Now for the remaining analysis regions and strata where Landsat TM data were not used, a direct expansion estimator can be expressed as

$$\hat{Y}_{ç\alpha(dir)} = \sum_{h=H\alpha+1}^{H} N_{\alpha h}/n_{\alpha h} \sum_{j=1}^{n_{\alpha h}} y_{ç\alpha hj}$$

$$\mathrm{Var}(\hat{Y}_{ç\alpha(dir)}) = \sum_{h=H\alpha+1}^{H} (N_{\alpha h}^2 - N_{\alpha h} n_{\alpha h})/(n^2_{\alpha h} - n_{\alpha h}) \sum_{j=1}^{n_{\alpha h}} (y_{ç\alpha hj} - \bar{y}_{ç\alpha h})^2$$

Where $y_{ç\alpha hj}$ is the reported acreage of land cover ç from segment j in stratum h from analysis region α. The state level estimate of land cover ç using ancillary Landsat TM data is written

$$\hat{Y}^{ç}_{TMç} = \sum_{\alpha=1}^{k} \hat{Y}_{ç\alpha(reg)} + \sum_{\alpha=1}^{k} \hat{Y}_{ç\alpha(dir)} + \sum_{\alpha=k+1}^{A} \hat{Y}_{ç\alpha(dir)}$$

$$Var(\hat{Y}_{TMç}) = \sum_{\alpha=1}^{k} Var(\hat{Y}_{ça(reg)}) + \sum_{\alpha=1}^{k} Var(\hat{Y}_{ça(dir)}) + \sum_{\alpha=1}^{A} Var(\hat{Y}_{ça(dir)})$$

RESULTS

For 1991 and 1992, the Remote Sensing Section submitted Landsat crop acreage indications to the NASS Agricultural Statistics Board and the Arkansas State Statistical Office early in December. NASS's Annual Crop Production Report, published in early January, contained crop acreages from the December board.

Before submission, the acreage indications are assessed through examining statistics from each of the main processing steps. Classification accuracy, exclusion error, and inclusion error are assessed using the kappa statistic, percent correct and percent commission. The regression relationship of acres with classified pixels is analyzed for fit, outlier segments and appropriate slope. Since the Landsat TM pixel is approximately 0.201 acres, then b_1 should be near 0.201. Also, the relative efficiency (RE) of the state level Landsat regression estimator to that of the direct expansion (JAS) estimate is noted.

Table 3 gives the kappa statistic, and percent correct and percent commission for rice in Arkansas for 1991 and 1992. Commission errors were better in 1992 with substantially better classification accuracy for 1992 central region.

For both 1991 and 1992 the central and eastern areas of Arkansas were covered by TM scenes. Weather conditions in each year were the final determinate for TM scene selection. In 1991 acceptable TM data were obtained only for mid-summer over the central analysis region while early spring and mid-summer data were available for the eastern region. Consequently, the 1991 central region was analyzed with unitemporal TM data while the eastern region was multitemporal. In 1992, spring as well as summer imagery was available, so that multitemporal TM data sets were created for all regression analysis regions. But the 1992 eastern region had differing summer image dates for northeast and southeast and was therefore divided into two analysis regions to control for atmospheric and crop progress effects. In general, classification accuracy was higher in the multitemporal analysis regions than in the unitemporal regions.

Table 4 shows the stratum level sample sizes ($n_{\alpha h}$) and $R_{çah}^2$ values for those strata where regression was used for rice. Table 5 shows state level direct expansion CV's (CV_{DE}), Landsat regression CV's (CV_{TM}), and the RE's for rice. Table 6 shows the difference of total planted rice acres estimated by direct expansion only from the estimate produced through using the regression estimator scaled by standard error. The state level and analysis region acreage indications (unofficial estimates) cannot be shown due to confidentiality restrictions.

In 1991, both state level direct expansion and regression method indications for planted acres of rice were below the 1991 official NASS estimate, while for 1992 the official estimate was between these two 1992 indications. In 1991, $\hat{Y}_{DE}$ was closer to the official estimate, but in 1992 $\hat{Y}_{TM}$ differed very little from the official NASS estimate. $\hat{Y}_{TM,1991}$ was 1.28 standard errors ($SE_{TM,1991}$) below the 1991 official rice estimate, and $\hat{Y}_{TM,1992}$ was 0.53 standard errors ($SE_{TM,1992}$) above the 1992 official estimate.

Table 3: Kappa values (k), percent correct (ct) and percent commission (cm) for sample segments' classification -- All Rice.

	Analysis Regions		
	Northeast[1]	Southeast[1]	Central[2]
Cover	k ct cm	k ct cm	k ct cm
rice (1991)	71 75 27		67 68 27
rice (1992)	74 79 19	81 84 14	83 87 14

Table 4: Regression of Reported Segment Acreage with Segment Categorized Pixels for All Rice.

	Analysis Regions								
Stra-	Northeast[1]			Southeast[1]			Central[2]		
tum[5]	n	R^2	b	n	R^2	b	n	R^2	b
1991[3]									
11	98	.94	.194	---[1]			23	.96	.222
12	13	.99	.204	---[1]			9	---[4]	
1992[3]									
11	54	.95	.195	53	.98	.203	37	.98	.191
21	1	---[4]		10	.84	.174	7	.97	.190

Table 5: Arkansas State Level Relative Efficiency (RE) for All Rice.

Crop	CV_{DE}(%)	CV_{TM}(%)	RE
Rice (1991)	10.1	5.4	3.9
Rice (1992)	6.8	4.1	3.2

Table 6: Difference in Total Planted Acreage. Direct Expansion Estimate minus Regression Method Estimate Scaled by Standard Error.

Crop	$(\hat{Y}_{DE}-\hat{Y}_{TM})/SE_{DE}$	$(\hat{Y}_{DE}-\hat{Y}_{TM})/SE_{TM}$
Rice (1991)	0.50	0.99
Rice (1992)	1.32	2.36

SUMMARY

In 1991 and 1992, the NASS Remote Sensing Section estimated planted rice acreage in Arkansas using NASS June Agricultural Survey area frame data and ancillary Landsat TM data in a regression estimator. To control for phenological effects, Arkansas was divided into analysis regions based on TM scene overpass dates. Each analysis region was analyzed separately. A regression estimator was used within the intensively cultivated land use strata for the TM covered analysis regions; otherwise, direct expansion was used. The state level acreage estimate was the sum of the analysis region estimates. For 1991, the regression estimator produced a state level indication (unofficial estimate) which was 1.28 standard errors below the NASS official planted acres estimate for rice. In 1992, the indication was 0.53 standard errors above the official estimate. For each year, the regression method indication and variance were less than the corresponding direct expansion indication and variance.

---

[1] The northeast and southeast regions were analyzed as one region in 1991 and as two in 1992.

[2] The central region was analyzed unitemporally in 1991.

[3] The Arkansas area sampling frame was reconstructed for 1992.

[4] Direct expansion was used.

[5] Direct expansion was used in strata which are not listed.

[DE] Direct expansion method--no ancillary satellite data used.

[TM] Method using regression estimator with satellite data where possible and direct expansion where not.

ACKNOWLEDGEMENTS

I would like to thank Paul Cook of USDA NASS for his 1991 analysis of central Arkansas.

REFERENCES

Allen, J.D., 1990a, A Look at the Remote Sensing Applications Program of the National Agricultural Statistics Service, Journal of Official Statistics, 6(4):393-409.

Allen, J.D., 1990b, Remote Sensor Comparison for Crop Area Estimation Using Multitemporal Data, in Proceedings of the IGARSS '90 Symposium, College Park, Md., pp. 609-612.

Allen, J.D. and Hanuschak, G.A., 1988, The Remote Sensing Applications Program of the National Agricultural Statistics Service: 1980-1987, U.S. Department of Agriculture, NASS Staff Report No. SRB-88-08.

Bellow, M.E. and Graham, M.L., 1992, Improved Crop Area Estimation in the Mississippi Delta Region Using Landsat TM Data, in Proceedings of the ASPRS/ACSM Convention, Washington, D.C.

Bellow, M.E. and Ozga, M., 1991, Evaluation of Clustering Techniques for Crop Area Estimation using Remotely Sensed Data, in American Statistical Association 1991 Proceedings of the Section on Survey Research Methods, Atlanta, Ga., pp. 466-471.

Cochran, W.G., 1977, Sampling Techniques, John Wiley and Sons, New York, NY, ch. 7, pp 189-203.

Cook, P.W., 1982, Landsat Registration Methodology Used by U.S. Department of Agriculture's Statistical Reporting Service 1972-1982, USDA/NASS/Remote Sensing Section.

Congalton, R.G., 1991, A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data, Remote Sensing of the Environment, 37:35-46 (1991).

Cotter, J. and Nealon, J., 1987, Area Frame Design for Agricultural Surveys, U.S. Department of Agriculture, NASS Area Frame Section.

Gong, P. and Howarth, P.J., 1992, Frequency-Based Contextual Classification and Grey-Level Vector Reduction for Land-Use Identification, Photogrammetric Engineering and Remote Sensing, Vol. 58, No. 4, April 1992, pp 423-437.

Johnson, R.A. and Wichern, D.W., 1988, Applied Multivariate Statistical Analysis, Prentice Hall, Englewood Cliffs, N.J., ch. 11, pp. 501-513.

Ozga, M., Mason, W.W. and Craig, M.E., 1992, PEDITOR - Current Status and Improvements, in Proceedings of the ASPRS/ACSM Convention, Washington, D.C.

U.S. Department of Agriculture, 1992, Crop Production - 1991 Summary, Agricultural Statistics Board, NASS.

APPLICATION OF SATELLITE DATA TO CROP AREA ESTIMATION AT THE COUNTY LEVEL

Michael E. Bellow, USDA/NASS
Research Division, 3251 Old Lee Highway, Room 305, Fairfax, VA 22030

KEY WORDS: Battese-Fuller model, county effect, combined ratio estimator

I. INTRODUCTION

The National Agricultural Statistics Service (NASS) of the United States Department of Agriculture has published county estimates of crop acreage, crop production, crop yield and livestock inventories since 1917. These estimates assist the agricultural community in local decision making and are also useful to agribusinesses. The primary source of data for agricultural commodity estimates has always been surveys of farmers, ranchers and agribusinesses who voluntarily provide information on a confidential basis. However, surveys designed and conducted at the national and state levels are often inadequate for producing reliable information at the county or small domain level. Therefore, supplementary data sources such as NASS list frame control data, previous year estimates and Census of Agriculture data are often used to improve county estimation. Earth resources satellite data represents a useful ancillary data source for county level estimation of crop planted and harvested area. The basis for improved estimation accuracy using satellite data is the fact that, with adequate coverage, all of the area within a county can be classified to a crop or ground cover type. The accuracy of the estimates depends upon how accurately the satellite data are classified to each crop.

NASS has used or considered several regression based estimators for small area crop acreage estimation with ancillary satellite data. These estimators use stratum level counts of pixels classified to crops. From 1976 to 1982, NASS used the Huddleston-Ray estimator (Huddleston and Ray, 1976). In 1978, the Cardenas family of estimators (Cardenas, Blanchard and Craig, 1978) was considered but not adopted. From 1982-87, the Agency used the Battese-Fuller estimator (Battese, Harter and Fuller, 1988) for county level estimation of major crops in the Midwestern grain belt with Landsat Multispectral Scanner (MSS) data. The same method was used to calculate county estimates of rice, cotton and soybeans in the Mississippi Delta region in 1991-92 with Landsat Thematic Mapper (TM) data. Research has recently begun to consider non-regression estimators based on overall (across strata) counts of classified pixels. This report discusses two such estimators and compares them with the Battese-Fuller estimator.

Graham (1993) provides a description of the methodology used to obtain classified pixel counts and generate state and regional level crop acreage estimates. Some knowledge of those concepts is helpful in the upcoming discussion.

II. BATTESE-FULLER ESTIMATOR

The Battese-Fuller approach to crop area estimation at the county level is an extension of the regression methodology used for state level estimation. The Battese-Fuller estimator (BFE) utilizes the analysis district (multi-county) level regression, but incorporates an additional term that accounts for county (random) effects.

The Battese-Fuller model was first developed in the general framework of linear models with nested error structure (Fuller and Battese, 1973), and later applied to the special case of county crop area estimation (Battese, Harter and Fuller, 1988). In state level estimation, a group of counties and parts of counties covered by one or more satellite scenes comprises an analysis district. Analysts compute regression relationships between NASS survey reported acreages and counts of classified pixels, using area frame sample units (segments) within each analysis district. The Battese-Fuller model assumes that segments grouped by county have the same slope relationship with classified pixels as the analysis district, but the intercept term is different. One can apply the model within an analysis district for any land use stratum where a valid regression relationship has been found. The analyst computes stratum level Battese-Fuller area estimates for all counties and subcounties within each analysis district. For land use strata where regression is not feasible due to lack of adequate satellite coverage or too few segments, a domain indirect synthetic estimator is used.

For a given analysis district, the strata where regression is done are here referred to as regression strata and the remaining ones as synthetic strata. For convenience, the regression strata are labelled $h=1,\ldots,H_r$ and the synthetic 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. If a given county is partially contained in the analysis district, then the estimation formulas given below apply only to the included portion.

For each sample segment within a given stratum h in county c, the Battese-Fuller model specifies the following relation:

$$y_{hci} = \beta_{0h}+\beta_{1h}x_{hci}+\nu_{hc}+\epsilon_{hci},\ i=1,\ldots,n_{hc}$$

where:

n_{hc} = number of sample segments in stratum h, county c

y_{hci} = reported acreage of crop of interest in stratum h, county c, sample segment i

x_{hci} = number of pixels classified to crop of interest in stratum h, county c, sample segment i

ν_{hc} = county (random) effect for stratum h, county c

ϵ_{hci} = random error in stratum h, county c, sample segment i

β_{0h}, β_{1h} = analysis district level regression parameters for stratum h

The county effect and random error are assumed to be independent and normal, with mean zero and variances σ^2_{vh} and σ^2_{eh}, respectively. The random errors for segments within the district are assumed to be mutually independent. The county mean residuals are observable and given by:

$$\bar{u}_{hc.} = \bar{y}_{hc.} - \hat{\beta}_{0h} - \hat{\beta}_{1h}\bar{x}_{hc.}$$

where:

$$\bar{y}_{hc.} = (1/n_{hc}) \sum_{i=1}^{n_{hc}} y_{hci}$$

$$\bar{x}_{hc.} = (1/n_{hc}) \sum_{i=1}^{n_{hc}} x_{hci}$$

$\hat{\beta}_{0h}$, $\hat{\beta}_{1h}$ = least squares regression parameter estimators for stratum h

For a given county, the stratum level mean crop area per population unit (segment) is estimated by:

$$\bar{y}_{(BF),hc.} = \hat{\beta}_{0h} + \hat{\beta}_{1h}\bar{x}_{hc} + \delta_{hc}\bar{u}_{hc.}$$

where:

$\bar{x}_{hc}$ = mean number of pixels per population unit classified to crop in stratum h, county c

$0 \le \delta_{hc} \le 1$

The range of allowed values of the parameter δ_{hc} defines a family of Battese-Fuller estimators. If δ_{hc}=0, then the estimate lies on the analysis district regression line for the stratum. The value commonly used is the one that minimizes the mean square error for stratum h in county c (Walker and Sigman, 1982):

$$\delta^*_{hc} = n_{hc}\sigma_{vh}^2/(n_{hc}\sigma_{vh}^2+\sigma_{eh}^2)$$

In general, the variance components σ_{vh}^2 and σ_{eh}^2 are unknown and must be estimated. The Appendix gives estimators that are a special case of the unbiased estimators derived by Fuller and Battese (1973), using the "fitting-of-constants" method. They require that a given stratum contain at least two sample segments within the county in question; otherwise δ_{hc} is set to zero in the computation of the Battese-Fuller estimate.

The (unadjusted) stratum level estimator of total crop area in county c is:

$$\hat{T}_{(uBF),hc} = N_{hc}[\hat{\beta}_{0h} + \hat{\beta}_{1h}\bar{x}_{hc} + \delta_{hc}\bar{u}_{hc.}]$$

where:

N_{hc} = number of population units in stratum h, county c

The county estimates are often adjusted to sum to the district totals obtained in state level regression estimation. The adjusted stratum level Battese-Fuller estimator is:

$$\hat{T}_{(aBF),hc} = \hat{T}_{(uBF),hc} - (N_{hc}/N_h)\sum_{c=1}^{C} \delta_{hc}\bar{u}_{hc.}$$

where:

N_h = number of population units in stratum h

C = number of counties in analysis district

The adjusted Battese-Fuller estimator of total crop area in the regression strata of county c is:

$$\hat{T}_{(aBF),c} = \sum_{h=1}^{H_r} \hat{T}_{(aBF),hc}$$

Estimation of the variance of the BFE is described by Walker and Sigman (1982). Their estimator of mean square error, used to derive the variance estimator, is known to have a downward bias due to estimation of the variance components. A correction due to Prasad and Rao (1990) may be implemented in the future.

As mentioned previously, synthetic estimation is done in strata where regression is not viable. Since a county usually contains few segments in a given stratum, the stratum level sample mean crop acreage over the entire analysis district is used to compute a synthetic estimate. The estimate of crop area in synthetic stratum h, county c is:

$$\hat{T}_{(SYN),hc.} = N_{hc}\bar{y}_{h..}$$

where:

$\bar{y}_{h..}$ = mean reported crop area per sample segment in stratum h

The domain indirect synthetic estimator of total crop area in the synthetic strata of county c is then:

$$\hat{T}_{(SYN),c} = \sum_{h=H_r+1}^{H} \hat{T}_{(SYN),hc}.$$

with estimated variance:

$$\hat{\sigma}^2[\hat{T}_{(SYN),c}] = \sum_{h=H_r+1}^{H} N_{hc}^2 s_{yh}^2 (N_h - n_h)/N_h n_h$$

where:

$$s_{yh}^2 = (1/n_h - 1) \sum_{i=1}^{n_h} \sum_{c=1}^{C} (y_{hci} - \bar{y}_{h..})^2$$

The final county estimate is obtained by summing the regression and synthetic components:

$$\hat{T}_c = \hat{T}_{(BF),c} + \hat{T}_{(SYN),c}$$

The estimated variance of the final county estimate is computed by summing the variance estimates of the regression and synthetic components. The use of the analysis district level average to estimate county totals ignores county effects, so the synthetic component of a county estimate can have a significant bias.

Walker and Sigman (1982) studied the Battese-Fuller model using Landsat MSS data over a six county region in eastern South Dakota. At that time, NASS was using the Huddleston-Ray estimator (Huddleston and Ray, 1976), which simply replaced the analysis district level pixel mean in each stratum with the county level pixel mean in the regression equation. The county effect parameter of the Battese-Fuller model was highly significant for corn, the most prevalent in the region of the four crops considered. The study showed robustness of the Battese-Fuller family against departure from certain model assumptions, and provided the justification for replacing the Huddleston-Ray estimator with the Battese-Fuller estimator for operational county crop estimation.

III. PIXEL COUNT ESTIMATORS

As improved satellite sensors enable higher classification accuracy, the overall (across strata) count of pixels within an area classified to a given crop or cover type becomes more interesting. The overall pixel count represents a census of pixels covering the area in question and therefore is not subject to sampling error. However, there is a nonsampling error due to pixel misclassification. As a result, the overall pixel count (converted to area units) is generally a biased estimator of crop area. Adjustment factors based on sample level information can reduce the bias. Although a pixel count estimator could be a function of counts of pixels classified to many different cover types, this discussion will be restricted to estimators based on the number of pixels classified to the crop of interest only. A general expression for such an estimator is:

$$\hat{T}_c = \eta X_c$$

where:

X_c = number of pixels classified to crop of interest in county c

η = adjustment term

The adjustment term may be a function of the sample level classification data. The choice of adjustment term determines the specific estimator used. If the term is simply set to the area on the ground corresponding to one pixel, then the Raw Pixel Count Estimator (RPCE) is obtained:

$$\hat{T}_c^{(RPC)} = \lambda X_c$$

where λ is the conversion factor (area units per pixel) for the satellite sensor being used.

The RPCE is biased if the theoretical commission error (probability that a pixel classified to the crop of interest is from another cover type) and omission error (probability that a pixel from the crop of interest is classified to another cover type) are not equal. The combined ratio estimator (CRE), based on the estimator of the same name described in Cochran (1977), attempts to adjust for the bias. This estimator is conceptually simple, uses stratum level information to compute the adjustment term and has a readily available formula for estimating the variance. The CRE can be expressed as follows:

$$\hat{T}_c^{(CR)} = [(\sum_{h=1}^{H} N_h \bar{y}_{h..})/(\sum_{h=1}^{H} N_h \bar{x}_{h..})] X_c$$

$$= \hat{R} X_c$$

An estimator for the variance of the combined ratio estimator is derived from Cochran's population variance formula, valid for large samples:

$$\hat{\sigma}^2[\hat{T}_c^{(CR)}] = [X_c/X]^2 \sum_{h=1}^{H} [(N_h^2(1-f_h)/n_h](s_{yh}^2 + \hat{R}^2 s_{xh}^2 - 2\hat{R} s_{xyh})$$

where:

$$s_{xh}^2 = (1/n_h - 1) \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_{h..})^2$$

$$s_{yh}^2 = (1/n_h - 1)\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_{h..})^2$$

$$s_{xyh} = (1/n_h - 1)\sum_{i=1}^{n_h}(x_{hi} - \bar{x}_{h..})(y_{hi} - \bar{y}_{h..})$$

$f_h = n_h/N_h$

y_{hi} = reported area of crop of interest in stratum h, sample segment i

x_{hi} = number of pixels classified to crop of interest in stratum h, sample segment i

$\bar{x}_{h..}$ = mean number of pixels per sample segment classified to crop of interest in stratum h

X = total number of pixels classified to crop of interest

IV. EMPIRICAL EVALUATION

This section describes an empirical evaluation of the satellite based county crop area estimators described above, performed using data from Iowa and Mississippi. The Iowa data were from a 1988 research project, while the Mississippi data were from NASS's 1991 operational project in the Mississippi Delta region (Bellow and Graham, 1992). The quantity estimated was acreage planted to a crop.

The first application area is a nine county region in western Iowa with a high concentration of corn and soybeans. Ground data from NASS's 1988 June Agricultural Survey (JAS) were used for estimation, with a total sample size of 30 segments from two strata. The region was covered by one TM scene with an image date of July 25, 1988. The second area, a twelve county region in northwestern Mississippi, comprises two contiguous crop reporting districts that accounted for most of the state's cotton and rice production in 1991. Ground data from the 1991 JAS were used for estimation, involving 73 segments in four strata for cotton and 59 segments in two strata for rice. The analysis used multitemporal satellite data with image dates of April 1 and August 23, 1991. Two TM scenes from each date were needed to cover the region. For both regions, all seven spectral bands from each scene were utilized. The adjusted version of the Battese-Fuller estimator was computed in all cases.

For Iowa, the analysis used 30 segments, with 28 coming from stratum A (agricultural) and the other two from stratum B (agri-urban). Data from the segments in stratum A were used for the BFE, which was computed within the subset of that stratum covered by the TM scene. Parts of Calhoun, Crawford and Ida counties lay outside the TM scene. For the BFE, CRE and RPCE, synthetic estimation was applied within stratum A for the areas outside the scene. For the BFE, synthetic estimation was used in stratum B for all areas.

The strata in Mississippi where Battese-Fuller estimation was used for cotton were strata A (75-100% cultivated), B (51-75%), C (15-50%) and D (0-15%). The BFE was applied only in strata A and B for rice. Synthetic estimation was used in the other strata for each crop. The TM scenes covered all areas except for a small part of Yazoo county.

Tables 1 and 2 give the computed values of the satellite based BFE, CRE and RPCE for Iowa and Mississippi, respectively. For comparison, the survey based estimate (SYN) obtained by using synthetic estimation in all strata is also shown. Estimated standard deviations are given for the SYN, BFE and CRE. The official county planted acreage estimates issued by NASS's Iowa and Mississippi State Statistical Offices are also listed. These published estimates are based on additional survey and administrative data. The official county figures for Iowa are believed to be highly accurate indicators of corn and soybean acreage. Rice figures are not given for Issaquena, Quitman and Yazoo counties since Mississippi did not issue official rice estimates for those counties in 1991. Tables 3 and 4 give measures of estimator accuracy for the two states, computed based on the final official figures. The mean deviation (MD), root mean square deviation (RMSD), mean absolute deviation (MAD) and largest absolute deviation (LAD) are shown.

Comparing the standard deviations of SYN, BFE and CRE given in Table 1, it is seen that CRE had the lowest value for both corn and soybeans in all Iowa counties considered. BFE had lower variance than SYN in all counties for corn and all but one county for soybeans. Table 2 shows that in Mississippi, CRE had lower variance than BFE in eight of twelve counties for cotton and eight of nine counties for rice. For both cotton and rice, SYN had higher variance than BFE and CRE in each county.

Table 3 shows that for corn in Iowa, BFE had the lowest MAD and RMSD among the four estimators studied. However, RPCE had the lowest RMSD and MAD for soybeans. From Table 4, BFE showed the lowest MAD and RMSD for cotton in Mississippi, but CRE had the lowest MAD and RMSD for rice. For all four crops, the survey based estimator SYN showed the highest values of RMSD, MAD and LAD and is therefore clearly inferior to the other three estimators. The mixed results suggest that the relative performance of the three satellite based estimators may depend to a large degree on the specific crop. The mean deviation of BFE was negative for all four crops, suggesting a possible downward bias of this estimator.

V. SUMMARY

This paper described the current status of satellite based county crop area estimation in

NASS. The Battese-Fuller model is currently applied to compute county acreage indications provided to certain NASS State Statistical Offices. Estimators based on overall pixel counts have recently begun to receive attention. Empirical results for Iowa and Mississippi suggest that the CRE has lower variance than the BFE, while relative performance of estimators appears to be crop specific. The BFE and CRE both showed a negative bias in the study. Future research will explore properties of these estimators for different crops and other regions.

REFERENCES

Battese, G.E., Harter, R.M. and Fuller, W.A., (1988) "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data," Journal of the American Statistical Association, vol. 83, no. 401, pp. 28-36.

Bellow, M.E. and Graham, M.L. (1992) "Improved Crop Area Estimation in the Mississippi Delta Region using Landsat TM Data," Proceedings of the ASPRS/ACSM/RT92 Convention, Washington, D.C., vol. 4, pp. 423-432.

Cardenas, M., Blanchard, M.M. and Craig, M.E., (1978) "On The Development of Small Area Estimators Using LANDSAT Data as Auxiliary Information," Economic, Statistics, and Cooperatives Service, U.S. Department of Agriculture.

Cochran, W.G., (1977) "Sampling Techniques," New York, NY: John Wiley & Sons, pp. 165-167.

Fuller, W.A. and Battese, G.E. (1973) "Transformations for Estimation of Linear Models with Nested-Error Structure," Journal of the American Statistical Association, vol. 68, no. 343, pp. 626-632.

Graham, M.L. (1993) "State Level Crop Area Estimation using Satellite Data in a Regression Estimator," Proceedings of the International Conference on Establishment Surveys, Buffalo, NY.

Huddleston, H.F. and Ray, R. (1976) "A New Approach to Small Area Crop Acreage Estimation," Annual Meeting of the American Agricultural Economics Association, State College, PA.

Prasad, N.G.N. and Rao, J.N.K. (1990) "The Estimation of the Mean Squared Error of Small-Area Estimators," Journal of the American Statistical Association, vol. 85, no. 409, pp. 163-171.

Walker, G. and Sigman, R. (1982) "The Use of LANDSAT for County Estimates of Crop Areas - Evaluation of the Huddleston-Ray and Battese-Fuller Estimators," SRS Staff Report No. AGES 820909, U.S. Department of Agriculture.

APPENDIX. ESTIMATION OF BATTESE-FULLER VARIANCE COMPONENTS

The estimators of the Battese-Fuller variance components at the analysis district level represent a special case of the more general unbiased estimators derived by Fuller and Battese (1973). The variance component estimators are as follows:

$$\hat{\sigma}_{eh}^2 = [1/(n_h-C-1)] \times \sum_{c=1}^{C} \sum_{i=1}^{n_{hc}} [y_{hci}-\bar{y}_{hc.}-\hat{\alpha}_h(x_{hci}-\bar{x}_{hc.})]^2$$

$$\hat{\sigma}_{vh}^2 = \max[0,(s_{uh}^2-(n_h-2)\hat{\sigma}_{eh}^2)/(n_h-T_h)]$$

where:

$$\hat{\alpha}_h = \frac{\sum_{c=1}^{C} \sum_{i=1}^{n_{hc}} (x_{hci}-\bar{x}_{hc.})(y_{hci}-\bar{y}_{hc.})}{\sum_{c=1}^{C} \sum_{i=1}^{n_{hc}} (x_{hci}-\bar{x}_{hc.})^2}$$

$$s_{uh}^2 = \sum_{c=1}^{C} \sum_{i=1}^{n_{hc}} (y_{hci}-\hat{\beta}_{0h}-\hat{\beta}_{1h}x_{hci})^2$$

$$T_h = \frac{n_h \sum_{c=1}^{C} n_{hc}^2 \bar{x}_{hc.}^2 + (\sum_{c=1}^{C} n_{hc}^2)(\sum_{c=1}^{C} \sum_{i=1}^{n_{hc}} x_{hci}^2) - Q_h}{(n_h \sum_{c=1}^{C} \sum_{i=1}^{n_{hc}} x_{hci}^2) - n_h^2 \bar{x}_{h..}^2}$$

$$Q_h = 2n_h\bar{x}_{h..} \sum_{c=1}^{C} n_{hc}^2 \bar{x}_{hc.}$$

The value of the quantity δ_{hc} that minimizes the mean square error of the Battese-Fuller estimator can then be estimated by:

$$\hat{\delta}_{hc}^* = n_{hc}\hat{\sigma}_{vh}^2/(n_{hc}\hat{\sigma}_{vh}^2 + \hat{\sigma}_{eh}^2)$$

Walker and Sigman (1982) provide expressions for the mean square error and mean square conditional bias of the stratum level Battese-Fuller estimator. Separate formulas are required depending upon whether the regression parameters are known or estimated. Variance estimators are derived from these formulas.

Table 1: County Estimates for Iowa 1988 (1000 Acres)

CORN:

County	Official	SYN	SD	BFE	SD	CRE	SD	RPCE
Audubon	100.0	112.4	6.5	92.2	3.2	93.6	2.1	100.6
Calhoun	133.0	144.9	8.3	133.2	3.9	134.4	2.9	144.2
Carroll	141.0	146.2	8.4	141.4	4.5	142.1	3.1	152.6
Crawford	147.0	183.2	10.6	152.7	4.7	155.1	3.2	164.9
Greene	125.0	145.9	8.4	130.0	3.9	132.8	2.9	142.7
Guthrie	98.0	151.3	8.7	106.3	5.2	107.8	2.4	115.8
Ida	112.0	111.4	6.4	107.0	4.0	107.0	3.8	110.3
Sac	136.0	148.1	8.5	138.3	4.0	139.6	3.1	150.0
Shelby	155.0	149.4	8.6	140.7	4.0	141.5	3.1	152.1

SOYBEANS:

County	Official	SYN	SD	BFE	SD	CRE	SD	RPCE
Audubon	70.7	74.0	7.5	69.9	4.6	70.4	2.1	74.8
Calhoun	150.0	95.4	9.6	145.0	5.8	136.9	4.0	145.2
Carroll	117.0	96.1	9.7	106.7	9.7	106.4	3.1	113.0
Crawford	106.0	120.4	12.1	106.9	5.8	108.1	3.1	113.8
Greene	143.0	96.1	9.7	117.5	5.4	109.6	3.2	116.3
Guthrie	77.5	99.5	10.0	64.4	7.0	78.8	2.3	83.7
Ida	75.2	73.3	7.4	76.4	5.3	76.1	4.3	78.2
Sac	124.0	97.3	9.8	112.9	5.5	108.8	3.2	115.5
Shelby	94.9	98.3	9.9	81.0	6.0	91.1	2.7	96.7

Table 2: County Estimates for Mississippi 1991 (1000 Acres)

COTTON:

County	Official	SYN	SD	BFE	SD	CRE	SD	RPCE
Bolivar	65.5	106.2	15.4	61.6	6.1	60.6	3.9	80.6
Coahoma	105.7	59.2	8.4	88.3	4.2	82.6	5.2	109.8
Humphrey	61.6	53.2	7.2	57.3	3.4	54.2	3.4	72.1
Issaquena	38.0	42.6	8.6	34.6	3.9	27.5	1.8	36.6
Leflore	79.2	68.8	9.6	87.8	3.5	83.4	5.3	111.0
Quitman	31.0	48.1	7.2	46.4	4.0	44.5	2.8	59.3
Sharkey	47.0	43.2	6.9	48.6	3.4	42.5	2.7	56.6
Sunflower	100.0	95.6	15.0	79.3	5.5	73.9	4.7	98.3
Tallahatchie	64.2	68.9	10.5	67.9	4.9	60.3	3.8	80.3
Tunica	45.6	47.1	6.9	38.0	2.5	36.5	2.3	48.6
Washington	95.7	84.4	11.6	102.4	4.0	93.2	5.9	124.1
Yazoo	94.5	89.3	23.4	93.9	7.5	81.9	5.2	108.9

RICE:

County	Official	SYN	SD	BFE	SD	CRE	SD	RPCE
Bolivar	74.0	50.8	11.9	66.2	3.6	66.9	6.1	60.9
Coahoma	15.8	20.3	4.7	10.4	2.5	10.7	1.0	9.7
Humphreys	3.6	22.8	5.2	7.1	2.3	4.7	0.4	4.3
Leflore	16.6	30.7	7.1	19.4	3.6	17.3	1.6	15.8
Sharkey	5.0	18.0	4.1	7.8	1.7	6.5	0.6	5.9
Sunflower	36.0	51.1	12.0	37.8	3.5	36.7	3.4	33.4
Tallahatchie	9.6	20.9	5.1	8.5	3.0	8.1	0.7	7.4
Tunica	17.5	17.6	4.3	9.9	2.6	13.0	1.2	11.9
Washington	30.5	39.6	9.0	22.6	3.5	28.0	2.6	25.4

Table 3: Iowa Estimator Accuracy

	CORN				SOYBEANS			
EST	MD	RMSD	MAD	LAD	MD	RMSD	MAD	LAD
BFE	-0.6	6.8	5.4	14.3	-8.6	11.9	9.1	25.5
RPCE	9.6	12.6	10.6	17.9	-2.3	10.3	7.4	26.7
CRE	0.8	7.4	6.3	13.5	-8.0	13.5	9.0	33.4
SYN	16.2	23.8	17.6	53.3	-12.0	28.0	21.6	54.6

Table 4: Mississippi Estimator Accuracy

	COTTON				RICE			
EST	MD	RMSD	MAD	LAD	MD	RMSD	MAD	LAD
BFE	-1.8	10.0	7.8	20.7	-2.1	5.2	4.5	7.9
RPCE	13.2	17.2	13.7	31.8	-3.8	5.6	4.1	13.1
CRE	-7.2	12.5	10.2	26.1	-1.9	3.5	2.7	7.1
SYN	-1.8	19.4	13.2	46.5	7.0	13.9	12.2	23.2

A SURVEY SAMPLE DESIGN TO ESTIMATE NUMBERS AND YIELD OF FRUIT TREES

by
Adnan M. Haider
Department of mathematics
Howard University
Washington D.C. 20059

KEY WORDS: Strata, enumeration, ratio, random.

ABSTRACT:

There are various methods of planting fruit trees and various methods of yield collection. The need for a sample survey design(SSD) to obtain efficient estimates of numbers of trees and their yield is an existing open problem.

This paper presents a procedure for construction of an effective SSD. The procedure uses concepts of" Region Random Selection" , "Multiple Stratification", and " Simple Random Sampling " to construct the SSD. The procedure utilizes existing information from agricultural census.

The procedure derive random strata to construct a frame for the SSD. The frame covers primary sampling units (PSUs) of a selected region of a given a proportion size. The region is considered a representative portion of the nations farms planted by the same fruit trees.

The paper closes with an application of the SSD.

THE PROCEDURE:

The procedure starts by a selection of a suitable representative region of predetermined proportion of size α% of the national area planted in the specific fruit trees.

Farms of the region are used to define the region's PSUs. The PSU is defined to contain a fixed number of fruit trees as reported by the existing agricultural census.

The PSU's areas and their identification numbers (IDs) also are reported by the agriculture census.

The key step in the construction procedure of the SSD is the derivation of a set of random ratios "RR" that represent the numbers of fruit trees per unit area of the PSUs of the region.

The establishment of a one-one correspondence between the region's PSUs (or their given IDs) and random elements of the derived set of "RR" is the essential step in the applicability of the proposed SSD.

The domain of the elements of the set of "RR" has a positive range. This range ought to be divided into a suitable number (between 10-and 15) of equal sized intervals.

These intervals cover all random elements of the set of "RR". Thus, by applying the one-one concept between these elements and the PSUs IDs we can distribute the IDs of the PSUs(or the region's PSUs) in randomly among the defined strata (intervals).

By following the above path we obtain two distinguished sets of random independent disjoined strata.

The first set covers the random "RR" derived values for the region's PSUs, and the second set covers the randomly distributed(located) ID's for the region's PSUs.

For each stratum from the first set, there exists a corresponding stratum in the second set that relate to the first. The relationship is defined in terms of the 1-1 correspondence between elements of the first stratum(say the one covers the IDs of the region's PSUs that are randomly located) and the corresponding elements of the second stratum (say of the one covers the IDs of the region's PSUs that are randomly located).

Accordingly, we can define a frame for the SSD that covers above distinguished strata .

The source of information needed to define the above frame is the agricultural census. Such information includes the PSUs IDs numbers, the PSUs areas and the recorded numbers of fruit trees included in defining the PSUs of the region.

We need to define the following set of symbols,

N: represents the number of PSUs in the region.

N_j: the size of j-th strata (number of PSUs in the j-th strata)

k: number of available strata.

$$\text{and,} \quad N = \sum_{j=1}^{k} N_j$$

Let,

n_1: be the sample size to be selected in the first stage of sampling where, n_1 is the efficient sample size of size α% (ESS)(Haider, 1992).

where,

$$n_1 = \alpha_1\% \text{ x } N.$$

$$n_{1,j} = \alpha_1\% \text{ x } N_j$$

$$j=1,..,k.$$

$$\text{for} \quad n_1 = \sum_{j=1}^{k} n_{1,j}$$

The n_1 PSUs selected in the first stage of sampling will be subjected to a 100 % enumeration.

n_2: sample size to be selected in the second stratified sampling.(using the same principle as that implemented in selecting the first stage sample.

$n_{2,j}$: number of PSUs selected from $n_{1,j}$ PSUs of the first stage.

and, for $j=1,\ldots,k$,

$$n_{2,j} = \alpha_2\% \text{ x } n_1.$$

$m_{.j}$: represents the number of trees randomly selected the third stage of sampling for harvesting from the j-th strata.

$m_{i,j}$: represents the number of harvested trees from the i-th PSU of the j-th strata selected in the third stage of sampling.

$$\text{where,} \quad m_{.j} = \sum_{i=1}^{n_{2,j}} m_{i,j}$$

$e_{i,j}$: represents the enumerated number of fruit trees in the i-th PSU from the j-th stratum.

$r_{i,j}$: represents the number of fruit trees as reported to be in the PSU from the j-th strata, as reported by the agriculture census.
for

$$m_{i,j} = \alpha_2\% \text{ x } e_{i,j}$$

and

$$m_{.j} = \alpha_2 \sum_{i=1}^{n_{2,j}} e_{i,j}$$

Let,

M: to represent the number of fruit trees to be harvested in the third stage of sampling.

$$\text{where,} \quad M = \sum_{j=1}^{k} m_{.j}$$

$Y_{i,j,s}$: to represent the yield of the s-th tree from the i-th PSU from the j-th stratum.

$$Y_{i,j} = (1/m_{i,j}) \sum_{s=1}^{m_{i,j}} Y_{i,j,s}$$

$E_{i,j}$: to represent the "RR" ratio of enumerated number of fruit trees per unit area of the i-th PSU in the j-th stratum.

$R_{i,j}$: represents the "RR"ratio of reported number of fruit trees per unit area for the i-th PSU of the j-th stratum.

$P_{i,j}$: the probability of selecting (with replacement) the i-th PSU from the j-th stratum.
Then, for

$$Z_j = E_j / E \; . \; P_j$$

and

$$V_j = R_j / R \; P_j$$

for

$$E_j = \sum_{i=1}^{n_{1,j}} E_{i,j}$$

$$\text{and} \quad E = \sum_{j=1}^{k} E_j$$

$$R_j = \sum_{i=1}^{n_{1,j}} R_{i,j}$$

$$\text{and} \quad R = \sum_{i=1}^{k} R_j$$

$$P_j = \sum_{i=1}^{n_{1,j}} P_{i,j}$$

A_N: the area of the region.

Define,

$$W_{n1} = \sum_{j=1}^{k} Z_j \; / \; \sum_{j=1}^{k} V_j$$

define,

$$\bar{R} = R / n_1 ,$$

An estimate of the mean number of trees per unit area of the PSUs of the region is given by :

$$\bar{Y}_{n1} = W_{n1} \; \bar{R}$$

Also the estimate of total number of fruit trees $T_N i$, in the region is given by:

$$\hat{T}_N = A_N \; \bar{Y}_{n1}$$

with ,

$$\hat{\text{Var}}(\bar{Y}_{n1}) = 1/n_1(n_1-1) \sum_{J=1}^{k} (Z_j - W_{n1} \; V_j)^2$$

and

$$\hat{\text{Var}}(\hat{T}) = A_N^2 \; \hat{\text{Var}}(\bar{Y}_{n1})$$

The estimate $\hat{TT}$ of the national number of fruit trees is given by,

$$\hat{TT} = I.F \times \hat{T}_N ,$$

for an inflation factor I.F-defined by

$$I.F. = 100/ \alpha_1$$

YIELD ESTIMATE:

To obtain an estimate of the national production of fruit trees, we may accomplish it by harvesting "M" random trees selected randomly in the third stage where M represents the minimum efficient sample size MESS(Haider, 1987).

The "M" random trees are selected from the n_2 PSUs of the second stage in proportion to the PSUs accumulated RR.

The procedure is to select $m_{.j}$ trees for harvesting from the j-th strata, that are in proportion to $e_{.j}$,

$$\text{where, } e_{.j} = \sum_{i=1}^{n_{2,j}} e_{i,j}$$

let,

$$Z_{j1} = EE_{.j} / EE,$$

where

$EE_{i,j}$: the random values of the i-th selected PSU from the j-th stratum calculated by using the enumerated numbers for only n_2 PSUs of the region, from which some trees have harvested.

$$EE_{.j} = \sum_{i=1}^{m_{i,j}} EE_{i,j}$$

$$EE = \sum_{j=1}^{k} EE_{.j}$$

and, for

$RR_{i,j}$: the RR using the reported numbers of trees provided by agriculture census. The PSUs are the same n_2 PSUs that provided trees for harvesting.

$$RR_{.j} = \sum_{i=1}^{m_{i,j}} RR_{i,j}$$

$$RR = \sum_{i=1}^{k} RR_{.j}$$

$$V_{j1} = RR_{.j} / RR .$$

Define,

$$U_{i,j} = (V_{j1} / Z_{j1}) Y_{i,j} ,$$

$$\bar{U} = (1/n_2) \sum_{j=1}^{k} \sum_{i=1}^{m_{.j}} U_{i,j}$$

the estimate for Y the mean yield per tree of the region is given by,

$$\hat{\bar{Y}} = \bar{U}$$

with a variance estimate

$$Var(U) = [1/n_2(n_2-1)] \times \sum_{j=1}^{k} \sum_{i=1}^{m_{i,j}} (U_{i,j}(m_{i,j}) - U_j)^2$$

and, the estimate of the total yield production "P" in the region is given by,

$$\hat{P} = \hat{T}_N \bar{U}$$

where,

$$\hat{Var}(P) =$$

$$\hat{T}_N^2 \hat{Var}(U) + \bar{U}^2 \hat{Var}^2(T) .$$

It is important to notice that if $Y_{i,j}$ is expressed in pounds, then $\hat{P}$, is also expressed in pounds, and if $Y_{i,j}$ is expressed in item numbers then so is $\hat{P}$.

THE NATION'S ESTIMATE OF FRUIT PRODUCTION:

By using the $\hat{P}$ estimate to obtain the national estimate NP for the production of fruit, which is:

$$\hat{NP} = \hat{P} / (\alpha_1) \times 100$$

APPLICATION:

Buhroze is a village in the northeast of Iraq a well known for the production of oranges. A sample survey was conducted to estimate the total number and the production of the orange trees for the State.

The village represents only (five per cent of the state areas planted by orange trees.

The village consists of 450 holdings as reported by the past census of agriculture. Also, the (reported) numbers of trees of oranges was supplied by the holders. However, in the actual sample survey, count of the trees may be made in a sample of holdings.

The holdings were grouped to form the primary sampling unit (PSUs), each consisted of at most 650 orange trees as reported by the agricultural census. If a holding contained more than 650 fruit trees, it formed a separate PSU, without being partitioned. There were 250 PSUs in all the village. Only twenty-five PSUs (α_2 %). Sampling were conducted with replacement.

THE FIELD WORK:

The counting of trees in the sample PSUs was carried out during the harvesting time. The third stage of sampling, was carried out to select orange trees to be harvested. (the number was 110 orange trees selected in a systematic random selection) from the third stratified sample of optimal sample of size 25.

THE RANDOM SELECTION OF ORANGE:

For the i-th PSU randomly selected PSU(in the third stage of sampling), we pick a first orange tree randomly, then, we continue systematically to pick every k-th orange tree of orange trees for harvesting.

RESULTS:

The estimates of the total numbers of orange trees and the total fruit production for total number of orange had been obtained.

CONCLUSIONS:

This paper presents a procedure to estimates the number and the yield of fruit trees in the country. The applicability of the procedure does not require highly trained enumerators and expertise to complete the task. The procedure gives the most efficient sample size to be selected for estimation.

BIBLIOGRAPHY:

Curran, P.S. and Adway, N.(1982)."Landsat data, its availability and suitability to monitoring the density of data palm trees in Saudi Arabia." The First Symposium on Date trees. Saudi Arabia.1982.

Cochran, William G. (1977)."Sampling techniques." 3rd edition. Wiley and Sons Publishing Company. 1977.

Haider, Adnan M.(1992b)."Haider's Extension of the Maximum Likelihood Estimation Procedure and Incomplete Multivariate Data". The Proceedings of the Annual Symposium of the American Statistical Association. Boston, Mass. U.S.A.

Haider, Adnan M. (1987)."Derivations of optimal sample sizes for efficient sample survey designs." Technical report, The American Institute for Technical Technology Transfer." Washington, D.C. 1992.

Haider, Adnan M.(1986). "Efficient estimate of yield and numbers of orange trees in Iraq. Project number 420, The Central Statistical Organization, Baghdad, Iraq. 1986.

Kish, Leslie."Survey sampling." Published by Wiley and sons Company. 1965.'

A COMPARISON OF FOUR ALTERNATIVE WEIGHTED ESTIMATORS TO THE OPEN ESTIMATOR FOR USE IN THE AGRICULTURAL LABOR SURVEY

Cheryl L. Turner, USDA/OASS
200 N. High, Room 608, Columbus, OH 43215

KEY WORDS

June Agricultural Survey, Agricultural Labor Survey, non-overlap, open estimator, weighted estimator

INTRODUCTION

The National Agricultural Statistics Service (NASS) within the United States Department of Agriculture (USDA) annually conducts a June Agricultural Survey (JAS). The JAS is a multiple frame survey, consisting of both a list frame and an area frame. The area frame is stratified according to land usage or the percent of cultivation. The area frame is further subdivided into overlap (OL) and non-overlap (NOL) domains. The overlap portion of the area frame is composed of farming operations which are also found on the list frame. The non-overlap contains those farming operations which are not found on the list frame.

The JAS begins the survey year and is the largest survey of the year for NASS. Follow-on survey samples are derived from a list sampling frame and a sample of the area frame. The Agricultural Labor Survey (ALS) is a multiple frame follow-on survey. It provides estimates of the number of farm workers and of the wage rates paid to those farm workers. Currently, the non-overlap estimate for the ALS is derived using an open estimator. The ALS open estimator is based on a sample of NOL Resident Farm Operators (RFO's) from forty percent of the area segments used in the JAS. (A segment is a piece of land that is the primary sampling unit in the NASS area frame sampling plan.) By definition, the open estimator excludes all non-Resident Farm Operators. An alternative to the open estimator is a weighted estimator. The weighted estimator is generated from a sample of all NOL farm operators, both RFO and non-RFO. The weighted estimator has historically had a smaller coefficient of variation (CV) than the open estimator because the weighted estimate is generated from a larger group of farm operators.

Four weighted estimators were evaluated for possible use in the ALS. They were the operational, modified weighted (modified), Hanuschak-Keough strata mean (H-K mean), and the Hanuschak-Keough strata median (H-K median). Each weighted estimator was compared against the current open estimator.

This report represents the comparative analysis done on these alternative weighted estimators. All estimators used the "peak number of hired workers" from 1991 JAS data. The JAS area questionnaire obtains the expected "peak number of hired workers" for the survey year. This number is then used to define the NOL strata for the follow-on ALS. This study was done independently on both the 17 labor regions and the eleven monthly and seasonal states.

STUDY DESIGN

Data for this survey were collected during the 1991 JAS and represent the NOL domain. The item of interest was the peak number of hired agricultural workers for the survey year. The data were evaluated at the regional level and at the state level (for the eleven monthly and seasonal states). There are 17 labor regions within the United States. They are defined as follows:

Region

Northeast I:
Connecticut, Maine, Massachusetts, New Hampshire, New York, Rhode Island, Vermont
Northeast II:
Delaware, Maryland, New Jersey, Pennsylvania
Appalachian I:
North Carolina, Virginia
Appalachian II:
Kentucky, Tennessee, West Virginia

Southeast:
Alabama, Georgia, South Carolina
Lake:
Michigan, Minnesota, Wisconsin
Cornbelt I:
Illinois, Indiana, Ohio
Cornbelt II:
Iowa, Missouri
Delta:
Arkansas, Louisiana, Mississippi
Northern Plains:
Kansas, Nebraska, North Dakota, South Dakota
Southern Plains:
Oklahoma, Texas
Mountain I:
Idaho, Montana, Wyoming
Mountain II:
Colorado, Nevada, Utah
Mountain III:
Arizona, New Mexico
Pacific:
Oregon, Washington
Florida:
**Florida
California:
**California

** Note that Florida and California are single state regions.

The monthly states are California, Florida, New Mexico, and Texas. Michigan, New York, North Carolina, Oregon, Pennsylvania, Washington, and Wisconsin are the seasonal states.

THE WEIGHTED ESTIMATORS

Two types of estimators were being evaluated, an open and a weighted estimator. For an open estimator, the location of the operator's residence is used to uniquely associate every farm with only one segment. A weight of one is assigned if the tract operator lives within the selected segment (if the tract operator is an RFO), and a weight of zero is assigned otherwise. Conversely, the weighted estimator apportions a farm's activities to a segment by weighing the data relative to the fraction of the farm's acreage that lies within the segment boundary. Therefore, one farm may contribute to the data in several segments.

As stated earlier, the ALS open estimator is based on a sample of NOL RFO's from forty percent of the area segments used in the JAS.

In contrast, an ALS weighted estimator would be based on the same sample size being selected from all NOL operations (both RFO's and non-RFO's) from the same forty percent of area segments. The respondents selected using an ALS weighted estimator would have been selected from a larger pool of potential respondents. In sampling from the larger pool of respondents, there is the potential for a reduction in the CV.

The operational, H-K mean, H-K median, and the modified weighted were the four weighted estimators being evaluated.

Operational

The operational weighted estimator is the weighted estimator traditionally used in NASS surveys. It merely assigns an "operational" weight of tract acres divided by total farm acres for each farming operation even partially contained within the segment. (Where the tract acres are the acres residing within a sampled segment.) This estimator prorates farm level data to the segment level.

Hanuschak-Keough strata mean and median

These two weighted estimators are similar to the operational weighted estimator, but they attempt to limit potential outliers by controlling the value of the weight. There are occasions when the exact farm acreage is neither obtainable nor known. This happens when the respondent either would not or could not give the correct farm acreage. In these instances the tract acreage and farm acreage may be recorded as equal (plus perhaps a token acre for the farmstead) on the JAS. Although this problem has been recognized and emphasized at training schools, it still exists (but to a lesser degree). Hanuschak and Keough proposed a solution for this specific type of problem. In some cases the equality of the tract and farm acres is accurate. However, if the farm acres should have been substantially larger than the tract acres, the "operational" weight would be nearly or equal to one when it should have been considerably lower. This problem leads to a great

overexpansion of the survey data. And conversely, there could be an underexpansion of the survey data if tract acres were underreported.

Hanuschak and Keough recommended a more robust estimator than the standard "operational" weight. A robust estimator is relatively insensitive to slight departures from the assumptions of normality. The Hanuschak-Keough estimators replaced the "operational" weight with a robust weight for all NOL tracts (or observations) in which someone other than the operator or the operator's spouse responded. The Hanuschak-Keough estimators will guard against large overexpansions or underexpansions of the survey data. Consider the following respondent codes as defined in the JAS survey:

Respondent Code
1 = Operator/Manager
2 = Spouse
3 = Other
4 = Observed Refusal
5 = Observed Non-refusal

The Hanuschak-Keough estimators replaced the "operational" weight for all NOL observations containing respondent code 3, 4, or 5 with a more robust weight. Within each land use strata, the Hanuschak-Keough strata mean estimator replaced the denominator of the "operational" weight for those observations containing respondent codes 3, 4, or 5 with the average farm acreage from the respondent code 1 and 2 observations. The Hanuschak-Keough strata median estimator replaced the denominator within each land use strata for those same observations with the median farm acreage from the respondent code 1 and 2 observations.

Modified Weighted

The modified estimator was originally proposed by Bosecker and Clark. It is an effort to eliminate screening for farm operators in densely populated segments. In reducing the amount of survey screening, the cost of conducting the survey is greatly reduced.

The modified estimator is especially suited to the measurement of rare populations, and the number of farm operators among the general population (particularly in residential areas) certainly qualifies as rare. The modified weighted estimator will exclude up to one half acre for non-agricultural land devoted to residential purposes (such as the house and yard). For residential agricultural tracts, the residential area would be subtracted from the weight's numerator and denominator; for non-resident agricultural tracts, the residential area would be subtracted just from the weight's denominator. Since the modified weight would be zero for small tracts consisting of only a house and yard, screening for farm operators in residential areas would be unnecessary. The modified weight assumed 1/2 acre for all residences, except where it was known that the farmstead was less than 1/2 acre.

The expanded peak number of hired workers was calculated using the open estimator and each of the alternative weighted estimators.

ANALYSIS

NOL estimates were generated for the peak number of hired workers. Both the open and the weighted estimators were generated using the same number of tracts and the same tract information. Identical analyses were used to independently compare each of the four alternative estimates with the current open estimate of the peak number of hired workers. Univariate paired t-tests were conducted at the regional level for the 17 regions and at the state level for the eleven monthly and seasonal states on each alternative estimator versus the open estimator. These t-tests will determine if the alternative estimate was significantly different from the open estimate. The paired t-test will test the following hypotheses for each alternative estimate:

H_0: $Y_{diff} = 0$ versus H_A: $Y_{diff} <> 0$

where Y_{diff} = alternative estimate - open estimate

RESULTS

Univariate paired t-tests were performed on the variable peak number of hired workers. The t statistics were calculated for both the 17 labor

regions and the eleven monthly and seasonal states for each of the four weighted estimates versus the open estimate.

Labor Region Results

The test results indicated that most of the comparisons yielded insignificant differences (alpha = .05) at the regional level. Therefore, there were negligible differences between each of the four alternative estimators and the open estimator for these regions.

The test results also indicated that some significant differences (alpha = .05) did exist at the regional level. Significant differences between each of the four alternative estimates and the open estimate existed in the Delta region and the Southern Plains region. In the Appalachian II region and the Southeast region, significant differences existed for all comparisons but the H-K mean estimate and the open estimate. Significant differences existed in the Pacific region between each the operational and modified estimates and the open estimate. And lastly, the Northern Plains and California regions obtained significant differences between the H-K median estimate and the open estimate.

As stated above, both the Delta and Southern Plains regions obtained significantly different results for the four alternative estimators as compared to the open estimate. Further examination of these two regions shows that Arkansas, Louisiana, and Texas were the dominating states within their respective regions. All states were significantly different with respect to the alternative estimate vs. the open estimate. When Arkansas, Louisiana, and Texas were evaluated individually, one tract often accounted for the majority of difference between the alternative estimates and the open estimate.

For example, within Texas there was one tract which made no contribution to the peak number of hired workers for the open estimate. But for each of the four alternative weighted estimates, this tract alone contributed between four and eight percent of Texas' state level expansion for the peak number of hired workers. The differences in these estimates were due in part to the farmer living outside of the selected segment (and therefore having an open weight of 0), while at the same time having a positive number of hired workers.

In following with previous findings, the open estimate was the lowest estimate (due to a downward bias) in 12 of the 17 regions, while the H-K median was the highest estimate in 11 of the 17 regions. The operational, H-K mean, and modified estimates were most often found between these two extremes.

The CV for the open estimator was the largest CV in 13 of the 17 regions. This supports the notion that sampling from a smaller sample size (only the RFO's) will increase the CV. The CV's for the four weighted estimators were (overall) considerably smaller than those for the open estimator, but none of the alternatives distinguished itself as having the lowest CV.

State Level Results

Mostly insignificant differences (alpha = .05) also existed at the state level. And as with the regional level results, this indicated that there were negligible differences between each of the four alternative estimators and the open estimator for the monthly and seasonal states.

The test results at the state level also indicated that some significant differences (alpha = .05) did exist. Significant differences between all four of the alternative estimates and the open estimate existed only in Texas. There were significant differences in Washington between the operational estimate and the open estimate and also between the modified weighted estimate and the open estimate. A significant difference also existed between the H-K median estimate and the open estimate in California.

Also, as with the regional results, the estimates were lowest for the open estimator in 7 of the 11 states and the estimates were highest for the H-K median estimator in 8 of the 11 states. The operational, H-K mean, and modified estimators were barely distinguishable from each other, each lying between the two extremes. And, again the open estimator CV as the largest CV in 7 of the 11 states. The four weighted estimator CV's again obtained smaller CV's than the open CV, while not substantially differing from one another.

CONCLUSIONS AND RECOMMENDATIONS

This paper evaluated four alternative weighted estimators (the operational, Hanuschak-Keough strata mean, Hanuschak-Keough strata median, and the modified operational) of the peak number of hired workers and compared them to the current open estimator approach. These evaluations were made at both the labor region and state level. When considering only the estimates and their corresponding CV's, it was evident that the open estimate was biased downward, while at the same time having an increased CV. This indicated that there was a need for a "better" estimator with a smaller CV.

The analyses indicated that, for the most part, insignificant differences existed between the open estimator and any of the four alternative weighted estimators. However, significant differences were also found. The Delta and Southern Plains regions were both significantly different for all four comparisons. Further review of these two regions indicated that one state within the region was primarily responsible for the significant difference. And, in reviewing that state, one (or several) tracts accounted for a substantial percentage of the estimation difference. This indicated that one (or several) tracts within a state could make a region (or state) significantly different.

When there was no significant difference between the alternative and the open estimate, any of the weighted estimators could be considered as a viable selection. Each of the alternative weighted estimators has a smaller CV than the open estimator. But the H-K median estimator also has a strong upward bias, which greatly overestimates the peak number of hired workers. This upward bias negates the H-K median as an adequate alternative to the open estimator. When selecting between the remaining weighted estimators, significant differences were considered. Of the three remaining alternative weighted estimators, more research is recommended on the Hanuschak-Keough strata mean. While the original prognosis on the H-K mean was positive, this is the first study done utilizing this estimator and more positive results are needed before a conclusion can be reached. The operational estimator is a tried and proven estimator. It had a smaller CV than the open estimator and also improved upon the downward bias of the open estimator. But the recommended alternative is the modified weighted. This estimator achieved the accuracy levels of the operational estimate, while also eliminating the JAS screening for farmers in the more densely populated segments, and thus reduced the overall survey cost. More research is also recommended on a combined estimator based on the modified weighted estimator and the H-K mean. This new combined estimator would merge the strong points of both estimators. It would reduce the screening requirements for potential farm operators within residential areas while, at the same time, lessening the effect of any potential outliers.

YIELD MODELS FOR CORN AND SOYBEANS BASED ON SURVEY DATA

Thomas R. Birkett, National Agricultural Statistics Service, USDA
USDA/NASS, Rm.4813-South, Washington, D.C. 20250

Abstract

The National Agricultural Statistics Service uses survey data to forecast yields for major agricultural commodities, including corn and soybeans. The survey data contains variables that become the independent variables in linear forecasting models. This paper describes the forecasting models, showing what the key survey variables are and examining how they are related to final yield.

Introduction

The National Agricultural Statistics Service (NASS), an agency of the United States Department of Agriculture, conducts monthly field surveys in the late summer and fall to forecast corn and soybean yields. Summarized data from the survey forms the independent variables for a statistical model that predicts the current season final average yield. The survey data include variables correlated with the final average number of ears or pods that will be harvested, along with variables correlated with the final average grain weight per ear or weight per pod. This paper gives a short description of these variables and how they are used to forecast final average yield.

Description of the Objective Yield Surveys

In June, NASS conducts a very large survey of agricultural land use in the U.S. to estimate the current season's acreage planted to corn and soybeans. From the base generated by this survey, NASS draws a random sample of corn and soybean plots. This is done through a two stage process, in which fields are selected and then random locations are designated within each selected field. The procedure is carried out so that a simple random sample is obtained, and each planted acre of corn or soybeans has an equal chance of being included in the sample. This simple random sample property is an important assumption for the statistical models to be applied to the survey data.

The randomly located plots are a few square feet in area. Within the plots, enumerators count and measure variables that are positively correlated with final yield. Among the variables collected for soybeans are number of plants per acre, number of nodes per plant, number of lateral branches per plant, number of blooms, dried flowers and pods per plant, and number of pods with beans per plant. For corn the NASS enumerators count the number of stalks per acre, number of stalks with ears, number of ear shoots, and number of ears with kernels per acre. They also husk a random sample of ears near the plot and measure the length of a typical kernel row on each ear. Just prior to farmer harvest of the corn or soybean field in which the sample is located, the enumerator harvests the plot and obtains the final yield. The same sample plots are revisited each month starting in August until farmer harvest.

Samples are laid out in all the major corn and soybean producing states. Data are collected during the period from the 21st of the previous month until the first of the month. Starting in August and continuing through November, around the 10th of each month the USDA releases yield estimates for each state based on the survey.

Variables in the Regional Models

The best relationship between the survey data and final yield is found at the regional level, the region being the set of states in the survey. Consequently, the plot level data is summarized to the state and then to the region level, where it is modeled against the region yield. Each monthly regional model normally has one independent variable X.

The form of the regional linear model is either

$$Y = \alpha + \beta X + \epsilon$$

or

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \epsilon$$

where

Y = average regional yield and

α, β's are unknown model parameters.

X is the known independent variable, and

ϵ is the difference between Y and its expected value.

In the examples used in this paper, the soybean model has the quadratic term while the corn model is limited to the linear term.

The values for X for corn and soybeans are shown in the following tables.

SOYBEAN VARIABLES BY MONTH	
August	estimated number of lateral branches per acre
September	Estimated number of pods with beans per acre
October-December	(estimated number of pods per acre) X (net weight per pod)

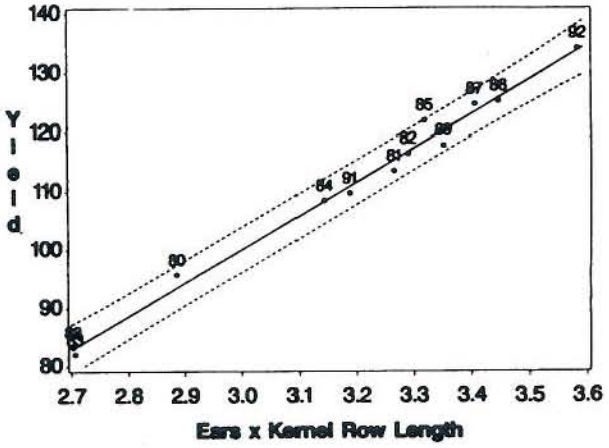
CORN VARIABLES BY MONTH	
August	(stalks with ears + ears with kernels per acre) X (average kernel row length per ear)
September	(Ears with kernels per acre) X (average kernel row length per ear)
October-December	(Ears with kernels per acre) X (average grain weight per ear)

Maturity Adjustment

While NASS conducts the survey during the last ten days of each month, the overall maturity of the crop at that time will vary from year to year, depending on when it was planted, subsequent weather, etc. The forecasting power of the model is enhanced by classifying each plot by stage of maturity and limiting the independent variable calculations to data from pre-selected stages. This adjustment allows the independent variables to be more comparable across years. Variables not used directly in X (such as nodes and blooms, dried flowers and pods) are used for maturity classification. Consequently, the predictor variable is not a function of all the data, but only those plots in a stage that has exhibited good predictive power for final yield. This criteria normally means the exclusion of very immature samples in the first month of the survey. After that the vast majority of the samples are used directly in X.

A plot of the data in the September 1 corn regional model is shown below. (The digits plotted represent the years 1980-1992).

September 1 Corn Model

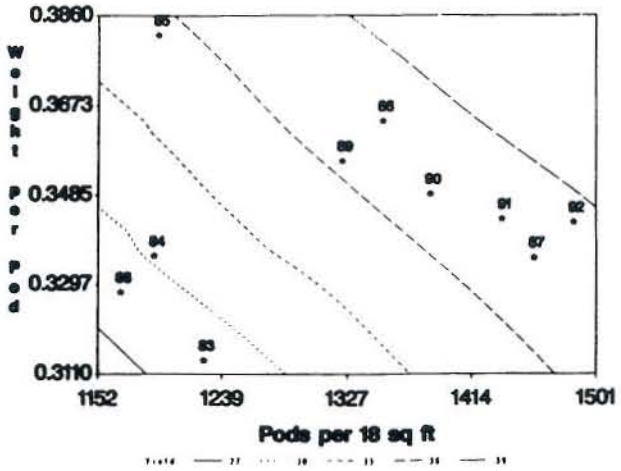


Relationship of the number and weight variables to final average yield

As mentioned above, the survey variables are selected to correlate with the components of final yield, which are number of ears or pods and weight per ear or pod. It is quite illuminating to view the 3-dimensional distribution of final yield and the factors of the independent variables to see how they explain the yield

level. Since the independent variable in the model can usually be factored into the product of a variable correlated with final weight and one correlated with final number, we can plot the fitted model surface over the weight X number plane. The projection of selected levels of the fitted yield surface onto this plane is easier to analyze. An October example for soybeans and a November one for corn are shown below.

Soybeans - October 1, 1983-1992



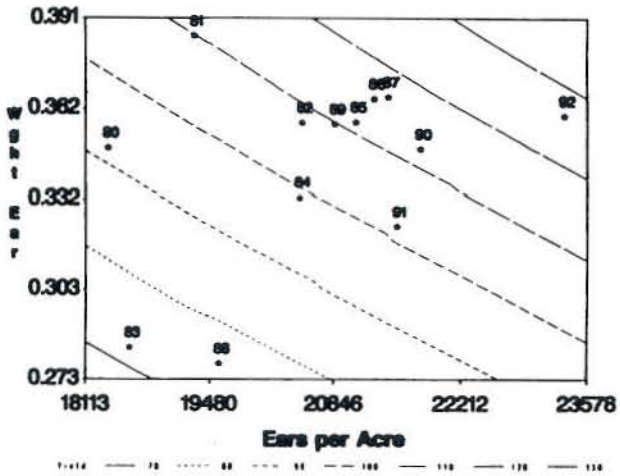
For soybeans, the weight per pod is in grams, and the yield contours projected from the fitted model surface onto the plane are 27, 30, 33, 36 and 39 bushels per acre. On October 1, usually about half the crop has been harvested, and the weight is for just those harvested samples. The pods per 18 square feet is for all samples as of October 1.

This graph contains a great deal of information about soybean yields. The years divide into two distinct

groups, with 1983, 1984 and 1988 in the lower left corner, and the remaining years distributed along the 36 to 39 bushel contour region. The years 1983, 1984 and 1988 were severe drought years in the corn belt, and both pod counts and weight were depressed to the point that yields averaged around 30 bushels. In the remaining years, conditions were more normal, and average yields were generally around 36 to 39. So far there has not been a year where weight and numbers of pods were simultaneously near record levels. There is an obvious negative correlation between average weight and number. The two variables interact inversely with each other to produce approximately the same yield, even though the weight and number variables are varying quite widely. The heaviest average weight occurred in 1985, but it had drought-like numbers of pods. At the other extreme, 1987 had the lowest October 1 weight of the normal years, but its pod counts were the second highest. 1992, which has the record yield to date, had the highest number of pods on October 1.

Since the surface is based on a model with a quadratic term, one can see the spacing between the contours increases as the yield level increases. This implies that there are diminishing increases in yield as the average weight and numbers increase. Also, since the contours are at roughly 45 degree angles, one can deduce that increases in weight or numbers will increase yield. However, this is survey data, and numbers and weight do not vary independently (they vary inversely) so an increase in one will normally be associated with a decrease in the other and vice versa.

Corn - November 1, 1980 - 1992



For corn, usually about two-thirds of the crop is harvested by November 1. The grain weight, in pounds, is just for the harvested samples. The ear counts are for all samples as of November 1. The projected yield contours from the fitted surface are 78, 88, 98, 108, 118, 128 and 138 bushels per acre.

Here we see the two drought years, 1983 and 1988, in the lower left corner. There appears to be less dependence between the weight and number variables for corn than there was with soybeans. Some years,

such as 1985, 1986 and 1987 are pushing the limit on both ears and weight. In 1992, ear density increased dramatically, while the ear weights maintained an average level for non-drought years. 1992 set a new record for yield by a large margin, driven by the large ear counts.

Since the corn model has no quadratic term, the spacing between the contour levels is constant. The 45 degree contours indicate both weight and numbers drive final average yield. If conditions are generally good, it is possible to have both large ear counts and above average weights in the same year, something that is not generally seen with soybeans.

Conclusion

Average corn and soybeans yields can be predicted by observing variables that are correlated with final numbers and weights. In corn both counts and weights can be high at the same time, producing record yields. With soybeans, however, final counts and weights are inversely related, producing relatively constant average yields in non-drought years.

References

Birkett, T.R. (1990), "The New Objective Yield Models for Corn and Soybeans", National Agricultural Statistics Service, SMB-90-02, Washington, DC, 20250.

SAS Institute Inc., SAS/GRAPH Software: Reference, Version 6, First Edition, Volume 1 and 2, Cary, NC: SAS Institute, Inc., 1990.

Searle, S.R. (1971), Linear Models, New York: John Wiley & Sons, Inc.

The author can be contacted at

NASS/USDA, Room 4813,
14th and Independence, SW,
Washington,DC, 20250

202-720-5359

USING DIFFERENT PRECIPITATION TERMS TO FORECAST CORN AND SOYBEAN YIELDS

M. Denice McCormick
USDA/NASS/Research Division/3251 Old Lee Hwy., Room 305, Fairfax, VA. 22030

KEY WORDS: Precipitation, regression models

INTRODUCTION

In 1990, the National Agricultural Statistics Service (NASS) introduced new models to forecast yield for corn and soybeans on the regional and State levels in a plan to phase out the older, less accurate models (Birkett 1990). An annual survey collects data from randomly selected sample plots in randomly selected fields. The old regression models predicted the components of yield such as number of pods per plant and weight per pod at the plot level based on five years of previous data. Plot level data were then aggregated to the State level. The new models are also regression models, and have initially been developed to predict yield directly rather than the components of yield using survey data aggregated to the regional level. Regions are constructed from the set all States that participate in the annual survey. A longer period of years in the historic data set must be used since only one data point is used to represent each year.

McCormick and Birkett (1992) tried to improve the accuracy of early season soybean yield forecasts by adding a term that represented total accumulated precipitation throughout the growing season from April 1 until the forecast date at a six-State regional level. The analysis indicated that soybean forecast accuracy at the regional level was not improved using this particular term. Based upon this result, two recommendations were made. One was to evaluate alternative time frame terms, such as monthly precipitation totals. The other was to use them to forecast other major agricultural crop yields. This paper reports results when separate monthly precipitation terms were added to corn and soybean yield forecast models. It considers data for thirteen years, 1980 to 1992. The soybean States included in the study are Arkansas, Illinois, Indiana, Iowa, Missouri, Minnesota, Nebraska, and Ohio. The corn States are Illinois, Indiana, Iowa, Michigan, Minnesota, Missouri, Nebraska, Ohio, South Dakota, and Wisconsin. The performance of each model is compared to official operational model performance.

This study evaluates multiple regression models which use precipitation and survey variables to forecast end-of-season crop yields. In previous research, the models showed improved performance using aggregated survey variables at the regional level. Therefore, this method was also used to aggregate the precipitation variables.

DATA

Precipitation Data

Precipitation variables used in the models represent total precipitation for a particular month at the regional level. The data are provided from a network of National Weather Service weather stations in each State. The variable is constructed as follows:

$$P_t = \frac{\sum_{s=1}^{S} A_{ts} R_{ts}}{\sum_{s=1}^{S} A_{ts}}, \qquad (1)$$

where

P_t = the average total precipitation within selected month for the region for year t,
S = the number of States covered,
A_{ts} = the acres for harvest for year t, State s, and
R_{ts} = the average total precipitation within selected month for year t, State s,

where

$$R_{ts} = \frac{\sum_{d=1}^{D_s} A_{tsd} E_{tsd}}{\sum_{d=1}^{D_s} A_{tsd}},$$

A_{tsd} = the acres for harvest for year t, State s, district d, and
D_s = the number of districts per State s,
E_{tsd} = the average station total precipitation within selected month for year t, State s, district d,

$$E_{tsd} = \frac{1}{W_{tsd}} \sum_{w=1}^{W_{tsd}} U_{tsdw}$$

where

W_{tsd} = number of weather stations for year t, State s, district d, and

U_{tsdw} = total precipitation within selected month for year t, State s, district d, weather station w.

Survey Data

The construction of the independent variables for the regional regression models for both soybeans and corn is discussed by Birkett (1990, 1993). For soybeans for the month of August, the independent variable (Z_t) is the estimated number of lateral branches per eighteen square feet. For September, the independent variable is the estimated number pods with beans per eighteen square feet. These regional-level estimates for soybeans are constructed as follows:

$$Z_t = \frac{\sum_{s=1}^{S} A_{ts} F_{ts}}{\sum_{s=1}^{S} A_{ts}} \qquad (2)$$

where

A_{ts} = the acres for harvest for year t, State s, and

F_{ts} = number of lateral branches per 18 sq. feet year t, State s,

$$F_{ts} = \frac{1}{m_{ts}} \sum_{j=1}^{m_{ts}} B_{tsj} L_{tsj} ,$$

where

m_{ts} = the number of samples in J_{ts} year t, State s,

J_{ts} = the subset of samples classified in maturity categories 2-6 (or 1-6 in southern States), year t, State s,

B_{tsj} = plants per 18 square feet for year t, State s, sample j,

L_{tsj} = lateral branches per plant year t, State s, sample j (for August) and

= estimated pods with beans per plant per 18 sq. feet, year t, State s, sample j (for September).

Corn independent variables (Z_t) are more complex as they are a function of both plant counts and average kernel row length per square foot. C_{ts} is substituted for F_{ts} in equation (2). In August, it is calculated as:

$$C_{ts} = \frac{1}{m_{ts}} \sum_{j=1}^{m_{ts}} (U_{tsj} + V_{tsj}) \bar{K}_{tsj},$$

where

C_{ts} = a function of the number of stalks with ears, the number of ears with kernels, and the average kernel row length per square foot,

U_{tsj} = number of stalks with ears per sq. ft., year t, State s, sample j,

V_{tsj} = number of ears with kernels per sq. ft., year t, State s, sample j, and

$\bar{K}_{tsj}$ = the average kernel row length per ear, year t, State s, sample j.

In September, C_{ts} is calculated as:

$$C_{ts} = \frac{1}{m_{ts}} \sum_{j=1}^{m_{ts}} (V_{tsj}) \bar{K}_{tsj} .$$

For both forecasts, data are used from the subset of samples in maturity categories 3-6 for year t, State s.

Yield Data

The regional yield values included in this study were calculated as follows:

$$Y_t = \frac{\sum_{s=1}^{S} A_{ts} Y_{ts}}{\sum_{s=1}^{S} A_{ts}} , \qquad (3)$$

where

Y_t = final regional yield for year t, and

Y_{ts} = NASS State yield year t, State s.

METHODOLOGY

Regression analysis was used to evaluate the performance of precipitation data in combination with survey data. Multiple linear regression models with associated diagnostics for model fit and forecast

accuracy were examined. The basic regression models analyzed were:

1: $Y_t = \beta_o + \beta_1 Z_t + \epsilon_t$

2: $Y_t = \beta_o + \beta_1 Z_t + \beta_2 Z_t^2 + \epsilon_t$

3: $Y_t = \beta_o + \beta_1 Z_t + \beta_2 P_t + \epsilon_t$

4: $Y_t = \beta_o + \beta_1 Z_t + \beta_2 Z_t^2 + \beta_3 P_t + \epsilon_t$.

Model 2 is the official model used by NASS to forecast August corn and soybeans and September soybeans. However, Model 1 is the official model used to forecast September corn. Models 3 and 4 use one monthly precipitation term. Analysis was conducted to determine which month from the growing season provided optimal forecasting capability. Also, models with multiple monthly precipitation terms were examined.

Model Evaluation Criteria

The primary model evaluation criterium is the set of prediction intervals (PI) for the minimum, median, and maximum yielding years over 13 years in the study. For soybeans, these years were 1988, 1981 and 1990, and for corn, they were 1983, 1989 and 1992, respectively. A second criterium is the adjusted coefficient of determination, R_a^2 which provides a measure of correspondence between predicted and actual yields. Both the PI and R_a^2 are based on the sum of squared differences from the least squares analysis used to derive the model parameters.

1. The prediction interval (PI) refers to half the confidence interval length for the predicted value of a future Y for a given future year o. That is, at the α significance level,

$$PI = t(1-\frac{\alpha}{2};n-1-p)SD(\hat{Y}_o),$$

where

$$SD(\hat{Y}_o)=s[(x_o'(X_o'X_o)^{-1}x_o) + 1]^{\frac{1}{2}},$$

s = (residual MSE)$^{1/2}$,
x_o = relevant p-dimensional row vector of independent variables for year o (for example, in Model 3: p= 3,
x_o = [1, Z_o, P_o]),
X_o = relevant (n-1 x p) matrix of independent variables (excludes x_o),
n = number of years, and
p = number of parameters.

The X_o matrix excludes the row vector x_o, so that the PI reflects the accuracy expected in an operational model where current year data are not included in the model development. A significance level of 0.32 was used for this study, which provides t values near 1.0. Consequently, the future Y will fall within the calculated PI of the predicted Y approximately 68% of the time.

2. R_a^2 is used as a goodness-of-fit test for each model with an adjustment made for the corresponding degrees of freedom (Draper and Smith 1981).

R_a^2 is calculated as:

$$R_a^2 = 1-\frac{(RSS_p)/(n-p)}{(CTSS)/(n-1)},$$

where

RSS_p = the residual sum of squares taking the changing number of parameters into account,
CTSS = the corrected total sum of squares,
n = the number of years, and
p = the number of parameters.

Outlier Identification

Since the purpose of the models is to make forecasts, the rstudent statistic (also called the studentized residual) was used to help identify outliers to be excluded from the model. This statistic was recommended in Belsley, Kuh and Welsh (1980). It is similar to the standardized residual:

$$r_{si} = \frac{r_i}{s\sqrt{1-h_i}},$$

where

r_i = i^{th} residual,
s = (residual MSE)$^{1/2}$, and
h_i = $x_i'(X'X)^{-1}x_i$.

Here, s is replaced by s(i). S(i) is the estimate of σ with the i^{th} observation deleted. In a forecasting model, rstudent measures how many prediction standard errors the forecast is from the observed Y. Observations with absolute values of rstudent greater than 3.0 were identified as outliers. The rstudent statistic is distributed closely to the t-distribution with n-p-1 degrees of freedom.

RESULTS

Regression analysis was conducted on a number of different models using different monthly precipitation terms. Tables 1 and 2 present the prediction intervals and R_a^2 for the official linear or quadratic model using survey data only and then results adding the optimal monthly precipitation term. In both tables, the prediction intervals relate to the years with minimum, median, and maximum regional yields.

Table 1: August Results

Model	R_a^2	Prediction Intervals		
		min	med	max
CORN:				
OFFICIAL	.87	7.0	5.7	6.2
P_t=JULY	.93	5.4	4.3	4.9
SOYBEANS:				
OFFICIAL	.70	2.8	2.3	2.7
P_t=JULY	.74	2.3	2.1	2.3

Note: August corn: both models have outlier year 1988 removed.

Table 2: September Results

Model	R_a^2	Prediction Intervals		
		min	med	max
CORN:				
OFFICIAL	.97	3.6	3.2	3.4
P_t=JUNE	.98	2.3	2.0	2.2
SOYBEANS:				
OFFICIAL	.89	1.7	1.6	1.6
P_t=AUGUST	.88	1.9	1.7	1.9

Note: September corn: Official model removed 1990; Precip model removed 1988.

CONCLUSIONS

Except for the September soybean forecast, the precipitation models performed better than the official forecast models since their prediction intervals were consistently smaller. Contrary to previous indications, the August forecast models demonstrated that the addition of a monthly precipitation term with a survey term does improve forecasts for both crops. For both periods, the corn forecast seemed to benefit the greatest. There is no evidence that a change from the official model is warranted for September soybeans.

BIBLIOGRAPHY

Belsley, David A, Kuh, Edwin, Welsh, R.E., (1980), Regression Diagnostics, John Wiley & Sons.

Birkett, Thomas R., (1990) "The New Objective Yield Models for Corn and Soybeans", SMB Staff Report Number SMB-90-02, USDA.

Birkett, Thomas R., (1993) "Yield Models for Corn and Soybeans Based on Survey Data", USDA, Proceedings, ICES Conference, Buffalo, New York.

Draper, N.R., Smith, H., (1981), Applied Regression Analysis, John Wiley & Sons Second Edition.

McCormick, M. Denice, Birkett, Thomas R. (1992) "Evaluating the Addition of Weather Data to Survey Data to Forecast Soybean Yields", SRB Research Report No. SRB 92-11, USDA.