United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Research Division

SRB Research Report
Number SRB-93-10

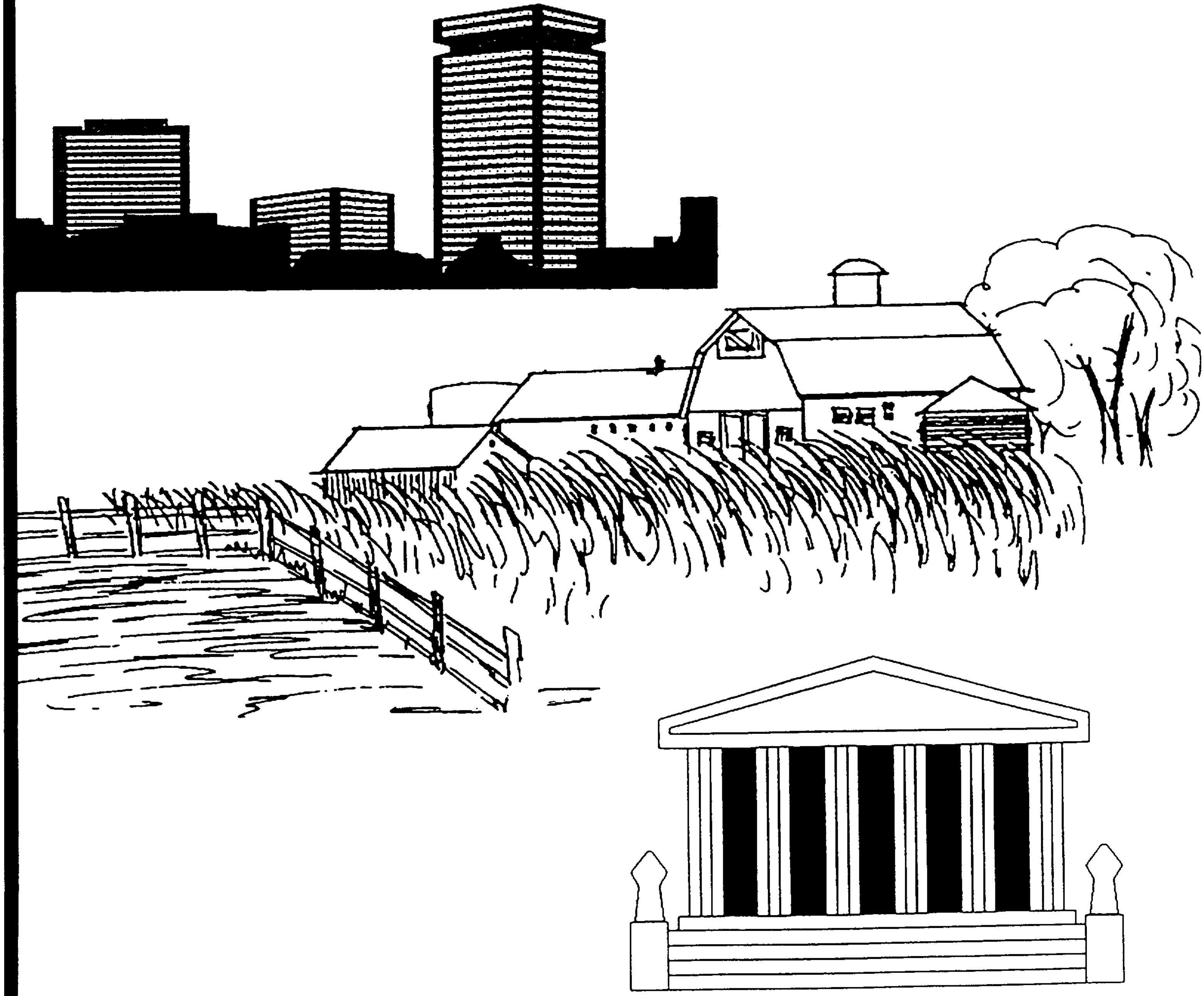
September 1993

SURVEY METHODS FOR BUSINESSES, FARMS, AND INSTITUTIONS

INTERNATIONAL CONFERENCE ON ESTABLISHMENT SURVEYS

Part I

NASS Participants



SURVEY METHODS FOR BUSINESSES, FARMS, AND INSTITUTIONS by Participants, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, DC 20250-2000, September 1993, Part 1 of 2, Report No. SRB-93-10.

ABSTRACT

Part 1 of this report is a compilation of all contributed papers presented at the International Conference on Establishment Surveys in Buffalo, New York, June 28-30, 1993. They have been organized by general subject matter. Several of these will be printed as separate and more detailed research reports. Part 2 of this report will include monograph and invited papers.

This paper was prepared for limited distribution to the research community outside the U.S. Department of Agriculture. The views expressed herein are not necessarily those of NASS or USDA.

ACKNOWLEDGMENTS

This report has been the culmination of the efforts of many individuals in NASS. Special thanks go to Yolanda Grant and Shawna McClain for their many hours of assistance in preparing for the conference and this report.

TABLE OF CONTENTS

I. NEW TECHNOLOGY

II. OBJECTIVE YIELD SURVEYS

III. NONRESPONSE TO SURVEYS

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 beer 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. Cne 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:

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

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 $\delta_{hc}=0$, 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:

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

$$= \ddot{R}X_c$$

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

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	**Multi-**	**__Overpass Date__**	
Region	**temporal**	**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

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

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

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

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

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

$\hat{R}_{\text{ç}\alpha h}{}^2$ 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

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

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

Where $y_{\text{ç}\alpha hj}$ 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

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

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

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 $R_{ç\alpha h}^2$ 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.

COMPUTER ASSISTED PERSONAL INTERVIEWING (CAPI) COSTS VERSUS PAPER AND PENCIL COSTS

Bruce Eklund
National Agricultural Statistics Service (NASS)
3251 Old Lee Highway, 3rd Floor, Fairfax, Virginia, 22030

KEY WORDS: Blaise, Computer Assisted Personal Interviewing (CAPI), Interactive Editing (IE)

Computer assisted personal interviewing (CAPI) promises improved data quality and timeliness, data quality through edits invoked *during* data collection and timeliness by telecommunication.

The purpose of this paper is to help an organization ask the right questions. First, is CAPI feasible? The organization must write easy-to-use instruments for its surveys and integrate CAPI into its survey process.

A more comprehensive question is "*should* an organization use CAPI?". Issues include feasibility, whether an organization can realize CAPI's potential benefits, and the topic of this paper: cost. This paper briefly reviews experience that shows NASS can realize CAPI benefits, describes initial cost comparisons, and documents a parallel CAPI and paper comparison.

CAPI Cost Analyses are Specific to the Organization

Britain's Office of Population Censuses and Surveys thought that computer assisted interviewing, including CAPI, would not only save money, but it was imperative to meet new budget constraints (Manners, 1991). How CAPI affects costs, however, depends greatly on an organization's survey program. Cost issues include the degree of centralization, number and frequency of surveys, consistency of surveys over time, length of data collection periods, and the amount and complexity of editing.

NASS's decentralized structure will challenge CAPI management. NASS trains enumerators from 44 state offices. In each state they gather each year for each major survey. Training costs already consume a considerable portion of survey costs. Costs include state office staff and enumerator salaries, lodging, and per diem. If CAPI training increases total training time, then CAPI increases NASS survey costs appreciably.

NASS collects survey data within short periods. Also, most surveys are quarterly or annual, not weekly or monthly. Thus, although NASS needs to write several survey instruments and train for several surveys, there are short, concentrated periods to realize CAPI benefits.

Conversely, during the concentrated periods, NASS spends considerable staff time on office editing after the interviews. Clerks verify arithmetic calculations; data entry clerks key and verify; and professional, subject matter specialists (agricultural statisticians) hand edit questionnaires. Afterward, NASS rents mainframe computer time to batch edit the data. Statisticians usually wait overnight for the output, make corrections, and then run another batch edit overnight. If CAPI reduces office editing time, it will greatly affect survey cost.

NASS Experience

NASS developed CAPI for a cross section of its surveys. The more of the survey program NASS can use CAPI, the more economically viable CAPI is. NASS would show that for a cross section of its surveys:

1) NASS can develop easy to use survey instruments

2) enumerators can learn and use CAPI well

3) telecommunication would speed data retrieval

4) CAPI would help clean data.

NASS first used CAPI operationally for a simple survey, the Livestock Prices Received. CAPI data were compared to paper collected data with a post-collection batch edit. CAPI flagged suspicious data (outside specified ranges) during the CAPI interviews. The enumerator could fix an error or verify valid data. CAPI collected data had 57% fewer suspicious "error" flags in the post-collection batch edit (Eklund, 1991).

More importantly, CAPI reduced "critical" errors by a factor of 16.5. Critical errors are non-sensible entries.

For this survey, batch edit critical errors often resulted from missing items that made data records unusable. CAPI required enumerators to answer these items before proceeding.

Post-Collection Batch Edit "Error" Rates[1]

Livestock Prices Received Survey

For Paper & Pencil

non-critical, suspicious "error"%= 3.74%

critical error%= 0.33%

n ~ 11000

For CAPI

non-critical, suspicious "error"%= 2.15%

critical error% = 0.02%

n ~ 12000

1) Pct. of records (data from livestock sales) with "errors".

In 1989, for the September Agricultural Survey, NASS found what other organizations have found: enumerators can use CAPI effectively and respondents accept CAPI. Two survey software products, CASES and Blaise were tested for CAPI.

NASS also used laptops and electronic calipers to size almonds to forecast production during the growing season. Enumerators entered data both from the key board and the caliper. The caliper, connected to a serial port, put data directly into the survey instrument. Enumerators then sent the interactively edited data to the State office by telecommunication, showing the ability to reduce the time between field assessment and crop forecasting. Furthermore, the technology can eliminate the need for a two-person team. With CAPI, one person can simultaneously measure and record.

Next NASS tested CAPI with a challenging survey, the Farm Cost and Returns Survey (FCRS). The paper questionnaire is over thirty pages, has complex branching, requires plenty of arithmetic, asks detailed financial questions, and averages 90 minutes per interview. It is difficult to convey the specific intent for some questions. Often farmers cannot respond precisely. NASS proved that Blaise CAPI software could handle this complex questionnaire. Enumerators showed that they could learn and use CAPI for one of NASS's most difficult surveys.

Relationship between CAPI & Interactive Editing (IE)

An important part of cost comparison is the relationship between CAPI, Computer Assisted Telephone Interviewing (CATI), and after data collection, Interactive Editing (IE). The Netherlands' Central Bureau of Statistics developed integrated survey software called Blaise. The user can write code and then compile in Pascal into either a CAPI, CATI, or an IE survey instrument. After CAPI or CATI, a subject matter specialist can use an IE instrument to review the data interactively, questionnaire by questionnaire, with automated edit checks. Programmers have only to write one set of code for CAPI, CATI, and IE. Cost estimates hereinafter assume NASS effectively integrates CAPI and IE software functions.

What Does CAPI Cost?

Understandably, NASS did not design its survey cost accounting with categories conducive to contrast paper and CAPI costs. Thus ten state offices estimated costs for several sub-categories of survey preparation, enumerator training, data editing, data entry, batch editing, and mailing costs.

A framework, starting with the FCRS project, was built for estimating and comparing CAPI cost. Spreadsheets can ease updates of cost estimates as costs change and people learn to use CAPI more effectively.

Corresponding costs were estimated or projected for CAPI based on previous CAPI and IE applications. Although survey specific cost estimates were developed for the FCRS, hardware costs were amortized both across surveys within a year and across years for hardware life expectancies. Thus, one can extrapolate the cost estimates for the gamut of CAPI targeted surveys. The chart depicts major (large X) and minor (small x) cost increases. An assumption is that people use CAPI/IE effectively, which excludes some initial costs.

CAPI COSTS
vs.
PAPER COSTS

SURVEY COSTS	*PAPER*	*CAPI*
Enumerator Training		X
Hardware		X
Survey Preparation	x	
Mail Costs	x	
Telephone Costs		x
Data Entry	x	
Data Editing	X	

Not only hardware but also training cost is likely to increase significantly with CAPI. Effective *and* cost efficient CAPI training is challenging. The organization must balance or integrate *effective* methods such as low student/teacher ratios and field observation with cost *efficient* methods such as home self training and practice. Also, survey trainers should build CAPI practice into the survey schools.

Most savings will come from office editing, which is examined in following sections. CAPI can save some survey preparation cost. Automated distribution of samples is cheaper than the present method of writing a program to print labels in each state office and then manually distributing properly labeled questionnaires to each enumerator. CAPI also saves mail and data entry costs but increases telephone costs somewhat with telecommunication.

Total CAPI cost was estimated close to or slightly less than paper collection in 1992 (Eklund, 1993). The margin of error of some estimates was such that the paper method may or may not have still been slightly less expensive.

The important points are to have reasonable estimates and to understand cost trends. Paper intensive costs such as mail and office labor are increasing. CAPI costs, such as hardware and telecommunication are decreasing (Clayton and Harrell, 1989). An organization should develop, test, and gain CAPI experience on a small scale to prepare for large scale use when cost, data quality, and timeliness, point to CAPI use.

CAPI and the June Area Frame Survey

Although this survey is considerably shorter than the FCRS, enumerators are much more likely to interview farmers outside. They often stand to interview as farmers work. Enumerators also must carry a 2' by 2' aerial photograph of the farm. Smaller computers enabled enumerators to use CAPI successfully in June, 1993.

The Blaise CAPI instrument was also coded for statisticians to use as an IE instrument after data collection. The IE instrument invoked more rigorous edits then CAPI. For example, the Blaise software requires one to "fix" edits deemed critical, but one can override non-critical errors. Programmers coded some critical IE edit checks as non-critical CAPI edit checks. Like the FCRS, enumerators handled this difficult survey well.

CAPI/IE Costs and the June Area Frame Survey

Once enumerators sent data to the State office, the editing flow was divided into four processes.

Process 1: Traditional method

1) paper interview => clerical edit => 1st statistician edit => 2nd statistician edit => key entry => key entry verification => batch edit.

Process 2: IE with one statistician hand edit

2) paper interview => clerical edit => 1st statistician edit => key entry => key entry verification => IE => batch edit.

Process 3: IE with no statistician hand edits

3) paper interview => clerical edit => key entry => key entry verification =>IE => batch edit.

Process 4: CAPI

4) CAPI => IE => batch edit.

The office staff divided into three teams that rotated to edit under each of the first three processes. The office processing time per interview is in the following table. The CAPI/IE combination shows the potential to reduce the office process by a factor on more than three.

Survey Process Times in State Office

(Minutes per "Interview")

Process	1	2	3	4
Clerical Edit	2.5	2.4	3.8	-
1st Statistician Edit	5.2	3.9	-	-
2nd Statistician Edit	3.3	-	-	-
Key and Verify	5.1	5.0	5.0	-
Interactive Edit	-	3.4	4.7	≤5
Total Time	**16.1**	**14.7**	**13.5**	**<5**
n - Interviews	757	341	391	86

Statisticians edited using the IE instrument more slowly than they will in the future because of: 1) a learning curve and 2) imperfections in the operationally untested instrument.

The IE part of process 4 was not in a production mode as processes 2 and 3 were. CAPI data were compared item by item to the paper backup to scrutinize enumerators' CAPI. Thus, the IE part of processes 2 and 3 were conservatively extrapolated to IE for process 4.

The project did not measure some survey management costs such as post-collection file handling. Conversely, CAPI/IE should reduce waiting periods for batch edits, reduce time editing those batch edits, and reduce out of pocket costs from the batch mainframe edits. Also, more savings should come as programers, statisticians, enumerators, managers, the entire organization; gains CAPI experience.

Interview Time

Differences in interview times are a negligible part of survey costs. Interviewing takes a small portion of the enumerators' time. Enumerators spend much more time traveling and finding sampled farmers. Interview length is a respondent burden issue, not a cost issue.

The instrument measured CAPI interview lengths automatically for the entire interview and by sections within the questionnaire. Enumerators do not record paper interview times so no comparison was made.

From observation, CAPI slowed the interview within a section that comprises a large table. For other parts of the questionnaire, particularly parts with automated skips, CAPI was faster. CAPI also saved time because it automated calculations.

Enhanced software or faster computers will soon further speed CAPI. Enumerators also suggested specific features that programmers could use to speed CAPI. Relative CAPI and paper interview times depend upon the survey, software, hardware, the intensity of interactive edits, and the imagination of survey instrument designers.

Conclusions

Although CAPI has high initial costs for hardware and training, it shows potential in reducing office work during a peak work period. CAPI can greatly reduce office keying and hand editing. CAPI/IE shows substantial gains over both the paper method and over IE without CAPI.

Results from the 1993 CAPI/IE June Area Frame supported key parts of previous projections of CAPI reductions in office work load. The unpolished instrument was facing its first operational test. Future CAPI/IE development should yield even more savings.

This promising work is fledgling cost analysis, not the final word. Cost estimates should be revaluated on an ongoing basis as the organization refines and expands the CAPI/IE process. The organization should use these cost analyses as a tool to help plan CAPI use.

This work focused on a particular aspect of the CAPI decision: cost. It did not quantitatively include timeliness and data quality. The organization must decide how to weigh all considerations before choosing whether to use CAPI.

Acknowledgments

Mark Pierzchala and Ann Romeo were co-conspirators with the author in coding the June Area Frame survey instrument.

References

Clayton, R. and Harrell L., Bureau of Labor Statistics, "Developing a Cost Model for Alterative Data Collection Methods: Mail, CATI, and TDE". Military and Government Speech Technology Conference. November 1989, Arlington, Virginia.

Eklund, B. Mobile Computerized Data Entry and Edit for the Livestock Prices Received Survey. page 5, SRB Research Report No. SRB-91-10, Washington D.C., USDA, National Agricultural Statistics Service, November, 1991.

Eklund, B. Computer Assisted Personal Interviews: Experience and Strategy. page 11, STB Research Report No. STB-93-01, Washington D.C., USDA, National Agricultural Statistics Service, February, 1993.

Manners, T. "The Development and Implementation of Computer Assisted Interviewing for the British Labour Force Survey". Computer Assisted Survey Information Collection: International Progress, Special Contributed Papers for the American Statistical Association 1991 Joint Statistical Meetings. August 1991, Atlanta, Georgia.

History and Procedures of Objective Yield Surveys in the United States

Eric Waldhaus, Eddie Oaks, Mike Steiner, National Agricultural Statistics Service
Eric Waldhaus, NASS, USDA, Room 4162, South Building, Washington, D.C. 20250

KEY WORDS: Crop cutting, Enumerators, Objective Yield, Yield Surveys

This paper reviews the efforts made by the National Agricultural Statistics Service (NASS) in the United States Department of Agricultural (USDA) to measure crop production by direct measurement of plant characteristics. NASS implies the current agency and all of its predecessors. These efforts are collectively known as Objective Yield Surveys (OYS). This program is similar to Crop Cutting surveys done in many parts of the world. The major difference is that OYS includes non-destructive field counts prior to harvest to facilitate yield forecasts during the growing season. Yield is defined as the weight of targeted crop, at standard moisture, produced per unit of harvested area. Sampling for NASS Objective Yield Surveys have been on a probability basis from the inception. This is, however, not the implication of the word 'objective' in the title. Objective refers to the direct collection of plant characteristic measurements, instead of subjective estimates of yield reported by an observer.

NASS is a recognized world leader in the use of objective yield technology. Objective yield surveys produce the primary indications for yield forecasts and estimates for the major feed and food grains in the United States. Additionally, NASS has made long term commitments to make this technology available internationally. Through cooperative arrangements NASS has demonstrated or helped implement objective yield programs in many countries of Asia, Africa, and Central and South America.

Three aspects of NASS' objective yield program for major field corps are considered: The history and evolution of the program, current sampling procedures, and general concepts of objective yield survey field procedures. Specifically not included is a discussion of the use of survey data in preparing yield forecasts. This major topic is covered in another paper presented at this conference.

HISTORY

Yield and production of major field crops in the United States have been forecasted and estimated by USDA since President Abraham Lincoln's administration in the 1860's. Crop condition surveys were prepared monthly by the Statistics Division, USDA as early as 1863, the year following the creation of the Department. Until 1884 pre-harvest reports were in terms of condition as compared to an 'average' crop. In 1884 the reporting concept changed. Condition began being asked as a percent of a 'normal' crop, given no adverse effects of weather, disease or pests.

Although crop area changes from year to year, some of the largest variations in crop production are caused by fluctuations in production per unit area or yield. For more than a century, yield forecasts were based solely on voluntary producer appraisals of expected yield. It was recognized early that actual changes in yield were not fully reflected in subjective grower appraisals. By 1898 traveling agents supplemented farmer-crop reporters' information with on site observations of crop conditions. By 1903 more than 100,000 agriculture related business operators, including cotton ginners, millers, elevator operators, and transportation agents were paneled to gain insight into the agricultural situation.

In 1910 a shift began in the practice of reporting crop condition to forecasting actual production during the growing season. By 1915 cotton production forecasts became available during the growing season. The transition from condition to yield forecasts required regression modeling. This was almost entirely done by visual interpretation of charts prior to the use of computers in the late 1960's.

Objective measurements for forecasting yield started with cotton in 1928. These early efforts involved statisticians driving along the perimeter of cotton fields, making boll counts at predetermined locations in fields. There appears to have been no effort made to relate the field counts to yield. Thus, it may be more appropriate to think of this early effort as 'Objective Condition' surveys. Later corn and wheat were added to this program, but this early effort in objective methods was discontinued at the start of the World War II. Research into objective measurements of wheat, corn, and cotton resumed in 1954.

The 'birth' of probability sampling for agricultural statistics and objective yield methods came in 1957 when the United States Congress funded an initiative titled "A Program for the Development of the

Agricultural Estimating Service". The project provided for an annual enumeration of a large area frame probability sample for crop area estimates. This area frame survey evolved into the current June Agricultural Survey (JAS). Target crop fields identified during the JAS provide the sampling universe for the OYS (except winter wheat).

Cotton and corn objective yield programs became operational in 1961. Wheat came on line a year later. Soybeans joined the national program in 1967 and potatoes in the early 1970's. Grain sorghum, sunflowers and rice were added in the 1980's, but due to budget constraints grain sorghum and sunflowers were dropped in 1988. The rice program was reduced then and finally discontinued in 1993.

OBJECTIVE YIELD SURVEYS OVERVIEW

NASS is organized into 45 State Statistical Offices (SSO). There is one in each state except in New England where six states are combined. There is a centralized Headquarters in Washington, D.C. Sample design and selection, planning and coordination between states, centralized data processing, and quality assurance are the major roles of Headquarters in the OYS. Headquarters prepares and distributes two major OYS manuals. The Supervising and Editing manual is focused on the tasks completed in the SSO. The Interviewers' Manual is a training and reference manual for enumerators in the field. Each SSO coordinates field work and other data collection activities independently within established guidelines.

Qualified, adequately trained field personnel, including SSO staff and field enumerators, are essential for a quality job. States send a survey statistician, designated the State Survey Statistician, to a National training workshop to learn and reinforce correct procedures. State Survey Statisticians return to their states to train field supervisors and enumerators.

Objective Yield Surveys begins with intensive training for field enumerators. Training is more intensive for OYS than many other NASS data collection operations. The need for rigorous training stems from the fact that data collection usually is accomplished in remote locations in the field where supervision is minimal and there is not the opportunity to clarify procedures. It is also recognized that data collection is often a very time sensitive process so it may be impossible to reconstruct an 'interview' when errors are discovered after the field work is complete. The cost of training is very high, but the need is critical. NASS has consistently recognized this need and continues to make a substantial resource commitment to training.

NASS field enumerators are part time employees of the National Association of State Departments of Agriculture (NASDA). NASDA contracts their services to NASS. There are approximately 600 NASDA enumerators who work on OYS. OYS enumerators are almost exclusively rural people, and most are from farm families, typically retired or part-time farmers or farm spouses. Understanding agricultural practices is a prerequisite for a successful OYS enumerator. Enumerators also have to demonstrate literacy and computational skills about equivalent to a high school graduate.

In addition to training, the State Survey Statistician is charged with monitoring survey progress, and is the resource person for enumerators. Responsibly also extends to oversight of all SSO processing of survey data, and supervising laboratory processes. The State Survey Statistician and assistants review all edit and summary output. In most states the final yield recommendations (proposed estimates) submitted to Headquarters are not prepared by the Survey Statistician, but a Commodity Specialist.

Field Quality Control is conducted by supervisory enumerators and statisticians from the State office. A random sample of each enumerator's field work is selected for personal inspection. The sample selected for quality control is unknown to the enumerator and the supervisor in advance to insure an accurate assessment of quality. The sample pattern is such that at least one quality check for each enumerator is insured, and multiple checks throughout the survey cycle are possible. Supervisors may inspect additional work of the enumerators in their charge on an 'as needed' basis.

Occasionally, deficiencies in field procedures are discovered by the quality control process. When this occurs remedial action is taken, both to correct errors in a particular sample and to re-train the errant personnel. Discovery of deliberately falsified survey results is another potential benefit of the Quality program. The authors, with about twenty years of objective yield experience each, have no personal knowledge of this ever occurring.

Objective yield surveys are timed for making crop production estimates which are released to the public in the monthly *Crop Production* report. *Crop Production* is published during the second week of the month, between the 8th and the 12th. To complete field work, process all data, and remain timely, the OYS adheres to a very rigid schedule. Data collection starts on the 22nd of the month prior to the survey reference date, and must be completed by the first of the reference month. Laboratory work, data processing, and summary review are completed, and recommendation submitted to the NASS Agricultural Statistics Board in Headquarters by the second day before the *Crop Production* release.

Concepts and methodology used in the OYS for forecasting and estimating yields are similar for all field crops. Two components of yield -- weight of the fruit and number of fruit -- are used to forecast a yield. Various plant characteristics are used to predict these components during the growing season. Harvest losses, estimated by gleaning small plots in the sample fields after harvest, are deducted to obtain a net yield.

During the early growing season, crop maturity varies considerably by region. Plant characteristics and measurements made to forecast yield change as the season and plant maturity progresses. The enumerator determines the maturity stage of the crop in the sample field during each visit and makes the appropriate counts and measurements for the growth stage.

Observations for each sample are made on two randomly selected plots (units) in each of the selected fields. Each plot consists of a specified number of parallel rows of predetermined length, or a rectangular unit drawn to specification if crop rows are indistinguishable.

SAMPLING

OYS samples are selected from acreage reported of the target crop in the March Agricultural Survey (MAS) or the June Agricultural Surveys (JAS). Spring and durum wheat, corn, cotton, potatoes, and soybean samples are selected from the JAS. The winter wheat sample comes from the MAS.

Winter wheat samples are unique as they are selected from the March Agricultural Survey using a multiple frame (combined list and area survey) design. Also, winter wheat varies in that samples are drawn from 'fields to be harvested for grain', while other crops are sampled from fields 'planted and to be planted' on the parent survey.

The objective yield sample for each crop is allocated to the most important production states such that 80 percent or more of the nations crop is included. Allocations are made to minimize production estimate coefficient of variation (CV). Until about 1990 allocations were made to maintain minimum harvest level CV's. As estimation models have improved, an effort has been made to allocate samples to maintain a minimum CV across the growing season.

The JAS, which is the parent survey for OYS, is the major, once a year, multiple frame survey conducted by NASS. Nationally, the area frame component includes approximately 15,500 segments, each about 1 mile square, representing about 52,500 farms which are enumerated in early June to identify land use. The area of target crop planted is expanded by the associated expansion factor for the area frame sample. OYS samples are then selected proportional to the expanded acreage. Proportional sampling insures that the distribution of the OYS sample will approximate the distribution of the crop as discovered in the JAS. Sampling procedures are similar for winter wheat except MAS is the base survey.

Survey States, sample size, and sample distribution are reviewed annual, but NASS has attempted to maintain consistent State involvement and sample sizes to maintain year to year comparability. In 1993 1,670 winter wheat samples were selected in 13 States. Spring wheat samples totaled 380 in four States, and 150 durum samples in one State were selected. Corn samples equaled 2,010 spread over 10 States, while 1,360 Cotton samples were drawn in six States. Soybeans samples totaled 1,330 in eight States, and 2,080 Potatoes samples were distributed over 11 States.

FIELD PROCEDURES

Enumerators are provided aerial photograph with the area frame segment containing the selected sample field outlined in red. Operators of land in these segments were interviewed during the JAS. Within the segment there may be more than one tract (farm). The enumerator locates and interviews the operator of the tract which contains the selected target crop field for OYS.

Six reporting forms are used through the growing season to collect information from the farm operator or to record counts and measurements. The reporting forms are identified by letter initials, which reflect the chronological order of use of the forms during the growing season. The data collected on each form are similar for all crops in the OYS program.

A convenient way to describe the field procedure for implementing the OYS is to describe each reporting form, and explain its use.

Form A - is an interview form, used to update the crop acreage intended for harvest and to identify the sample field. It shows which field (area frame) or how to select a field (list frame) that will be used for making actual field counts and measurements. The Form A is completed on the first visit to the selected farm. It is also used to gain permission from the farmer to enter the field to set out OYS sample units, and to query the farmer about pesticide usage so the enumerator can take appropriate personal safety precautions.

Pesticide usage has expanded over the years both in the crops treated and the variety of chemicals available. Consequently pesticide safety training and enumerator exposure monitoring has become an integral part of the OYS program. This is especially true for Cotton OY, where the use of organophosphorus pesticides is nearly universal.

Form H - also an interview form, is used to collect data on seed, fertilizer, and pesticide application rates and tillage practices. These data are used for further economic analysis, and are not part of the yield estimation program directly. It is completed at the same time as the Form A.

Form B - is a field observation recording form. It is used to record counts and measurements of the plants and fruits. This form also reiterates instructions for locating, constructing, and processing the sample units.

The following two sections: Locating the Sample, and Counts and Measurements are presented here because these activities are associated with completion of Form B. A separate Form B is completed each survey month until harvest time, when a final Form B is completed.

LOCATING THE UNIT:

After completing Forms A and H, the units are constructed in the sample field by the enumerator. Two units are laid out for each sample. Unit 1 and Unit 2 are located independently of each other (except in wheat where unit locations are dependent). The random number of rows and paces for locating Units 1 and 2 are computer generated and preprinted on a label on the Form B.

The point of entry into the field, or starting corner, is the first corner reached when approaching the field that allows the units to have a chance of falling anywhere within the field boundaries. The shape of the field must be considered to insure that the entire field has a chance of selection. Research has indicated that there is no statistical differences related to starting corners. Therefore, any field corner which does not exclude some part of the field is acceptable.

The following steps are followed when locating and laying out units:

Step 1: The enumerator marks the staring corner with a piece of plastic flagging ribbon so it will be clearly visible on later visits.

Step 2: The enumerator then walks along the end of the crop rows the number of rows (or paces for wheat and broadcast seeded fields) indicated for Unit 1. A piece of flagging ribbon is tied onto the first plant in Row 1. This helps locate the same row on later visits. The next row in the direction of travel will be Row 2 of Unit 1.

NOTE: The enumerator walks his or her normal paces when locating the units within the field. It is not necessary to measure the distance traveled as it is not necessary to locate a precise point in the field, only one determined by a random process.

Step 3: The enumerator then walks the required number of paces into the field between Row 1 and Row 2, starting the first pace 1.5 feet outside the plowed end of Row 1. This makes it possible for a unit to fall anywhere in the field including the very edge.

Step 4: After the last of the required paces is taken, a dowel stick is laid down so that it touches the end of the enumerator's shoe. The dowel is placed across Row 1 and Row 2, at a right angle to the rows. The unit is laid out in the direction of travel of the last pace.

Step 5: The zero end of a 50 ft. tape is anchored at the dowel stick directly beside the plants in Row 1. The sample number is written on a florist stake and inserted at the anchor point.

Florist stakes are colored lath about 6 to 8 inches long. They are highly visible markers commonly used in nursery and greenhouse operations to mark seed beds. Florist stakes deteriorate quickly so no hazard will be created if lost or abandoned in the field after the survey.

Step 6: In row 1 a starting florist stake is placed exactly 5 feet from the anchor point. It is marked "U1-R1". This measured 'buffer zone', helps insure that the unit location is not subjectively biased in its location by the enumerator. The florist stake should be placed beside the row about 2 inches from the base of the plants. The marker is placed outside the plant row to avoid any damage to the developing crop.

Step 7: Working outside the unit, the enumerator carefully measures the unit length and places a florist stake at the designated point. Corn, cotton and potatoes have larger unit lengths which are measured with a tape. For example, the corn count area is 15 feet long. A rigid metal frame is used for marking wheat and soybeans where the unit size is smaller. The wheat unit is 21.6 inches.

Not all fields are square or rectangle and other special situations may arise when locating and laying out a unit. The Interviewers' Manual gives details on how to handle most of these situations. Some of the problems that more commonly occur include: blank areas in the field that were known or unknown during the mid-year survey; the field is not large enough to accommodate the number of rows or paces specified; row direction changes; odd shaped fields are encountered as circular fields under pivot irrigation; fields planted on contours; or crop rows that are not distinguishable due to sowing practices. These situations are covered with precise instructions.

The Form B is the recording form for counts and measurements that are made at the units. Visits to these sample units will take place monthly during the growing season except for potatoes, when only one visit is made within 3 days of harvest or when vines are dead.

Because the same sample unit must be revisited monthly it is important the enumerator precisely mark the location of the unit. Plastic flagging ribbon is used. This is highly visible, but like the florist stakes, quickly disintegrates so it may be abandoned after the survey.

COUNTS AND MEASUREMENTS

Step 1: Measure 1-row space and then 4-row spaces. Measurements are made from the plants in row 1 to row 2 and then from row 1 to row 5. These measurements are used to calculate area of the unit.

Step 2: Count the number of plants in each row in the designated unit.

Step 3: Classify the unit by maturity category. Descriptive four page handouts with color picture examples are helpful in determining maturity.

Step 4: Make the specific counts and measurements of plant characteristics required. Different counts are made depending on the maturity level category. The crop and type of counts are as follows:

Soybeans: 1) plants; 2) nodes; 3) lateral branches with blooms, dried flowers, or pods; 4) blooms, dried flowers and pods; and 5) pods with beans.

Corn: 1) plants; 2) average length of kernel rows; 3) diameter of ear; 4) stalks with ears or silked ear shoots; 5) number of ears; 6) ears with kernel formation; and 7) cob length; and 8) field weight of corn.

Cotton: 1) plants; 2) burrs, open and partially opened bolls; 3) large unopened bolls; 4) small bolls and blooms; and 5) squares.

Wheat: 1) stalks; 2) heads in late boot; 3) emerged heads on all stalks; and 4) detached heads.

Potatoes: 1) hills; 2) tubers; and 3) field weight of tubers in the unit.

After completing Unit 1 counts and measurements go back to the beginning of the Row 1, and walk to the designated row, or number of paces, for Unit 2. Continue in the original direction of travel as when locating Unit 1 if Unit 2 count exceeds the Unit 1 count. After locating the Row 1 of Unit 2, walk the required paces into the field to set up Unit 2, and make the counts and measurements required.

A Form B is done for each month until very near harvest. Close contact is made with the operator so a sample field will not be harvested before a final Form B (just before harvest) can be completed. During this last visit before the farmer harvests, a sample of mature crop is sent to the laboratory. This sample is the basis for at harvest yield estimates.

FORM C-1 and C-2 - These forms record laboratory observations, and are not seen by the field enumerator. Form C-1 records data from pre-harvest field visits, while the C-2 is generated from the last field visit made at, or just before the farmer harvest.

FORM D - is used to record the actual number of acres harvested at the end of the year and the operator estimated yield of the field.

FORM E - is a field observation form used to collect data for determining field harvest loss so a net yield estimate can be made. The field visit to collect data must be within 3 days after harvest to determine harvest loss accurately as loose grain deteriorates quickly or is lost when left in the open. Harvest losses are subtracted from gross yield to arrive at a net yield. Finding the location of this post-harvest unit is similar to the original unit location. A measured rectangle is staked out and fruit from the crop is collected, and sent to the lab. There it is counted, weighed, and moisture tested to determine the field loss.

NON SAMPLING ERROR

Controlling non-sampling error is a major concern of the OYS program as in any large scale sampling survey project. Cause for OYS non-sampling error can be divided into two major categories: faulty procedures, and faulty procedure implementation. Additionally, as OYS use sub-samples from other surveys, non-sampling error present in the parent survey is passed on or magnified. This is out of the control of the OYS personnel except to monitor the larger survey for consistency. This source of error will not be considered further herein.

Non-sampling error which are the result of faulty procedures can be dealt with in a straight forward manner. The NASS research unit continuously reviews various aspects of the OYS program to insure survey validity. Validation surveys are conducted for each crop on a rotational basis. These surveys explore many aspects of OYS, such as the independence of the starting corner as noted earlier.

The survey quality program is also useful in discovering faulty procedures. Most often procedural difficulties that are discovered in the quality control program relate to some 'special case' which was not adequately considered when preparing manuals. Instruction changes that clarified selecting starting corners that do not exclude some part of the field developed largely through this route.

Insuring that procedures are consistently and accurately followed across the country is the greatest challenge in controlling non-sampling error. The most important control for non-sampling error is training. OYS training is continuous. The training cycle starts with training for State Survey Statisticians at National workshops. Usually there are three held in a year, one for Wheat, another for Corn, Cotton and Soybeans, and the third for Potatoes. Corn, Cotton, and Soybeans are combined for training because procedures, growing seasons, and States involved largely overlap.

Training continues with workshops for field enumerators, conducted by the State Survey Statistician. Assistance from the Headquarters OYS unit is available to the SSO's in conducting local training. This can be an important resource for a new State Survey Statistician, and gives Headquarter personnel the opportunity to observe local operations.

The formal quality control program in which individual enumerators have work inspected at random is an important part of the NASS non sampling error control program. While the potential is in place to discover an enumerator who is intentionally falsifying reports or 'table topping', this is not a major concern. The real value of the quality control program is to assess the level and effectiveness of training. Another important benefit of the program is its moral boosting effect on enumerators. The normal out come of quality control is that the enumerator is 'caught doing it right'. When fed back to the enumerator in a positive way this can be excellent reinforcement for continued quality field work.

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)

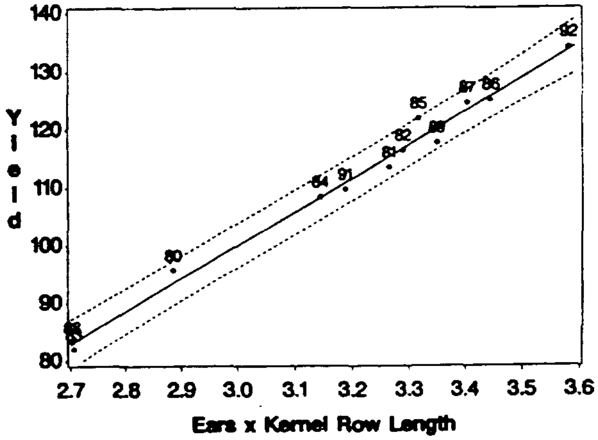
CORN VARIABLES BY MONTH	
August	(stalks with ears + ears with kernels per acre) X (average kernel row length per ear)
September	(Ears with kernels per acre) X (average kernel row length per ear)
October-December	(Ears with kernels per acre) X (average grain weight per ear)

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).

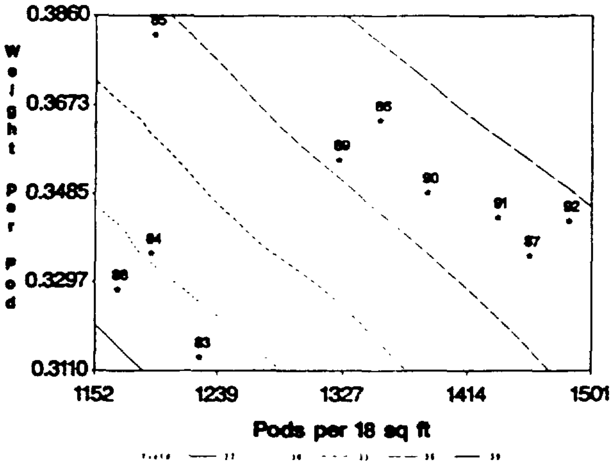
September 1 Corn Model



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

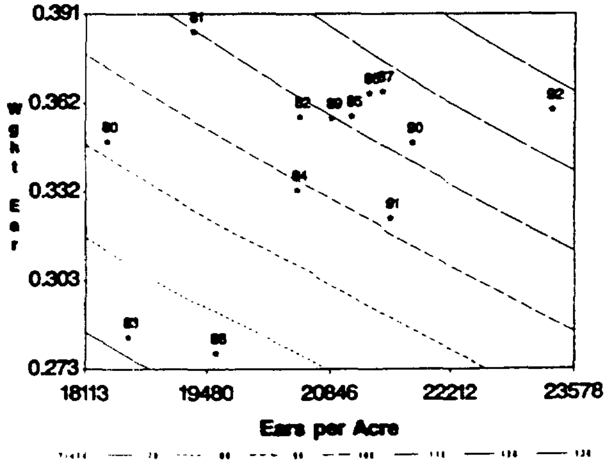


For soybeans, the weight per pod is in grams, and the yield contours projected from the fitted model surface onto the plane are 27, 30, 33, 36 and 39 bushels per acre. On October 1, usually about half the crop has been harvested, and the weight is for just those harvested samples. The pods per 18 square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

= 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,

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

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.

AN EMPIRICAL INVESTIGATION OF AN ALTERNATIVE NONRESPONSE MODEL FOR THE ESTIMATION OF HOG TOTALS

Matt Fetter, National Agricultural Statistics Service
USDA/NASS/Research Division/3251 Old Lee Hwy., Room 305, Fairfax, VA. 22030

KEY WORDS: Nonresponse model, reweighting, bias

1. THE SURVEY

The National Agriculture Statistics Service (NASS) conducts the Quarterly Agriculture Survey (QAS) to collect data on cropland acreage, grain storage, and various livestock items including hogs. The QAS is a "multiple frame" survey. Two independent frames are sampled, the "list" frame and the "area" frame. The "list" frame is a list of farm operations across the U.S. that NASS maintains. The "area" frame is composed of all land in the contiguous U.S. The QAS estimate for an item of interest is constructed by adding the estimate obtained from the list frame sample with the estimate obtained from the farm operations in the area frame sample that are not on the list frame. All estimators of interest in this paper are list frame estimators, thus the area frame portion of the multiple frame estimate will not be discussed further.

The QAS list frame is sampled using a stratified simple random sample design. Stratification is based primarily on each unit's control data for hogs, grain storage capacity, and acreage. A priority scheme is used to place each unit into exactly one stratum. The resulting stratification is not optimal for any one particular item of interest. For example, one stratum might be composed of units having similar grain storage capacity, but their hog characteristics might be quite different. Another stratum might be composed of units having similar hog characteristics but very different cropland acreage, and so forth.

Total nonresponse for QAS list frame samples typically range from 10 to 20 percent. Even for total nonrespondents there is often some information known about the sampled unit. For example, the interviewer or enumerator may be able to determine that the sampled unit is in business. Sometimes the presence or absence of hogs can be determined even though the actual number of hogs may be unknown. This partial information can be used to reduce nonresponse bias.

With the exception of certain self-representing strata, NASS currently uses sampling weight adjustment procedures (reweighting) to reduce nonresponse bias in its estimate of list frame hog totals. Two different estimators are used to model nonresponse. The first assumes that the nonrespondents can be reasonably well represented by the respondents. This is a strong assumption and its validity is seriously questioned. The estimator that is based on this assumption is not of significant interest and will not be formally discussed here. The other estimator is based on a model that uses the hog presence/absence information that is available on some nonrespondents. This estimator will be referred to as the Adjusted Estimator.

The Adjusted Estimator, developed by Crank (1979), was designed to take advantage of all partial information that was available on nonrespondents. At that time, the QAS was only designed to capture information regarding the presence or absence of hogs for nonrespondents. Currently, the QAS captures information regarding in/out of business status (ag-status) for nonresponding units. Cox (1993) described an alternative estimator that incorporates this additional information into the nonresponse model. The purpose of this paper is to describe this alternative estimator (referred to henceforth as the Revised Estimator) and to investigate the effect it has on the level of the estimates produced by the Adjusted Estimator.

The Revised Estimator, was applied to historic data so that a direct comparison of the two estimators could be made. Five major hog producing states (Georgia, Illinois, Indiana, Iowa, and North Carolina) were chosen for this purpose. The Revised Estimator was applied to 15 consecutive QAS surveys (June 88 - December 91) for each state. A comparison of the two estimates could then be made for each state in each quarter.

2. WEIGHTING CELL FORMATION

A weighting cell is defined as a group of sampled units within which nonresponse adjustments are computed and applied to the sampling weights. If the propensity to respond is linked to certain hog characteristics of the sampled units, it is desirable that weighting cells be composed of units that are similar in these characteristics. Under these conditions, all units within a weighting cell would be equally likely to respond. Thus the respondents would be representative of the nonrespondents and nonresponse bias would be minimal.

For the Adjusted Estimator, the weighting cells are the design strata. Because the stratification of the list frame is not optimal for hog estimation, design strata are not the most efficient cells for computing and applying nonresponse adjustments. Thus respondents are less likely to be representative of the nonrespondents within these cells. Through the use of poststratification, it is possible that improved weighting cells can be defined.

3. THE NONRESPONSE MODELS

In order to claim that a reweighted estimator is unbiased in the presence of nonresponse, some assumptions must be made about the nonrespondents. If all other factors are considered equal, the estimator based on the most sound set of assumptions would be judged as the estimator of choice for the reduction of nonresponse bias.

When considering the form of these estimators, it will be helpful to think of the estimation procedure as consisting of a sequence of three specific steps. For each sampled unit, three determinations need to be made. These are:

1) the sampled unit's status as an agricultural operation (ag-status). [(Is the unit in business or out of business)? This determination is only applicable in the case of the Revised Estimator.]

2) the sampled unit's status as a hog operation (hog-status). (Does the sampled unit raise hogs or not)?

3) the sampled unit's status as a hog-total respondent (hog-total status). (Is the number of hogs associated with the sampled unit known)?

A complete respondent will be defined as a sampled unit for which the number of hogs associated with that unit is known. A nonrespondent will be defined as any sampled unit for which any one of the above determinations can not be made.

In order to compare the nonresponse models implied by the estimators considered here, the underlying assumptions must be understood. At each modeled level of nonresponse, a valid assumption concerning the nonrespondents is required to claim that the estimator is unbiased in the presence of nonresponse.

The Adjusted Estimator adjusts for nonresponse at two levels, the hog-status level and the hog-total status level. Therefore, one assumption concerning the nonrespondents at each level must be valid. For the hog-status level, the required assumption is:

Assumption 1A. The probability that hog-status will be determined is the same for all sampled units in a particular stratum. This implies that hog-status nonrespondents represent a simple random sample of the stratum population.

For the hog-total status level the required assumption is:

Assumption 2A. Within a stratum, amongst all units which have been determined to be hog operations, the probability that the number of hogs associated with that unit will be obtained is the same for each unit. This implies that within a stratum, hog operations that are complete respondents represent a simple random sample of all sampled units which have been determined to be hog operations.

If N(h) represents the stratum h population size and n(h) represents the stratum h sample size, the Adjusted Estimator can be expressed in the following form at the stratum level:

$$\hat{Y}(h) = W_{samp}(h)\ A_{hog-st}(h) \sum_{e=1}^{2} A_{hog-tot}(he) \sum_{i=1}^{n(he)} y(hei)$$

(1)

where:

$\hat{Y}(h)$ represents the estimated number of hogs in stratum h.

$W_{samp}(h) =$ N(h) / n(h).

$A_{hog\text{-}st}(h) =$ $n(h) / n_{hog\text{-}st\ resp}(h)$, the hog-status nonresponse adjustment for stratum h,

where:

$n_{hog\text{-}st\ resp}(h)$ represents the number of hog-status respondents in stratum h.

$A_{hog\text{-}tot}(he) =$ $n_{hog\text{-}st\ resp}(he) / n_{comp\text{-}resp}(he)$, the hog-total status nonresponse adjustment for weighting class e in stratum h,

where:

$n_{hog\text{-}st\ resp}(he)$ represents the number of hog-status respondents in weighting class e within strataum h and,

$n_{comp\text{-}resp}(he)$ represents the number of complete respondents in weighting class e within stratum h.

y(hei) represents the number of hogs reported by complete respondent i in weighting class e within stratum h.

n(he) represents the number of units in class e in stratum h.

The subscript e denotes two distinct sets (classes) of hog-status respondents in stratum h; hog operations and non-hog operations. Once a sampled unit is identified as a non-hog unit, the number of hogs associated with that unit is immediately known to be zero. Thus all identified non-hog units are complete respondents. Let e= 1 denote this class. For this class, there is no nonresponse at the hog-total status level. Thus:

$A_{hog\text{-}tot}(h1) = 1$ since:
$n_{hog\text{-}st\ resp}(h1) = n_{comp\text{-}resp}(h1)$.

For the hog operation units (e=2), $A_{hog\text{-}tot}(h2)$ must be expressed in the general form stated above.

The Revised Estimator adjusts for nonresponse at three levels, the additional level being the ag-status level. For each of the three levels, one valid assumption is required for the estimator to be unbiased. These assumptions are:

<u>Assumption 1R.</u> The probability that ag-status will be determined is the same for all sampled units in a particular weighting cell. This implies that ag-status nonrespondents can be thought of as a random sample of the cell population.

<u>Assumption 2R.</u> Within a particular weighting cell composed of identified ag-operations, the probability that hog-status will be determined is the same for all units comprising that cell. This implies that hog-status nonrespondents can be thought of as a random sample of the units composing the cell.

<u>Assumption 3R.</u> Within a particular weighting cell composed of identified hog operations, the probability that hog-total status will be determined is the same for each unit in that cell. This implies that the hog-total status nonrespondents can be thought of as a random sample of the units composing the cell.

The assumptions on which these estimators are based are likely to be invalid unless the weighting cells are judiciously defined. In order to increase the likely validity of the underlying assumptions of the Revised Estimator, it was desirable to define the weighting cells in such a way that they would be composed of units having similar hog characteristics. Poststratification based on each unit's hog control data was used to form weighting cells. Thus the weighting cells were defined similarly to the way that design strata would be defined for a hog-specific survey. In order to further increase efficiency, the weighting cells (post-strata) were defined to insure that approximately 20 complete respondents would be contained in each cell. (The Adjusted estimator is not implemented in such a way as to insure reasonably high numbers of complete respondents).

Because the weighting cells cut across design strata, the Revised Estimator will be expressed at the final nonresponse adjustment cell level, e, e= 1,...,E. The general form of the Revised Estimator is:

$$\hat{Y}(e) = A_{hog\text{-}tot}(e) \sum_{i}^{n_e} W_{samp}(ei)\ A_{ps}(ei)\ A_{ag\text{-}st}(ei)\ A_{hog\text{-}st}(ei)\ y(ei) \qquad (2)$$

where:

$\hat{Y}(e)$ represents the estimate of the total for hog-total status weighting cell e, y(ei) represents the number of hogs reported by unit i in weighting cell e.

n_e represents the number of sampled units in weighting cell e,

W_{samp}(ei) represents the sampling weight for the ith unit in weighting cell e,

A_{ps}(ei) represents the poststratification adjustment for the ith unit in weighting cell e,

$A_{ag\text{-}st}$(ei) represents the ag-status nonresponse adjustment for the ith unit in weighting cell e,

$A_{hog\text{-}st}$(ei) represents the hog-status nonresponse adjustment for the ith unit in weighting cell e, and

$A_{hog\text{-}tot}$(e) represents the hog-total status nonresponse adjustment for the ith unit in weighting cell e. (Note all hog-total status respondents have the same hog-total status adjustment within class e).

All of the nonresponse adjustments have the usual form:

$$\frac{\sum_{\textit{all sampled units} \in \textit{cell}} W^{*}_{samp}}{\sum_{\textit{all responding units} \in \textit{cell}} W^{*}_{samp}}$$

where W* represents the sampling weight or an adjusted sampling weight, depending on the level of the adjustment. All nonrespondents have a nonresponse adjustment of zero by definition.

The poststratification adjustment has the following form:

$$A_{ps}(ei) - \frac{N(g)}{\sum_{i \in g} W_{samp}(gi)}$$

where N(g) represents the number of units on the list frame that fall in poststratum g and W_{samp}(gi) is the sampling weight for the ith sampled unit in poststratum g.

4. THE VALIDITY OF THE ASSUMPTIONS

Although the assumptions implied by the Adjusted Estimator are reasonable, they are not beyond justifiable criticism. As stated earlier, assumption 1A asserts that within a stratum, all sampled units are equally likely to be hog-status respondents. However, if the partial information concerning ag-status is considered valid, then the original sample can be divided into three mutually exclusive groups: 1) those units for which ag-status is not determined, 2) those units identified as non-ag units, and, 3) those units identified as ag units. All units in the first group have a zero probability of having hog-status determined because hog-status determination implies ag-status determination. Clearly, hog-status determination is certain for all units in the second group because all non-ag units have zero hogs. Therefore, one could argue that it would be desirable to augment the nonresponse model so that the probability of determining ag-status is the same for all sampled units, while the probability of determining hog-status is the same for all sampled units which are known to be ag-operations. If valid, this argument would imply that the Adjusted Estimator is based on a misspecified model.

If the Adjusted Estimator is based on a misspecified nonresponse model, it is of interest to understand the effect that this misspecification is having on the estimates of hog totals. First, an argument for the nature of the misspecification will be presented. Second, the effect of this misspecification on the level of the estimate will be described.

All ag-status nonrespondents are either: 1) non-ag units (out of business), 2) non-hog ag-operations, or 3) hog operations. Because every unit in the population must be one of these types, it is reasonable to assume that ag-status nonrespondents represent a random sample of the cell (stratum) population. However, the Adjusted Estimator is based on the stronger assumption that the hog-status nonrespondents as a whole represent a random sample of the cell (stratum) population (see figure 1). For a moment, let us assume that this assumption is valid. If we adopt as a premise that a subset of this set-- ag-status nonrespondents, represents a random sample of the cell population, then the compliment of this subset-- identified ag-operations that are hog-status nonrespondents, must also represent a random sample of the cell population. It will now be argued that the Adjusted Estimator's assumption is not reasonable under the adopted premise.

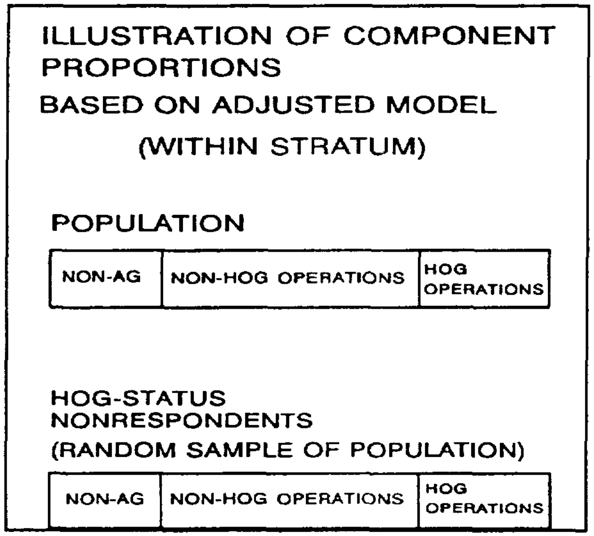


Figure 1

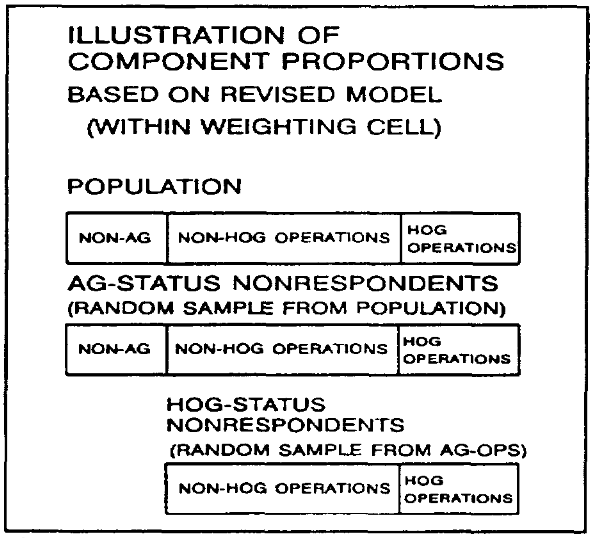


Figure 2

All identified ag-operations that are hog-status nonrespondents must either be non-hog ag-operations or hog operations. Because non-ag operations are missing from this group (all non-ag units are hog-status respondents-- they are non hog units), it is difficult to argue that identified ag-operations that are hog-status nonrespondents can be thought of as a random sample of the cell population (see figure 2). The effect of this misspecification is to bias the estimate downward. This can be explained as follows:

Identified ag-operations that are hog-status nonrespondents have only one source of zeros-- non-hog ag-operations, whereas ag-status nonrespondents have two sources of zeros-- non-ag units and non-hog ag-operations (see figure 2). It therefore seems reasonable to assume that identified ag-operations that are hog-status nonrespondents are more likely to be hog operations than ag-status nonrespondents. It is thus argued that the Adjusted Estimator essentially underestimates the proportion of hog operations in the population. It gives an unbiased estimate of this proportion for the ag-status nonrespondents but gives a downward biased estimate for those identified ag-operations that are hog-status nonrespondents.

The Revised Estimator is based on the augmented model referred to earlier. The difference between the underlying models of the Revised and Adjusted Estimators is that the Adjusted Estimator models all hog-status nonrespondents the same way (Assumption 1A). The Revised Estimator models ag-status nonrespondents as if they are a random sample of the cell population (Assumption 1R), and models identified ag-operations that are hog-status nonrespondents as if they represent a random sample of those units identified to be ag-operations (Assumption 2R).

Note that both estimators model identified hog operations that are hog-total nonrespondents as though they represent a random sample of those records identified to be hog-operations. (Assumption 2A is essentially the same as assumption 3R.)

5. RESULTS AND CONCLUSIONS

The main focus of the research was to observe how estimates obtained from the Revised Estimator would compare to those produced by the Adjusted Estimator using historical QAS data files. The observed effect of applying the Revised Estimator to historical data is an increase in the estimated total number of hogs. This supports the argument that the Adjusted Estimator is biased downwards. The average percentage increase relative to the Adjusted Estimator ranged from a low of 0.64 percent in Iowa to a high of 2.96 percent in Georgia. Across the five states studied, the increase averaged 1.53 percent over all quarters. There were several quarters for which the Revised Estimator produced a lower estimate than the Adjusted Estimator. This was not due to the nonresponse model, but was caused by the poststratification crossing design strata. The Revised Estimator tracked well with the other estimators for all states. Figure 3 shows the relationship between the estimators for Illinois.

The structure of this Revised Estimator is appealing because it provides separate assumptions for each of the three stages of nonresponse. A logical argument has been made that the distribution of the nonrespondent population is different between the ag-status and hog-status stages. The assumptions that nonrespondents are random samples at each stage serve as a reasonable baseline approach, but as yet have not been validated by empirical evidence. Further study is needed to determine the appropriateness of these assumptions.

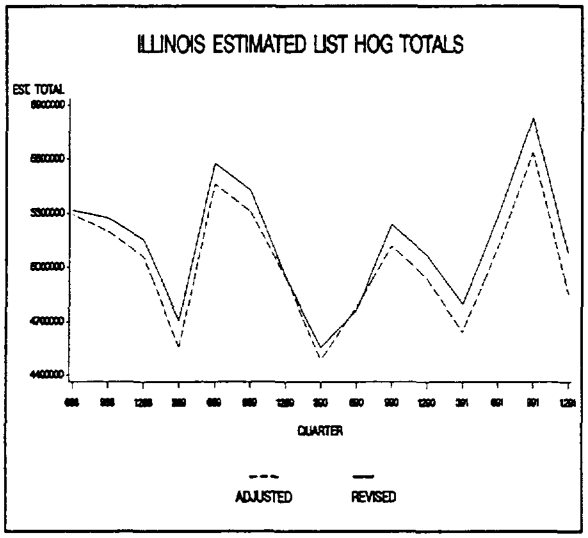


Figure 3

REFERENCES

[1] Agricultural Statistics Board (1991), "Hogs and Pigs" (September 1991 Report), National Agricultural Statistics Service.

[2] Cox, B. G. (1993), "Weighting Class Adjustments for Nonresponse in Integrated Surveys: Framework for Hog Estimation," National Agricultural Statistics Service.

[3] Cox, B. G. (Unpublished), "Weighting Survey Data for Analysis," National Agricultural Statistics Service.

[4] Crank, K. N. (1979), "The Use of Partial Information to Adjust for Nonresponse," Agriculture, Economics, Statistics, and Cooperatives Service.

[5] Fuchs, D. R., and Bass, R. T. (1990), "1989 Agricultural Surveys Survey Administration Analysis," National Agricultural Statistics Service.

[6] Kott, P. S. (1990), "Nonresponse Adjustments in NASS Agricultural Surveys," National Agricultural Statistics Service.

[7] Kott, P. S. and Thorson, J. (1989), "Improving Variance Estimates for Livestock Surveys," National Agricultural Statistics Service.

[8] Garribay, R. and Huffman, J. E. Jr. (1991), "1990 Agricultural Survey Administration Analysis," National Agricultural Statistics Service.

IDENTIFYING AND CLASSIFYING REASONS FOR NONRESPONSE ON THE 1991 FARM COSTS AND RETURNS SURVEY

Terry P. O'Connor, USDA/NASS
Research Division/3251 Old Lee Hwy., Fairfax, VA 22030

ABSTRACT

A research study was conducted during the 1991 Farm Costs and Returns Survey (FCRS) to identify and classify the reasons given to field interviewers by potential respondents for refusing to participate in the survey. The reasons given by field interviewers for coding a sampled unit as inaccessible during the survey were also identified and classified.

The research was conducted in all 48 surveyed states, and included 6 FCRS questionnaire versions. Upon receiving a refusal, interviewers were instructed to record the reason given on the face page of the questionnaire. If no reason was given, or in cases where more than one reason was given, the interviewers were instructed to discuss the concerns of the respondent in regards to completing an interview, and identify the main reason for refusing. When a sampled unit was coded as a inaccessible, interviewers were instructed to explain the reason for the inaccessible.

During the survey statistician's manual edit of the questionnaires, the reasons for refusal or inaccessible were reviewed and compared to a coded list of reasons for nonresponse compiled from previous research into this topic on the FCRS. Statisticians could consider the comments from the interviewers as a match to a pre-coded response, or add additional codes for unique comments.

The nonresponse rate on FCRS averages 30% per year. The reasons behind the nonresponse have been a source of speculation for many years, and previously only anecdotal evidence was available on which to base efforts to maximize response. This research shows the anecdotal evidence to have been on the mark in some cases an off in others.

INTRODUCTION

The Farm Costs and Returns Survey (FCRS) is a face to face interview survey conducted annually during February and March by the National Agricultural Statistics Service (NASS). It is a survey of the agricultural sector, and is conducted in the 48 conterminous states to collect detailed information on farm expenditures and income, costs of production and demographic data. The FCRS has a multiple frame design utilizing a list sample of medium and large ranches and farms, and an area nonoverlap sample of Resident Farm Operators (RFOs) not represented by the list, most of whom operate small farms (Rutz, 1991).

While all 48 FCRS states utilize the same survey procedures, the FCRS includes several questionnaire versions used in different combinations across the country. The versions used in a particular state for a given year depend upon the agriculture in that state and the areas of agricultural specialization being studied. Costs of producing the various agricultural commodities are studied on a year-to-year rotating basis. There are variations in geography, sample sizes, farm or ranch types and sizes, economic conditions and respondent attitudes about the survey across the country; therefore, many factors must be considered when making direct state to state comparisons of the survey results (Rutz, 1991).

The 1991 FCRS national response rate was 67.9 percent, with a refusal rate of 24.9 percent and an inaccessible rate of 7.2 percent. Response rates on the survey have declined slightly over time, despite extensive efforts to limit nonresponse. While NASS uses farm expense data from the FCRS in its reports, the primary user of the FCRS dataset is the Economic Research Service (ERS), which utilizes all of the FCRS data in producing economic analyses and cost of production reports (Rutz, 1991).

A benefit of collecting this type of information is that survey managers can make adjustments to the public's perception of a too long interview by testing a shortened version of the questionnaire (as is being planned for the 1992 FCRS). Headquarters can prepare materials to aid survey statisticians in training their interviewers to meet the challenges of the refusal types common across states. Survey statisticians should develop materials for use in their state workshops to prepare interviewers for situations common to their state. Experienced interviewers who have had success in converting refusals into respondents should share their techniques through panel presentations or group discussions. In this way, interviewers will maximize response rates on the initial contact by being prepared to discuss concerns and grievances brought up by the respondents, thus avoiding the additional time and money costs of a re-contact.

BACKGROUND

The research project to identify and classify nonresponse on the FCRS stems from four years of preliminary work which the author completed while on staff in the South Carolina and Indiana State Statistical Offices (SSOs).

Beginning with the 1985 FCRS, the author required that the South Carolina interviewers document the reasons given by respondents who refused to participate in the survey. Previously, interviewers were likely to simply write "refusal" across the questionnaire, and the comments the interviewer received from a refusal were discussed second or third hand if at all, and were sketchy at best.

Then on the 1986 FCRS, South Carolina was selected as one of six states to take part in a refusal conversion research project. All respondents who refused to participate in the survey during the initial contact were to be re-contacted with the purpose of convincing them to complete an interview. It was apparent that interviewers selected to re-contact a refusal in the current survey had an advantage if they were aware of the reason the respondent gave when initially refusing.

The information on "reasons for refusing" gathered during 1985 were discussed during the training workshop for the 1986 FCRS, and responses to the reasons were developed by the interviewers. To prepare for the re-contact required by the research, interviewers were again required to write on the questionnaire the exact reason or circumstances behind each refusal received on the FCRS. In this way, subsequent interviewers were made aware of the events of the initial contact.

The primary benefit of identifying the refusal types was that the interviewers could PREPARE for common situations before encountering them in interview situations. According to interviewer comments, this preparation improved their confidence in approaching interviews, and even when they could not prevent a refusal, they were able to set the stage for the respondent's cooperation on other upcoming surveys. The second benefit was that, when approaching a re-contact on the refusal conversion project, the subsequent interviewer could prepare for a specific situation. A third benefit was that interviewers (with their supervisor's approval) could eliminate re-contacts of certain refusal types (violent refusals, death in the family, etc.), saving money and time during the critical data collection period.

Perhaps because the refusal conversion project was new and received much attention, or perhaps because the refusal identification preparation worked, the FCRS response rate in South Carolina for 1986 was 17 percent higher than in 1985 (Dillard, 1987). The author attributes most of this increase to interviewer preparation on the initial contact since only a small number of refusal conversions were obtained.

Upon transferring to the Indiana SSO, the author again instructed the field interviewers to document the reasons given by refusals. While the refusal identification and interviewer preparation led to an initial decrease from 35 percent to 31 percent in the refusal rate in Indiana, no additional gains have been evident, with the refusal rate averaging 31 percent over the past five years. The list of refusal types compiled during this time served as the basis of the refusal list utilized for the nonresponse identification project on the 1990 FCRS.

This research was conducted during February and March, 1991. The six test states included two states that averaged high nonresponse rates, two states that averaged mid-level nonresponse rates, and two states that averaged low nonresponse rates on the FCRS. Comments from the FCRS post-survey evaluations completed by survey statisticians around the country alluded to problems with certain refusal types, but with only anecdotal information to support their impressions. Evaluations included the following comments:

* "Some farmers feel it's none of our business."
* "Many farm operators refused due to the length of the questionnaire."
* "Most of the second time contacts were refusals and didn't want to be contacted again."

Some...many...most. The 1990 FCRS nonresponse identification project was expanded to all surveyed states for 1991 in order to put some numbers on these valid concerns and to better determine what NASS is up against when trying to minimize FCRS nonresponse.

RESULTS

The results of the 1991 refusal identification and classification research are listed in Appendix A.

Refusal types coded 01 - 53 were provided in the survey instructions; codes 200 - 409 were initially left blank for state use, and states added refusal types based upon their data collection experiences with the survey.

The most frequent reason given by the farmers when refusing to participate in the survey was "Would not take the time / too busy". This response was given by 1,395 of the 5,663 refusals encountered (24.6%), and was recorded nearly twice as often as the next most frequent response. This seems to be strong evidence for those involved with the survey who believe that farmers perceive the interview to take too long.

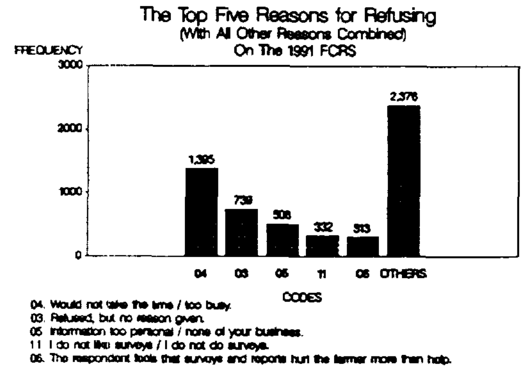
The second most frequent reason recorded was "Refused, but no reason given", mentioned 739 times, or 13.0 percent of the total refusals received. This category represents a difficult type of refusal to convert to a respondent: they just say NO. They may understand what NASS is and its mission, and may even recognize the interviewer from previous contacts, but cut off any attempt at an interview before their concerns can be identified and addressed.

The third most frequent reason recorded was "Information too personal / none of your business", mentioned 508 times, or 9.0 percent of the total refusals received. Together these first three reasons account for 46.7 percent of the total refusals received, and the top five reasons account for 58 percent, even though 52 different reasons for refusing were mentioned during this research.

Refusal reasons mentioned as frequently and as widespread as these five should be addressed on a national level. However, SSOs must review their state specific data to determine which less frequently mentioned reasons are important to their state.

This research also involved identifying and classifying the reasons given by an interviewer when coding a sampled unit inaccessible, shown in Appendix B. Inaccessible types coded 75 - 150 were provided in the survey instructions; codes 500 - 709 were initially left blank for state use, and the states added inaccessible types based upon their data collection experiences with the survey. While basically separate from the refusal identification, certain respondent situations (such as "Family illness / death") could be coded either as a refusal, an inaccessible or a valid zero out-of-business depending upon the circumstances encountered.

One benefit of this research is that the number of incomplete questionnaires, that is, those questionnaires for which the respondent could not or would not provide enough information for the interview to be completed, is evident for the first time. For the 1991 survey, 263 questionnaires were coded as incomplete and were not summarized. This amounts to 3.6 percent of the nonresponse, but is only 1.2 percent of the total survey contacts.



The most frequent inaccessible reason recorded by the interviewers was "Tried several times; could not reach anyone for an appointment. Just an extremely busy person.", given for 455 of the 1,653 inaccessibles encountered (27.5%). This is a surprising finding in light of the six week data collection period.

The second most frequent inaccessible reason recorded was "Illness / death in the family prevents the operator from responding", mentioned 182 times, representing 11.0 percent of the total. This is a difficult situation for an interviewer to encounter, and setting the stage to see a respondent under better circumstances in the future is the best that can be accomplished.

The third most frequent reason recorded was "Farm records are not available until after the survey period closes", mentioned 172 times, representing 10.4 percent of the total. Together these first three reasons account for 48.9 percent of the total inaccessibles recorded, with 23 different reasons for coding an inaccessible mentioned during this research.

SSOs must review their state specific data to determine which additional reasons are important to their state. For instance, "The operator is away on an extended vacation", normally thought to be a Midwest or Northern situation for escaping the snow, was also mentioned in California, Florida and other warm weather states.

DISCUSSION AND RECOMMENDATIONS

Data analysts, survey managers, statisticians and interviewers are concerned about the levels of nonresponse on the FCRS. Being close to the survey, they develop impressions about what factors are "driving" the nonresponse. The purpose of this research is to identify the reasons for nonresponse, and to attach some numbers to them in order to rank their relative importance. Considering the nature of the FCRS, that it is a long, detailed interview of a respondent's operating procedures, income and expenses, assets and liabilities and demographic information, many survey organizations would be thrilled to have a national response rate exceeding 70 percent. Rather than defend this position, the survey managers at NASS and ERS continually strive to improve the response rate on the survey.

Following a discussion of the preliminary results of this study and from previous consideration of the subject, NASS and ERS have agreed to test a shortened version of the questionnaire for the 1992 survey year. A detailed discussion of the benefits of a shortened questionnaire version can be found in Dillard (1991). NASS will provide training and materials to the survey statisticians at the regional workshops in January, 1993, to aid in training their field interviewers during state workshops. Additionally, the information is useful in the weighting of survey results and summarization.

According to Turner (1992) the FCRS nonresponse adjustment factor is based on an assumption that all nonrespondents are operating farms; that is, they would provide positive records if interviewed. Miss-coding valid zero reports as nonrespondents will positively bias the expanded indications. Turner (1992) states that, "Identifying these (nonresponse) reasons will enable enumerators to improve classification of cases where no farm appears to exist as a valid zero. Continued emphasis should be given to classifying only positives as refusals and inaccessibles. Those nonrespondents that have no indication of being in business should be coded as out of business."

Look at the pattern of nonresponse across the data collection period, and an interesting picture appears. In five of the seven survey weeks, more refusals occurred on Mondays than on any other single day, and during the other two weeks, the number of Monday refusals is near the peak for the week. This is probably a function of more interviews being attempted on Mondays, but it may also indicate that Mondays are not the best day to attempt a long interview without a prior appointment. Otherwise, the distribution of refusals seems normally spread throughout the survey period.

As might be expected, the number of inaccessibles peaks near the end of the data collection period when time constraints force the interviewers to begin to give up on respondents who either cannot be located or who continue to put off the interview when contacted. In general, the incomplete interviews seem normally spread throughout the data collection period.

As the results from the six test states in the 1990 research served as an excellent predictor of the 1991 results, there does not appear to be enough yearly variation to justify transferring this research into an operational aspect of the survey. I recommend that this research be repeated in three years. In this way, each SSO can be updated on the causes of nonresponse likely to be encountered, and patterns of nonresponse can be compared.

REFERENCES

Dillard, Dave and T. Gregory (1987). 1986 FCRS Analysis, Report II. Response Rates, Interview Times, and Data Collection Costs. United States Department of Agriculture, National Agricultural Statistics Service.

Gregory, Thomas L. (1990). 1989 Farm Costs and Returns Survey, Survey Administration Analysis. United States Department of Agriculture, National Agricultural Statistics Service, Agricultural Statistics Board, NASS Staff Report SMB Number 90-03.

Rutz, Jack L. and C.L. Cadwallader (1991). 1990 Farm Costs and Returns Survey, Survey Administration Analysis. United States Department of Agriculture, National Agricultural Statistics Service, Research and Applications Division, NASS Staff Report SMB Number 91-04.

Turner, Kay (1992). Modification of FCRS Nonresponse Adjustment Procedures. United States Department of Agriculture, National Agricultural Statistics Service, Research Division, SRB Research Report Number SRB-92-08.

APPENDIX A: Reasons Given By Respondents When Refusing To Participate on the 1991 Farm Costs and Returns Survey, All States and Versions Combined.

FREQUENCY	CODE	REASON
1,395	04.	Would not take the time / too busy.
739	03.	Refused, but no reason given.
508	05.	Information too personal / none of your business.
332	11.	"I do not like surveys / I do not do surveys."
313	06.	The respondent feels that surveys and reports hurt the farmer more than help.
255	02.	Contact attempted, but respondent refuses on all surveys, and refused on this one.
253	10.	"I will have nothing to do with the Government."
195	34.	Respondent will do other surveys, but not financial surveys.
135	20.	Family illness / death.
134	12.	Respondent only does compulsory surveys.
128	18.	The respondent feels the operation's records are inadequate to complete the interview.
120	16.	"My farm is too small to count / too small to be representative."
120	17.	"You contact me too often."
105	21.	Operator would not keep appointments.
97	19.	Farm records are at the tax advisors / lawyers.
95	07.	"I did this survey before, but not again."
89	01.	Known refusal, no contact attempted.
72	32.	"This is not a farm."
64	24.	Violent / threatening refusals.
58	52.	Questionnaire not sent to the field to avoid jeopardizing cooperation on other surveys.
56	27.	Respondent is quitting farming.
48	28.	Out of business now, will not answer for the previous year.
46	23.	Wants to be paid for interview time and effort.
42	08.	"I just did a different survey for your office."
40	22.	Spouse / secretary / etc. will not let the enumerator see the operator.
36	13.	The respondent does not think the information is kept confidential.
36	26.	Respondent does not want to report due to legal / financial problems.
30	25.	Respondent does not want to talk about farming.
29	14.	The respondent mentions a specific grievance with the SSO or NASS (other than confidentiality).
22	29.	Figures for the previous year were not typical.
18	09.	"I just did a survey for someone else."
18	53.	Would not answer the door even though they were home.
5	365.	Operator called the office after receiving the pre-survey letter, asked not to be contacted.
5	366.	The operator does not believe in statistics, so will not complete an interview.
4	15.	The respondent mentions a specific grievance with the state cooperator.
2	240.	Needed partner to provide some information; partner refused.
2	260.	Getting divorced, too upset to respond.
2	265.	Operator has a grievance with the IRS.
2	267.	Fed up.
1	215.	Water rights curtailed, will not cooperate.
1	250.	"The government is broke, how can we afford to send these people out?"
1	255.	NASS data is not accurate. Too political.
1	256.	Doing well financially -- does not want to respond.
1	257.	Operator has several operations and could not separate records for the sampled unit.
1	258.	Upset with the government -- has to spend $20,000 to dig up fuel tanks.
1	262.	Farmhouse and records lost in a fire, January, 1992.
1	269.	This survey is not needed.
1	270.	Responded previously on this survey, and asked to be excused this year.
1	335.	The respondent feels the operation is too complex for our survey.
1	340.	The respondent has a specific grievance with ASCS.
1	341.	The farm operation is in a blind trust for a national politician.
1	367.	His father would not do surveys, so neither will the son.

5,663 Total Responses

* Code numbers not listed were not used.

APPENDIX B: Reasons Given By Enumerators When Coding a Sample Unit as Inaccessible/Incomplete on the 1991 Farm Costs and Returns Survey, All States and Versions Combined.

FREQUENCY	CODE	REASON
455	116.	Tried several times; could not reach anyone for an appointment. Just an extremely busy person.
263	150.	INCOMPLETE -- Respondent provided partial information, but would not or could not provide enough information to make the questionnaire complete.
182	84.	Illness / death in the family prevents the operator from responding.
172	85.	Farm records are not available until after the survey period closes.
169	86.	Respondent postponed the interview beyond the end of the survey period.
142	79.	The operator is away on an extended vacation.
80	81.	The operator is away on business.
67	80.	The operator is away on a brief vacation.
27	76.	No respondent, as listed on the label, could be found.
26	94.	Inaccessible, but no reason given.
18	82.	The address on the label is summer-seasonal housing.
12	75.	No operation, as listed on the label, could be found.
9	83.	Access to the address on the label was denied by a gate / guard / etc.
7	78.	The address on the label is vacant / burned out / no structure exists.
7	87.	Enumerator workload prevented this operation from being contacted during the survey period.
5	591.	The operator moved away during 1991.
3	667.	The questionnaire was returned too late to be included in the summary.
2	92.	Non-English speaking respondent; interpreter not available.
2	119.	Enumerator mistake; caught it too late to complete an interview within the survey period.
1	120.	Operator has several operations and could not separate records for the sampled unit.
1	540.	Questionnaire from the enumerator lost in the mail.
1	561.	Operator had just gotten out of jail and would not talk with anyone from the government.
1	565.	Enumerator did not contact sufficiently; gave up too soon.
1	580.	Enumerator error, should not have collected the data.

1,653 Total Responses		

* Code numbers not listed were not used.

AN EVALUATION OF NONRESPONSE ADJUSTMENT WITHIN WEIGHTING CLASS CELLS FOR THE FARM COSTS AND RETURNS SURVEY

Kay Turner, USDA/NASS
Research Division Room 305, 3251 Old Lee Hwy Fairfax, VA 22030

INTRODUCTION and OBJECTIVES

The Farm Costs and Returns Survey (FCRS) is conducted by the National Agricultural Statistics Service (NASS) during February and March of each year. The data are collected in the 48 contiguous States from farm operators/managers for the preceding year via personal interviews. Various versions of the FCRS collect detailed and aggregate expenditure, income, asset, liability and cost of production data. The data from the FCRS are used to ascertain the financial status of the agriculture sector by supplying information such as: farmers' net income, costs of producing commodities, financial situation of farm operators, debt held by farm operators, and importance of production expense items. Farm organizations, agribusinesses, Congress, the Department of Agriculture, farmers, and ranchers are some of the groups that utilize FCRS data (NASS, 1989). Each year a sample is drawn for the FCRS using both list and area frames. The list frame includes mainly large and specialty operations. The area frame includes small operations not on the list frame, or nonoverlap (NOL) (NASS, 1991).

Nonresponse exists because all sampled farm operators do not respond to the survey. The two types of nonrespondents are refusals (the farm operator declines the interview) and inaccessibles (the farm operator cannot be contacted). Kalton and Maligalig (1991) note,

> "When total nonresponse occurs, the survey analysis may simply be carried out on the data provided by the responding elements. However, since responding and nonresponding elements may differ systematically in their survey characteristics, there is a risk with this approach that the survey estimators will be biased. It is therefore a common practice to attempt to compensate for the missing data arising from total nonresponse by some form of weighting adjustment".

Previous analysis (Turner, 1992) has indicated that FCRS direct estimates at the U.S. level for five major variables over the years 1987-1990 are biased downward as follows: three major expense items are biased downward about 10%, while land in farms and number of farms are biased downward about 20%. An inappropriate nonresponse adjustment for the list frame portion of the multiple frame (MF) estimate and undercoverage of farms are major causes of this bias. The 1990 FCRS nonresponse adjustment procedures will be referred to as the current procedure. Currently, FCRS data are collected under the following assumption.

> Assumption a: All nonrespondents would qualify for an interview and would have some positive responses to the survey (i.e., are positive records).

In the supervising and editing manual, field enumerators are instructed to code all out of business (zero) records, who would not qualify for an interview, as respondents. These instructions are intended to ensure that all nonrespondents would qualify for an interview, i.e., have an agricultural operation. Since all interviews are face to face, it is possible to determine if a record is in business or not. The underlying assumption of the current list frame nonresponse adjustment factor, which assumes nonrespondents are similar to all respondents, conflicts with Assumption a because the adjustment assumes nonrespondents can include positive and zero records. The current area frame nonoverlap (NOL) nonresponse adjustment factor, which is applied at the State level, assumes nonrespondents are all positive records and is consistent with Assumption a.

Objectives 1 and 2 of this study involved the application of a simple adjustment (which is consistent with Assumption a) to list frame sample records using the following weighting classes: 1) the design strata, and 2) type/size cells over strata. Objective 3 examined the effect of applying the adjustment at a type/size cell level to area frame NOL records. Weighting classes or cells based on farm type and economic size are intended to provide more homogeneity within weighting classes and heterogeneity across weighting classes than the current classes (strata for the list and States for the area NOL) provide. If the weighting classes are effective in capturing this homogeneity within and heterogeneity across classes with respect to response probabilities, they will help reduce nonresponse bias. Previously reported control data were used to place nonrespondents into appropriate type/size cells.

EXPANSION FACTORS

The area frame sampling unit is a segment of land, usually about one square mile in area, within a land use stratum. Area frame reporting units are residents of the sampled segments who reported agricultural activity on the previous June Agricultural Survey (JAS), and who are NOL with respect to the FCRS list. The list frame sampling unit is a name on the list sampling frame (LSF). The reporting units are all operating arrangements associated with the sampled names. In the following notation, let h denote a sampling stratum; c denote a type/size weighting cell within a State; and s denote a State.

Furthermore, let
t = h, c, or s as appropriate,
N(t) = number of sampling units in the population denoted by t,
n(t) = number of sampling units sampled from the population denoted by t,
g(t) = number of positive respondent reporting units in t,
f(t) = number of zero respondent reporting units in t,
r(t) = g(t) + f(t) = number of respondent reporting units in t,
e(t) = number of positive nonrespondent reporting units in t,
j(t) = number of zero nonrespondent reporting units in t, and
m(t) = e(t) + j(t) = number of nonrespondent reporting units in t.

Finally, let
$r^*(t)$ = the number of respondent sampling units in t, and
$m^*(t)$ = the number of nonrespondent sampling units in t.

For a sampling unit of the area frame to be classified as nonrespondent, the interviews of all qualifying residents in a land segment must be coded as refusals and inaccessibles. For the list frame, there is usually one reporting unit per sampling unit. If the reporting unit refuses or is inaccessible, then it is a nonrespondent sampling unit. When there is more than one reporting unit associated with a list frame sampling unit, these operating arrangements are referred to as multiple operations. A nonrespondent sampling unit exists in the case of multiple operations when all of the questionnaires corresponding to the sampled name are classified as refusals and inaccessibles.

The current list frame expansion factor is

$$EF = \frac{N(h)}{n(h)} * \frac{n(h)}{r^*(h)} . \qquad (1)$$

The FCRS summary currently has two methods for adjusting the list and area frames for nonresponse due to refusals and inaccessibles. Both procedures are described below. Each sampled unit is initially assigned an original expansion factor that would be applicable if there were no nonresponse, that is, if a usable report was obtained from each reporting unit. For both the area and list frames, the original expansion factor is the first term of Equation (1). The corresponding assumption of this term is the following.

> Assumption b: the n(h) sampled units in stratum h are a simple random sample of sampling units from the N(h) population units in the stratum.

This assumption is clearly true. Since all reporting units do not respond, the original expansion factor is multiplied by an adjustment factor to account for the nonrespondent reporting units. The second term of Equation (1) is based on the following assumption.

> Assumption c: the $r^*(h)$ respondent sampling units in stratum h are a simple random sample from the n(h) sampled units.

If Assumption c were true, then the $m^*(h)$ nonrespondent sampling units would also be a simple random sample of the n(h) sampled units in stratum h. This contradicts Assumption a, where all nonrespondents are assumed to be positive.

The following expansion factor is designed to be consistent with Assumption a and meet Objectives 1, 2, and 3 of this study. The level at which the nonresponse adjustment is calculated, which is represented by x, varies and will be described below.

$$EF = \frac{N(h)}{n(h)} * \frac{g(x)+e(x)}{g(x)} . \qquad (2)$$

The modified list frame expansion factors for Objectives 1 and 2 each have the form of Equation (2) where the nonresponse adjustment factor (term two) is **calculated at the stratum level (x = h) for Objective 1 and at the type/size cell level (x = c) for Objective 2.** These nonresponse adjustment factors are consistent with FCRS Assumption a (i.e. all nonrespondents are positives) since they are based entirely on positive records. The nonresponse adjustment factors of Objectives (1) and (2) are based on the following assumption.

> Assumption d: the positive respondent reporting units {g(h), g(c)} are a simple random sample from the positive reporting units in the stratum or weighting cell.

The current area frame expansion factor has the same form of Equation (2) where the nonresponse adjustment factor (term two) is calculated at the State level (x = s). The nonresponse adjustment factor (term two) of Equation (2) is **applied at the type/size cell level (x = c) for Objective 3.** The nonresponse adjustment factors of the current area frame expansion factor and Objective 3 are both based on Assumption d above.

DATA DESCRIPTION

For this project, 1990 FCRS data were used. The variables that were examined in the analysis are: total expenses, livestock expenses, labor expenses, land in farms, and number of farms. Nine States (Arizona, Colorado, Georgia, Illinois, Kansas, Montana, New York, North Carolina, and Wisconsin) could not be included for the list frame type/size cell analysis because the control data for size were missing. Control data for list records were obtained from the list sampling frame. For area NOL records, control information was collected on the previous June Agricultural Survey. Type categories were collapsed into two classes: crops and livestock. The following five size cells were chosen with respect to annual total gross value of sales: 1} 1 to 9,999, 2} 10,000 to 39,999, 3} 40,000 to 99,999, 4} 100,000 to 249,999, and 5} 250,000 plus. Since variance inflation can result when adjustment factors are not based upon adequate sample sizes, a goal of at least 20 positive respondent records with control data per weighting class was set. (Cox, 1991). To ensure uniform collapsing of cells, a priority scheme and logic flowchart were followed.

RESULTS

Objective 1

Expansions and CV's were obtained for the five variables using the current list frame nonresponse adjustment factor, term two of Equation (1), applied to each of the 281 strata in the 39 States. The modified nonresponse adjustment factor, term two of Equation (2), was applied to each of the 281 strata in the 39 States for Objective 1. The modified nonresponse adjustment factor by stratum produced expansions approximately 9% to 10% higher than the current expansions. Four of the CV's are slightly greater than those of the current method and one CV is the same.

Objective 2

Term two of Equation (2) was applied by type/size cell within State to evaluate Objective 2 for list frame estimates. A total of 212 cells were used over the 39 States. These estimates are 10% to 17% greater than the current estimates. The CV's tend to be slightly larger than those for the unadjusted expansions or for adjusted expansions at the stratum level.

Objective 3

To evaluate Objective 3, records were assigned to area frame NOL type/size cells within State using the same logic used for the list frame records. A total of 68 cells were used for Objective 3. The estimates of Objective 3 are very near the current NOL estimates. Three of the CV's are less than those of the current method and two are greater. Since the percentage change in the estimates is small for Objective 3, these results indicate that application of the nonresponse adjustment for the area frame NOL within cells has negligible effect.

MULTIPLE FRAME RESULTS

List and area NOL results have been considered separately. Multiple frame results show the effect of the list and area NOL results together. Agricultural Statistics Board numbers, which are considered to be truth, exist for number of farms and land in farms. For the three expense items, "Pseudo Board" values (Turner, 1992) were calculated that adjust somewhat for the FCRS undercoverage of farms. The Pseudo Board values represent a minimum value of truth since there are other factors that also contribute to the downward bias. Nonresponse adjusted MF estimates were calculated at the 48 State level using type/size cells within State for both the list and area NOL indications. The list data for the nine States with unknown size control data were expanded by stratum using the modified nonresponse adjustment factor, since type/size cells could not be created. The probable effect of using the modified nonresponse adjustment by stratum for these nine States on the 48 State MF indications, instead of using the modified nonresponse adjustment by type/size cell within each State, is to bring the indications downward. These nonresponse adjusted MF estimates as well as the current MF estimates arecompared to the Board and Pseudo Board estimates in Table 1. The nonresponse adjusted MF estimates for the expense items closely match their Pseudo Board values, ranging from 3.7% below to 1.3% above. Land in farms adjusted for nonresponse is still biased downward by about 13%. This bias is probably due in part to the tendency of farm operators to underreport total farm acreage (McClung, 1988). However, this bias is about 8 percentage points smaller than the current bias of 21%. This reduction in bias, represented by the last column in Table 1, for land in farms is comparable to the reduction for the expense items, indicating about an 8 to 11 percentage point effect on the MF estimates for these items. One important characteristic of these four items is that approximately 23% of the MF estimates are from the area frame NOL. The reduction in bias for number of farms is only about 4 percentage points, but approximately 58% of the MF estimate is from the area frame NOL. Since the nonresponse adjustment had negligible effect on the area frame NOL, the bias reduction for the MF estimate is also small.

Table 1: 1990 Current MF Estimates and Nonresponse Adjusted MF Estimates Using Type/Size Cells Within State at 48 State Level Compared to 1990 Board and Pseudo Board Estimates.

Item	**1990 Board & Pseudo Board Estimates (mil.)**	**Current MF (% of Board)**	**Nonresponse Adjusted Type/Size Cells MF (% of Board)**	**Nonrs. Adjstd. (% of Board) - Current MF (% of Board)**
Total Expenses	150,269	87.9%	96.3%	8.4%
Livestock Expenses	16,864	88.9%	97.1%	8.2%
Labor Expenses	14,828	90.1%	101.3%	11.2%
Land in Farms	985	78.8%	86.6%	7.8%
No. of Farms	2.1352	82.1%	85.8%	3.7%

CONCLUSIONS

Results indicated that the largest bias reduction for the list frame portion of the estimate occurred using type/size cells over strata. Evidently, these cells do a more effective job of grouping homogeneous records together than the current design strata. There was little effect, however, from using type/size cells for area frame NOL records primarily because cells could only be created in 17 of the 48 States because of the goal of at least 20 records per cell. A major factor to the remaining downward bias on all five items is the undercoverage of farms by FCRS. The CV's of the nonresponse adjusted estimates increased slightly as compared to the current CV's. This probably reflects more the failure of the variance approximation procedure than the nonresponse adjustment procedures.

RECOMMENDATIONS

Analysis of 1990 data indicated the adjustment should be made using type/size weighting classes within each State for the list frame records. The recommendation for Objective 3 was optional, since the impact of type/size cells within State was negligible on the area side. It was recommended that analysis be conducted on the 1991 FCRS data to determine if type/size weighting classes within each State were needed, or if the list frame strata were adequate weighting classes. The 1992 FCRS ·used the modified nonresponse adjustment at the design stratum level, since 1991 list frame stratification changes were expected to better account for type and size of farm and since the creation of type/size cells would have added complexity to the summary process. Since the nonresponse adjustment is based on the assumption that all nonrespondents have operating farms, survey training materials and instructions should continue to emphasize that refusal and inaccessible sampling units must be farm operators.

REFERENCES

(1) Cox, Brenda G. (1993), "Weighting Class Adjustments for Nonresponse in Integrated Surveys: Framework for Hog Estimation," SRB Research Report Number SRB-93-03, National Agricultural Statistics Service.

(2) Kalton, Graham and Dalisay Maligalig (1991), "A Comparison Of Methods Of Weighting Adjustment For Nonresponse," 1991 Annual Research Conference Proceedings, Bureau of the Census, pp. 409-428.

(3) McClung, Gretchen (1988), "A Commodity Weighted Estimator," Staff Report No. SRB-88-02, National Agricultural Statistics Service.

(4) National Agricultural Statistics Service (1989), "Interviewer's Manual 1989 Farm Costs & Returns Survey (FCRS)," Author.

(5) National Agricultural Statistics Service (1991), "1990 Farm Costs and Returns Supervising and Editing Manual," Author.

(6) Turner, Kay (1992), "Modification of FCRS Nonresponse Adjustment Procedures," SRB Research Report Number SRB-92-08, National Agricultural Statistics Service.

ACKNOWLEDGMENTS

The author appreciates the significant contributions of time and effort by Bill Iwig and Dave Dillard. Thanks to Fatu Wesley for her valued support and advice, to Jim Burt for variance analysis, to Fred Warren and Martin Ozga for downloading the data, to Dick Clark and Dan Ledbury for supplying FCRS and Board numbers, and to Phil Kott for his valuable review of the original NASS staff report (Turner, 1992). Sincere thanks to everyone who provided assistance on this project.

SELECTED RESULTS OF THE INCENTIVE EXPERIMENT ON THE 1992 FARM COSTS AND RETURNS SURVEY

Diane K. Willimack, National Agricultural Statistics Service, U.S. Department of Agriculture
Survey Management Division, Room 4151, South Building, Washington, D.C. 20250-2000

ABSTRACT: A split ballot experiment was conducted on the 1992 Farm Costs and Returns Survey (FCRS) in four States to test the effects of a prepaid nonmonetary incentive on response rates and related variables. Results showed a statistically significant improvement in response rates of 5.4 percentage points due to the incentive. The incentive appeared to be the most effective among farms stratified in the smallest and largest sales classes. In addition, the incentive appears to have enhanced identification of non-eligible sample units (non-farms) over the No Incentive group, reducing a potential nonsampling error.

KEY WORDS: Incentive, prepaid, nonmonetary, split ballot experiment, response rates

INTRODUCTION

The Farm Costs and Returns Survey (FCRS) is a nationwide multiple frame survey of U.S. farm operators, conducted annually by the National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture (USDA). The purpose of the FCRS is to gather data for estimating total expenditures, net farm income, cost of production for selected agricultural commodities, and other economic indicators of the financial condition of the agriculture sector. In addition, the FCRS provides a data base for price index construction and microeconomic analyses.

The FCRS is considered by NASS personnel, from field enumerators to headquarters staff, to be a challenging survey. Face-to-face interviews averaging just over 90 minutes in length request detailed expenditure and income information, which is highly sensitive in nature, from individual farm operators. It is not surprising that the FCRS has suffered low U.S. level response rates that have declined in recent years from 69 percent to 63 percent (Rutz, 1993).

Concern for declining response rates has prompted NASS, like other survey organizations, to search for methods by which potential respondents may be encouraged to participate in its surveys, particularly the FCRS. In recent years, extensive broad-based public relations materials have been disseminated in the popular farm press, and special refusal conversion efforts have been attempted. In addition, FCRS participants are routinely offered an Individual Farm Financial Analysis, which summarizes a number of economic characteristics of each respondent's own farm in comparison with farms of similar type and size in the same State, based on data from the survey. While none of these efforts is considered inconsequential, they have been met with little or no apparent success at actually increasing response rates.

In order to continue to explore methods for increasing response to the FCRS, NASS decided to investigate the use of incentives, a method often considered in survey research for the purpose of influencing survey participation. In such studies, typically a monetary or nonmonetary incentive is either prepaid unconditionally or promised for survey completion. In the study reported in this paper, a split ballot experimental design was utilized in four States to test the effect of a prepaid nonmonetary incentive on response rates on the 1992 FCRS.

LITERATURE REVIEW

Incentives have long been used in mail surveys, particularly of the general population, in order to increase response rates. A wide body of published literature suggests that monetary incentives are more successful at eliciting response than are nonmonetary gifts, and that prepaid incentives are more effective than rewards promised upon return of completed survey instruments. See for example qualitative review articles by Armstrong (1975) or Linsky (1975), or quantitative meta-analyses by Church (1993), Heberlein and Baumgartner (1978), or Yu and Cooper (1983). In particular, Church (1993) found promised incentives of any kind, monetary or nonmonetary, to be ineffective, asserting that "[t]hese types of incentive plans are simply not worth the energy involved" (p.75).

A recent study reports the results of an incentive experiment in an establishment survey, a mail survey of small construction subcontractors asking for detailed information about employees' health insurance coverage (James and Bolstein, 1992). Tested for their effect on response rates were prepaid monetary incentive amounts increasing from $1 to $40, along with a promised monetary incentive of $50 contingent upon survey return. Each of these was coupled with up to four follow-up mailings. The promise of $50 performed no better than the control condition of no incentive with multiple mailings. Although prepayment of $20 maximized survey response, it was not considered cost effective, and the authors concluded that token amounts such as $1 or $5 were sufficient to influence survey participation.

James and Bolstein attribute their positive incentive effects to social exchange theory: "By giving money the researcher extends a token of trust to the survey participant and initiates a social exchange relationship which invokes a social obligation for the participant to reciprocate in kind" (p.451). Social exchange theory and the principle of reciprocation are typically invoked to explain the positive effects of incentive use in surveys of households and among the general population (Dillman, 1978; Groves, Cialdini, and Couper, 1992). Applying this reasoning to establishment surveys is a nontrivial matter because it is unclear who in an establishment we are attempting to influence with the use of an incentive. Incentives are meant to influence the survey participation decision and to motivate the respondent. However, in an establishment survey, the decision-maker may not be the actual respondent, the most knowledgeable provider of the information being sought. This is typically the person who has access to and understanding of any records to be used as a source for responding (Edwards and Cantor, 1991).

Farmers, on the other hand, are likely to be both the decision-maker regarding survey participation and the most knowledgeable provider of the information sought on the FCRS. Hired accountants or record-keeping services were used routinely for keeping farm financial records by less than half of the farms with sales of $250,000 or more in 1987. Moreover, farms of this size account for only about 5 percent of all U.S. farms. Approximately 11 percent of all farms subscribed to record-keeping services in 1987. Although the proportion of farmers using outside services for financial record-keeping may have increased in recent years, it is still likely that the majority of farm operators are the most knowledgeable source of the financial information being requested by the FCRS, since evidence suggests that even service users have some close hands-on method of keeping track of finances, such as a workbook or ledger (Willimack, 1989). Furthermore, almost 99 percent of U.S. farms were sole proprietorships, legal partnerships, or family-held corporations in 1990. These operations are, by definition, family farms and are attached to farm operator households (Ahearn, Perry, and El-Osta, 1993). Indeed, "family farms .. meet either definition," households or establishments (Edwards and Cantor, 1991). Farmers may call upon a decision rule not unlike that of a householder regarding a request for survey participation. Thus using incentives to influence farm operators takes advantage of their characteristics as 1) the person most knowledgeable of the information requested, 2) the decision-maker regarding survey participation, and 3) a householder.

DESIGN AND METHODOLOGY

The FCRS utilizes a multiple frame design, consisting of a list frame and a complementary area frame. The list frame accounts primarily for medium, large, and specialty farms. For the FCRS, the operations on the list are stratified by type and size, including only those with "estimated value of sales" of $20,000 or more. The area frame, which is stratified geographically and by land use, covers farms that are missing from the list, called nonoverlap (NOL). NOL farms are typically, though not exclusively, small farms. The U.S. list and area NOL sample size for the 1992 FCRS was 21,000.

FCRS data are collected using several questionnaire versions in face-to-face interviews averaging 90 minutes in length. Data for the previous calendar year are collected during February and March of the current year, which coincides with the end of the traditional tax preparation period for U.S. farms. Any operation having agriculture qualifies for an interview, although it may not qualify as a farm. The official USDA definition of a farm is any establishment which sold or would normally have sold at least $1,000 of agricultural products during the previous year.

The 1992 FCRS Incentive Experiment utilized a "split ballot" experimental design in four States -- Georgia, Idaho, Kansas, and Michigan -- representative of historically low or declining response rates, as well as agricultural and geographic diversity. Sample elements were randomly assigned to the two incentive treatment groups by version and by stratum, resulting in 1181 sample units in the Incentive Group and 1183 sample units in the No Incentive Group.

The nonmonetary incentive item was a vinyl pocket portfolio containing a notepad and a removable solar calculator. It bore an imprint identifying the survey, along with the State's and the Agency's identification. It was token in nature and not intended to represent compensation for the respondent's time and survey participation. The item cost $7.06 per unit to produce.

The incentive item was prepaid; its receipt was not contingent upon survey participation. It was enclosed along with the pre-survey notification letter mailed to sample units prior to interviewer contact. Letters containing the incentive item acknowledged its enclosure, saying,"While this token in no way equals the value of your contribution to this survey, it represents our appreciation for your consideration." Letters to sample units in the No Incentive Group did not include the incentive item, nor did they mention it. Other than the above additional statement, letters in both groups were identical, written and signed by the State Statistician of each participating State. All pre-survey letters were addressed to the farm operator or operation by name and were mailed 1-2 weeks prior to the beginning of FCRS data collection in each State.

Survey interviewers were informed of the identity of incentive recipients and nonrecipients by way of a special code on each questionnaire label. They were allowed to use this knowledge in an appropriate manner upon contact with incentive recipients only. They were not allowed to personally carry or possess the incentive item, nor were they allowed to promise it as a gift for survey completion among nonrecipients.

On the FCRS, the various sample disposition rates have the following definitions:

$$\text{Response Rate} = \frac{\text{Number of Completed Interviews}}{\text{Number Eligible for FCRS}}$$

$$\text{Refusal Rate} = \frac{\text{Number of Refusals}}{\text{Number Eligible for FCRS}}$$

$$\text{Eligibility Rate} = \frac{\text{Number Eligible for FCRS}}{\text{Number of Sample Units}}.$$

NASS traditionally reports response rates in terms of sample counts, with each sample unit being given equal weight. That is, these calculations do not take account of the sample design, and sampling weights are not applied.

RESULTS

Although both unweighted and weighted analyses were conducted for this paper, only the unweighted results will be reported in detail, as this appears to be more common in the survey literature. (By our definition, "unweighted" analysis completely ignores the sample design, while a "weighted" analysis incorporates the design.) Reference will be made to weighted analysis in summary, and exceptions will be noted. In all cases, a one-tail test of significance is of primary interest. The decision criteria will be relative to a one-tail confidence level of alpha=0.10, since the wide variability of sampling weights in the NASS stratified sample design is not particularly beneficial to the estimation of proportions. Nevertheless, two-tail p-values will always be reported, allowing the reader to make his/her own judgments.

The response rates were 63.3 percent in the Incentive Group and 57.9 percent in the No Incentive Group, representing a statistically significant increase in response rates due to the incentive of 5.4 percentage points (two-sided p-value=0.009). The reduction of nearly 5 percentage points in refusal rates, from 29.9 percent in the No Incentive Group to 25.0 percent in the Incentive Group, is also statistically significant (two-sided p=0.011). See Table 1.

Furthermore, the eligibility rate of 90.8 percent in the Incentive Group is significantly lower than the comparable rate of 93.2 percent in the No Incentive Group (p=0.034). That is, sample units that had no agriculture, and thus were not eligible for the FCRS, were more likely to be identified among incentive recipients than among nonrecipients. The lack of eligibiity emanates primarily from the list frame. The farm status of individual list records has likely not been determined as recently as operations sampled from the area frame, which have been contacted at least once in the past year.

Table 1: Response Rates, Refusal Rates, and Eligibility Rates, by Incentive Group, 1992 FCRS Incentive Experiment 1/

	Incentive	No Incentive	Two-tail p-value 2/
Response Rate	63.3%	57.9%	0.009
(n)	(1072)	(1102)	
Refusal Rate	25.0	29.9	0.011
(n)	(1072)	(1102)	
Eligibility Rate	90.8	93.2	0.034
(n)	(1181)	(1183)	

1/ Multiple frame, unweighted, ratio estimates.

2/ H_0: Rate(Incentive) = Rate(No Incentive).

Table 2: State-level Response Rates and Refusal Rates, by Incentive Group, 1992 FCRS Incentive Experiment 1/

State	Incentive	No Incentive	Two-tail p-value 2/
Georgia			
Response Rate	73.5%	66.0%	0.070
Refusal Rate	18.5	23.0	0.227
(n)	(238)	(244)	
Idaho			
Response Rate	75.3	67.9	0.061
Refusal Rate	19.2	25.7	0.078
(n)	(255)	(265)	
Kansas			
Response Rate	44.0	41.2	0.472
Refusal Rate	37.2	42.4	0.183
(n)	(309)	(311)	
Michigan			
Response Rate	64.8	59.9	0.235
Refusal Rate	22.2	25.9	0.314
(n)	(270)	(282)	

1/ Multiple frame, unweighted, ratio estimates.
2/ H_0: Rate(Incentive) = Rate(No Incentive).

As can be seen in Table 2, the incentive appears to have increased response rates and reduced refusal rates in each of the four States. Even though the experiment was not specifically designed with sufficient sample sizes to support State level inferences, the response rate differences in Georgia and Idaho were large enough to reach significance. In addition, the refusal rate is significantly reduced (according to a one-tail test with alpha=0.10) by the incentive in Kansas, a State that has historically found FCRS survey participation to be particularly troublesome.

Table 3 shows the distribution of Incentive and No Incentive response rates and refusal rates by "estimated farm value of sales," the variable used explicitly for stratification purposes in the list frame and for classification purposes in the area frame. Here an interesting phenomenon appears. Increases in response rates due to the incentive are highly significant in the smallest size class and in the largest size class only, although small mid-size farms in the $40,000-$99,999 sales class appear to have been significantly influenced by the incentive as well. Two size classes, $20,000-$39,999 and $250,000-$499,999, exhibit nonsignificant reversals, where the Incentive Group suffered lower response rates than the No Incentive Group. However, these reversals become negligible as well as nonsignificant when the sample design is considered and appropriate sampling weights are applied.

Table 3: Response Rates and Refusal Rates, by Stratified Farm Value of Sales Class,by Incentive Group, 1992 FCRS Incentive Experiment 1/

Farm Value of Sales	Incentive	No Incentive	Two-tail p-value 2/
less than $20,000			
Response Rate	81.3%	64.3%	0.001
Refusal Rate	8.2	22.1	0.001
(n)	(134)	(140)	
$20,000-$39,999			
Response Rate	61.6	68.0	0.364
Refusal Rate	24.4	19.4	0.409
(n)	(86)	(103)	
$40,000-$99,999			
Response Rate	69.4	62.2	0.145
Refusal Rate	21.7	29.4	0.094
(n)	(180)	(177)	
$100,000-$249,999			
Response Rate	59.7	57.4	0.593
Refusal Rate	29.3	30.9	0.696
(n)	(273)	(256)	
$250,000-$499,999			
Response Rate	56.0	61.8	0.327
Refusal Rate	35.2	27.4	0.160
(n)	(125)	(157)	
$500,000 or more			
Response Rate	57.7	46.1	0.007
Refusal Rate	26.6	38.7	0.003
(n)	(274)	(269)	

1/ Multiple frame, unweighted, ratio estimates.
2/ H_0: Rate(Incentive) = Rate(No Incentive).

Estimation taking account of the sample design enables inference of these experimental findings to the population of U.S. farm operators. Although the level estimates change when sample weights are applied, the direction of the differences and their statistical significance are retained in all cases reported above, except in the largest sales class. Moreover, when analysis is limited to list frame sample elements, this largest size class exhibits a highly significant increase in response rates of more than 12 percentage points (p=0.006), due to the incentive. It appears that the loss of significance in the multiple frame estimate is due to the influence of an area frame refusal or inaccessible sample element with a large sample weight.

In order to further understand the effects of the incentive, additional analysis was undertaken using the list frame sample only. The greatest sampling efficiency is gained in the list frame since it accounts for the largest farms, which contribute the most to the FCRS estimates of expenditures and income. Thus, response from these list frame sample elements is crucial to data quality. However, it is well documented that FCRS response rates decline as farm size increases (Rutz, 1993). Therefore, it is important to evaluate the effect of the incentive in the list frame.

A regression was performed incorporating the sample design using data from list frame records only. Response behavior was regressed on incentive receipt, the stratification variable "estimated farm value of sales," as well as a term representing their interaction, along with variables that controlled for State and questionnaire version effects. While the incentive variable by itself proved nonsignificant, its interaction with the size variable significantly increased the probability of response of a list sample unit. The selected estimated equation is:

$$\begin{aligned}
&(\text{Probability of Response}) * 100\% = \text{Intercept} \\
&- \underset{(p=0.067)}{3.29\, \mathit{ln}(\text{Size})} + \underset{(p=0.091)}{0.56\, [\mathit{ln}(\text{Size}) * \text{Incentive}]} \\
&+ \text{State variables} \\
&+ \text{Questionnaire Version variables,}
\end{aligned}$$

where Size = Estimated farm value of sales

and Incentive = 1 if incentive,
0 if no incentive.

Incentive receipt appears to have significantly offset the decline in the probability of response among larger farms.

CONCLUSIONS

Clearly the inclusion of a nonmonetary incentive with the pre-survey notification letter resulted in a significantly higher response rate on the 1992 FCRS in the four States included in the experiment. The increase in response rates appears primarily due to a reduction in refusals among those who received the incentive. In addition, the incentive appears to have enhanced identification of non-eligible sample units (non-farms) over the No Incentive control group, reducing a potential nonsampling error. Incentive recipients who had no agriculture may have been more attentive to the survey request and more determined to notify the interviewer of their non-farm status, rather than to become refusals or inaccessible sample units.

Furthermore, the incentive appears to have had the largest and most significant effects on response rates among the smallest farms from the NOL area frame sample and among the operations sampled from the largest farm strata on the list frame. These two groups of farms have in common high likelihood of repeated contact for NASS surveys. Area frame NOL farms remain in the NASS sample for five years. Those selected for the FCRS are in their fourth and fifth years of their sample rotation and have likely been called upon repeatedly for a variety of NASS surveys, for it is their role to account for the incompleteness of the list. The largest list frame operations, likewise, are frequently subject to repeated contact for NASS surveys. As the largest operations, they account for a substantial portion of agricultural production. In addition, their higher degree of variability results in their being sampled at a higher rate.

The incentive, as a token item, cannot have compensated respondents, especially the large farms, for their FCRS participation. Thus these results cannot be interpreted using an argument based on economic exchange. Instead, the effects of the incentive must be interpreted in a social context. The incentive may have 1) drawn attention to the pre-survey letter, 2) legitimized the survey request, 3) identified and differentiated the survey sponsor (this is especially relevant since the U.S. Census of Agriculture was conducted just prior to the 1992 FCRS), 4) notified the farmer of the impending visit by an interviewer, and 5) offered a sign of appreciation, enabling the trust necessary for social exchange to occur.

Indeed, the latter explanation drawing upon social exchange theory, may be most appropriate, given that the two groups most affected by the incentive, the smallest NOL farms and the largest list frame farms, are those most frequently requested to participate in NASS surveys. It is likely that a relationship, rapport, has been built between these farm operators and the interviewers. Unconditional receipt of the incentive item may have been perceived by these FCRS respondents as a token of appreciation consistent with the ongoing social relationship of survey contact. This may have been particularly meaningful among the smallest farms, which are noncommercial in nature, and their operators may not even consider themselves to be farmers. The incentive may have symbolized the trust that, according to Dillman (1978), is necessary for social exchange to successfully occur, invoking social norms in the respondent consistent with survey participation. Thus, like James and Bolstein (1993), it seems reasonable to attribute positive incentive effects in the FCRS, an establishment survey of farmers, to social exchange theory.

ACKNOWLEDGMENT

The author wishes to acknowledge the technical assistance of several staff members of the Survey Research Branch of NASS, particularly Jim Burt and Phil Kott, who advised on analysis of the data taking account of the sample design.

REFERENCES

Ahearn, M. C., J. E. Perry, and H. El-Osta (1993) "Economic well-being of farm operator households, 1988-1990." AER-666, Economic Research Service, U.S. Department of Agriculture, Washington, D.C.

Armstrong, J. S. (1975) "Monetary incentives in mail surveys." Public Opinion Quarterly 39:223-250.

Church, A. H. (1993) "Estimating the effects of incentives on mail survey response rates: A meta-analysis." Public Opinion Quarterly 57:62-79.

Dillman, D. A. (1978) Mail and Telephone Surveys: The Total Design Method. New York: John Wiley and Sons.

Edwards, W. S., and D. Cantor (1991) "Toward a response model in establishment surveys," Ch. 12 in Measurement Errors in Surveys, edited by P. P. Beimer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman. New York: John Wiley and Sons, Inc., pp. 211-233.

Groves, R. M., R. B. Cialdini, and M. P. Couper (1992) "Understanding the decision to participate in a survey." Public Opinion Quarterly, 56:475-495.

James, J. M., and R. Bolstein (1992) "Large monetary incentives and their effect on mail survey response rates." Public Opinion Quarterly 56:442-453.

Linsky, A. S. (1975) "Stimulating responses to mailed questionnaires: A review." Public Opinion Quarterly 39:82-101.

Rutz, J. L. (1993) "Farm Costs and Returns Survey for 1991 and 1992: Survey Administration Analysis." Staff Report #SAB-93-01, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C.

Willimack, D. K. (1989) "The financial record-keeping practices of U.S. farm operators and their relationship to selected operator characteristics." Paper presented at the Annual Meetings of the American Agricultural Economics Association, Knoxville, Tennessee, August.

Yu, J., and H. Cooper (1983) "A quantitative review of research design effects on response rates to questionnaires." Journal of Marketing Research 20:36-44.

TIME RELATED COVERAGE ERRORS AND THE DATA ADJUSTMENT FACTOR (DAF)

Jeffrey T. Bailey, USDA/NASS
Research Division/3251 Old Lee Hwy., Fairfax, VA 22030

ABSTRACT

The National Agricultural Statistics Service (NASS) conducts quarterly surveys to estimate crop acreage, grain stocks and hog inventories. Sample replicates from the stratified sample design are surveyed on a rotating basis to allow for quarter to quarter overlap while bringing other operations into the survey. With this design, farming operations may be enumerated from one to four quarters in a particular year's survey cycle.

Operations are sometimes reported as "out-of-business" in one of the quarterly surveys when they were in business during a previous quarter. While this is not a problem if the questionnaires are correctly coded, a review of survey data reveals a significant number of coding errors in one quarter or the other. This between-quarter discrepancy in an operation's business status can change the coverage of the population (particularly if the change is due to incorrect coding) and have a major impact on the resulting indications.

This study looked at the effect of the coverage change on the indications and the reasons for questionnaires being coded as "out-of-business". From this research we hope to determine: 1) the extent to which those "out-of-business" changes represent data collection errors rather than real operation changes, 2) how to reduce the number of operations incorrectly being coded "out-of-business" and, 3) whether the data are increasing for operations remaining in business to offset operations legitimately going "out-of-business".

SUMMARY

The National Agricultural Statistics Service (NASS) conducts quarterly surveys to estimate crop acreage, grain stocks, and hog inventories. The replicated, stratified sample design results in sampled operations being surveyed in a rotating fashion, allowing for some quarter to quarter overlap while reducing respondent burden. A new sample begins in June with quarterly surveys in the following months of September, December, and March.

A dilemma arises as the year's survey cycle progresses beyond the June base survey, because the percentage of "out-of-business" operations increases. This creates a situation where the indications from the survey decrease and the population coverage may become incomplete. Observations show that approximately 4 to 6 percent of operations change from "in business" one quarter to "out of business" the next. Reviewing the questionnaires indicates that a substantial number of these were inaccurately coded or lacked complete information.

The Data Adjustment Factor (DAF) adjusts the data for duplication and eliminates data that should not be summarized. When an operation is "out-of-business" the DAF is zero. Calculations of the average DAF show that it continually decreases the further you get from June. The DAF reduced the December expansions relative to June by about 2 percent in 1991 and 1 percent in 1992. This drop from June is substantial, but how much of it reflects a legitimate change in the target population? What led to the reduction of the DAF impact in 1992 and how can we further reduce its effects?

The DAF should continue to be monitored and efforts be made to reduce its artificial impact upon the survey indications. Some suggestions to reduce the DAF decline are more training, changes in coding old replications, and the use of historic data to confirm "out-of-business" operations. These suggestions will likely not completely eliminate the DAF problem and more ideas should be developed and studied to lessen and monitor the DAF impact.

INTRODUCTION

The National Agricultural Statistics Service (NASS) conducts many surveys to estimate inventory and production of various agricultural commodities. As a part of its Agricultural Survey Program, NASS conducts quarterly surveys to estimate crop acreage, grain stocks and hog inventories. Analysis of December 1991 Agricultural Survey data showed that the December crop indications for planted acres were always lower than the June indications. Within a growing season the reported planted acreage of a crop should not change, unless intentions were reported in June and the crop was never actually planted. It was discovered that many operations which reported crops in June were now "out-of-business" in December.

Reviewing the data of these "out-of-business" operations focused attention on the Data Adjustment Factor (DAF). The DAF is used to adjust for duplication and to eliminate any data reported on an "out-of-business" operation. The value of the DAF is always inclusively between zero and one. The average DAF was calculated for successive quarterly surveys and found to decline as time passed. Several reasons can account for this and many ideas have been expressed.

This paper will begin with a description of the multiple frame surveys at NASS and how coverage errors can occur as time passes. Then the analysis of the DAF will be presented.

NASS MULTIPLE FRAME SURVEYS

NASS conducts many surveys and for each it is necessary to define the sampling population or frame of units to sample. For most NASS surveys the target population is all operations with the agricultural commodities of interest. NASS maintains a list frame of names thought to be farm operators in each state for its sampling. Considerable time and resources are spent in the state offices updating and maintaining these lists. In addition to the samples drawn from these lists, samples are drawn from an area frame of all land in the U. S. from which estimates are generated to measure list incompleteness. Together the two frames form a multiple frame survey design which NASS uses in many of its surveys.

This study focuses on NASS's quarterly multiple frame Agricultural Surveys. The list sample is selected in the spring with the surveys conducted during June, September, December and March. During the base survey in June a complete area sample is enumerated. For this survey, every operation in the U. S. has a chance to be sampled either from the list and area frame or the area frame alone. Names found in the area frame during June that are not on the list frame (NOL) will be used in subsequent quarters to represent those operations which had no chance of list frame selection.

The list sample consists of several replications which are selected each spring for use during the course of the survey year. These replications are rotated in and out from survey to survey to provide quarter to quarter comparability and to relieve respondent burden. With the rotation scheme used, farming operations may be enumerated from one to four quarters in a particular year's survey cycle.

TIME RELATED COVERAGE ERRORS

The samples for the Agricultural Survey are selected in the spring of each year. Before some samples are surveyed they will go "out-of-business". If an "out-of-business" operation is taken over by a new operation, this new operation must have a chance of selection. Any new operations taking over an "out-of-business" operation before June 1, will have a chance of inclusion in the area frame sample during the June Agricultural Survey. New operations starting up after June 1 can only be accounted for by substitution procedures, since there is no complete area frame survey done after June.

These substitution procedures provide a means to give everyone a chance of being selected to assure population coverage. Substitutions should be made when sampled units are "out-of-business" and the new operator was not farming on June 1, but there is concern that the procedures are not always executed properly and all needed substitution is not being done (Jones 1988). Furthermore, substitution only occurs when an operation is completely "out-of-business". If an operation sells off only part of its land to a new operator, that operation is not eligible for substitution and does not have a chance of selection (Dillard 1993). NASS is currently researching how effectively substitution procedures are being followed and the impact of the substitution process on survey indications.

For the follow-on quarterly surveys of September, December, and March, about 40% of the sample is from new replicates, with the remaining from old replicates that were surveyed in a previous quarter. For old replicate samples only those operations that were in business in the previous quarter will be surveyed in a following quarter.

Figure 1 shows the percentage of active samples from old replications that were coded "out-of-business". While over the course of time it is natural for some operations to go "out-of-business", the percentage coded as "out-of-business" is questionably high. It is doubtful that all operations so coded actually went "out-of-business" since the earlier quarter contact; some may be miscoded and others may have been refusals in a previous quarter.

This study looked at the errors of reporting and coding "business" status and their effect on coverage. While some operations legitimately go "out-of-business" between quarters, and these can be substituted for, a substantial number of changes from quarter to quarter are errors in coding. For example, an operation is

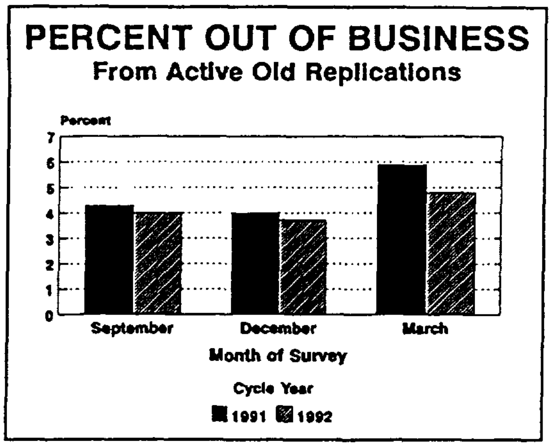


Figure 1

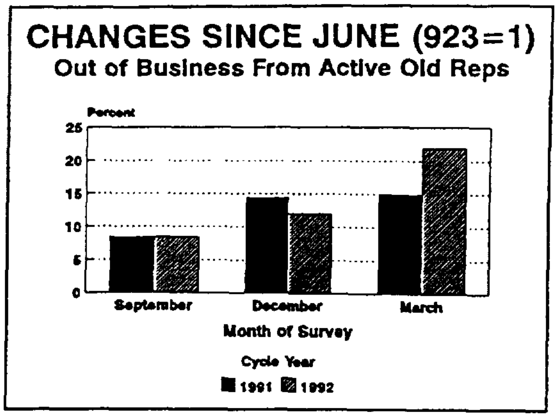


Figure 2

coded as "out-of-business" in a current quarter but "in business" for a previous quarter, when in fact it should have been recorded as "out-of-business" during the first quarter because the sample unit was a landlord. The converse can also happen when an operation is coded as "out-of-business" when it is really in business, since it continues to have potential for agricultural production.

In addition to being coded as "out-of-business", questionnaires are coded as to whether the sampled operation has changed since June 1. When an operation has gone "out-of-business" since June 1 item code box 923 on the face page of the questionnaire is coded a 1. Figure 2 shows the surprisingly low percentage of "out-of-business" operations from active old replicates that were coded as a change since June 1. Since all old replicates were reported in business during a previous quarter, we would expect nearly all current survey "out-of-business" reports to be changes since June 1. Therefore, if the current survey coding is correct, most operations were reported erroneously during the previous quarter. However, it is believed that code box 923 is frequently left uncoded. The coding of this box may be overlooked for old replications in part because it does not need to be coded for new replications.

Any operation that is reported as "out-of-business" is not surveyed again during that year's survey cycle. By NASS's definition, an "out-of-business" operation does not have any agricultural commodities and has no potential for agriculture during the rest of the year. Therefore, if correctly reported, it will have nothing to report in the following quarters and need not be surveyed. Each quarter more of these known zeros are accumulated, which creates problems when an operation is misreported as "out-of-business." State Statistical Offices (SSO) are instructed to review the known zero operations, but since not all are enumerated again some previous survey errors may go undetected. Any undetected misreporting of business status will cause a downward bias in the indications.

ANALYSIS OF THE DATA ADJUSTMENT FACTOR

In NASS's Agricultural Surveys, the Data Adjustment Factor (DAF) adjusts reported data for duplication and eliminates any positive data for operations that should not be summarized. Under normal situations the DAF is one, but it can have other values between zero and one. Common situations where the DAF is not one are: 1) an operation is duplicated in the same stratum (DAF=.5), 2) an operation is duplicated in a higher stratum (DAF=0), and 3) an operation is "out-of-business" (DAF=0). Table 1 shows the weighted (by the expansion factor for each design stratum) average of the DAF during the last two cycles of the Agricultural Surveys. The pattern of a decline is clear. One would expect to see some decline as operations go "out-of-business", but the amount of decline is of concern since it can have a large impact on survey results.

To determine the effect of the DAF on the expanded data, analysis was done comparing June to December expansions (Tables 2 & 3). The effects of the DAF, reported data, and the tract/farm weight factors were separated to assess the magnitude of each. This was done by calculating the normal June expansion, then using the information from those reporting in December to recalculate the June expansion. For example, the expanded data for an operation that was in business in June but not in December, would be positive in June

Table 1. **Average Data Adjustment Factor**

Cycle Year	Month of Survey			
	June	September	December	March
1991	.926	.899	.871	.849
1992	.946	.933	.907	.87

Table 2: **Data Adjustment Factor (DAF) Effect on the Corn Planted Acreage Expansion for Survey Years 1991 and 1992.**

Factor	June to December Comparable Reports for Factor					
	1991			1992		
	Ratio June to Dec.	Difference June - Dec. (000)	Difference as % of US	Ratio June to Dec.	Difference June - Dec.(000)	Difference as % of US
DAF	.95	-1,554	-2.0	.96	-1,108	-1.4
List Data	.99	-192	-0.2	1.00	-63	-0.1
Area Data and Weight	.99	-95	-0.1	1.07	464	0.6

Table 3: **Data Adjustment Factor (DAF) Effect on the Total Hog Inventory Expansion for Survey Years 1991 and 1992.**

Factor	June to December Comparable Reports for Factor					
	1991			1992		
	Ratio June to Dec.	Difference June - Dec.(000)	Difference as % of US	Ratio June to Dec.	Difference June - Dec.(000)	Difference as % of US
DAF	.95	-1,271	-2.3	.98	-615	-1.0
Data	.97	-685	-1.2	1.03	807	1.4
Area Weight	.99	-122	-0.2	.99	-99	-0.2

and zero for the recalculated June expansion with the December information. Comparable reports for a particular factor had to have usable factor information from both the June and December surveys.

Additionally, comparable reports for data and weight had to be in business both quarters. For the corn planted acreage expansion the area data and tract/farm weight factors can not be separated, because in June only tract data are reported while in December only farm data are reported. For total hogs, farm data are reported in both June and December, so comparisons between June and December of both data and weights can be made.

From Tables 2 & 3, we can see that in 1991 the DAF factor had a greater impact upon the difference in expansions between June and December than did the data or the weight. For example, the DAF factor resulted in a decrease in the U. S. expansion of 2 percent for corn planted acreage while the list/area data and weight factors decreased the expansion by only 0.2 and 0.1 percent, respectively. The situation for total hog inventory was similar, with the DAF decreasing the hog expansions by 2.3 percent. The size of the decrease due to the DAF factor is larger than the coefficient of variation for both estimates, illustrating the substantial effect the DAF has.

When we look at the 1992 analysis in Tables 2 & 3, we see that the effect of the DAF is about one half the size it was in 1991. This is encouraging, but the reason for the change in results is hard to determine. It is possible that training to make people aware of the DAF concerns has had a positive impact. One possible reason for the drop is the new list sampling unit/reporting unit association procedures, which half of the states used in 1992. These new procedures for associating reported data with sampled list names are called "operator dominant," as compared to the previous procedures which are referred to as "operation dominant." To see if this procedural change reduced the DAF impact, the effect of the DAF was compared between the two groups of states. Analysis showed there is only slight evidence that the DAF effect was smaller in the group with the new list dominant procedures.

To learn why operations were being coded as "out-of-business" we began to collect reasons. Observations made in Missouri during June 1992 were used to compile a preliminary list of these reasons. This list was used in Kansas during the December 1992 Agricultural Survey to code all questionnaires for which the reporting unit was coded "out-of-business" (i.e. item code 921=9). All old replications so coded were in business a previous quarter, while new replicates had not been surveyed. The reasons to be used in the coding were designed to differentiate between the situations expected between old and new replicate samples. The resulting list of reasons, while a starting point, turned out to be inadequate since too many reasons were grouped as "other."

To improve upon the reason coding, listings were sent to selected states after the December 1992 Agricultural Survey. State office personnel were to write out the reasons that operations changed their business status to "out-of-business". Table 4 is a compiled list of the reasons from four states. The most common reason was that incomplete information was obtained during the prior survey, because the respondent either refused or did not provide information about a partner involved in the operation.

Several of the reasons for operations being coded as "out-of-business" are related to the (small) size of operations and to whether they have agricultural potential. NASS defines as "out-of-business" an operation which has no potential for agricultural inventory or production during the remainder of the survey year. With this definition, no operation with potential for agricultural commodities should be coded as "out-of-business". While these operations may have nothing to report for any particular quarter they may have agricultural inventory or production during a subsequent quarter.

From the Table 4 list we can not tell directly whether the change in business status occurred after June 1 or was simply not picked up during a previous quarter. We can presume that some reasons, like 'landlord only', reflect situations which were not picked up in a previous quarter. Others, like 'sold farm', may or may not represent actual changes since June 1. If the change occurred after June 1 then the selected unit would be a candidate to be substituted for. If there is not an actual operation change, then there is a mistake in one quarter or the other. This may result from the respondent failing to answer correctly, some recording error, erroneous office coding, or one of many other possibilities.

Table 4: **Detail of Reasons for Old Replications Coded as "Out-of-Business"**

Number Times Occurred	Reason
37	Previously refusal and status not determined
15	Partner reported in higher strata
12	Partner reported in same strata
11	June with potential only.
8	Landlord only: incorrectly reported in previous quarter
7	Turned over to someone else
7	Sold farm
6	Name on label does not farm
5	Reported crops or livestock earlier, and reported none now

4	Minor crops or a few livestock only in previous survey
4	Turned over to son
4	Deceased
4	Retired
3	Land is now idle
3	Valid "out-of-business" (reason unknown)
3	Box 921 coded in error in current survey
3	Land is now rented, operated it previous quarter
2	CRP operator which should not be coded "out-of-business"
2	Miscoded multiple operations
2	Operator lied on previous report
2	Farm operated by someone else
2	Previously reported as 2 operations, actually only 1
2	Name correction on area frame, now OL
2	Partner strata boxes coded incorrectly
1	Chicken contractor only
1	Works on another farm only
1	Wrong name collected on June tract
1	Grain Co. only

DISCUSSION AND RECOMMENDATIONS

There are many causes for the DAF decline. Some of the decrease is valid and expected since operations will always be going "out-of-business", but some is due to survey error. The many causes increase the complexity of determining what needs to be done. The evidence suggests that the DAF decrease is large, meriting further analysis. Education and awareness can reduce errors. Procedural changes in coding to distinguish the difference between reporting errors and valid changes may provide better indications. Collecting more reasons for operations coded as "out-of-business" may give further insight, while measuring and adjusting for the DAF and the use of ratio estimates based on operations whose DAF did not change may need to continue.

There already have been efforts to educate people about the DAF. During the 1991 Midyear Survey Training, a session was conducted which provided DAF averages and comparisons between June and December expansions. This awareness may have made a difference since the decrease in the DAF in 1992 was about one half what it was in 1991.

Based on these results, I recommend continued, enhanced training with each state to examine their unique problems and further reduce the DAF dilemma. This education could be done during the advanced mid-year workshops. Statisticians in each state office could compile a list of reasons why some of their operations were coded "out-of-business". This list could then be the subject of small group discussions, probing for solutions.

I recommend the coding scheme for the "change since June box" be modified to improve the accuracy of its coding. Procedures that would require it be coded for all "out of business" operations would prevent it from being ignored. Once accurate information is obtained, ratios to a previous quarter could exclude illegitimate changes.

Another way to reduce the number of old replicate samples inappropriately being coded as "out-of-business" is by using historic data. When a respondent responds that they do not have the items of interest, we could then verify that they no longer have the items reported previously. This would be especially beneficial on CATI/CAPI.

I recommend we look more closely at the "out-of-business" operations and assess whether data compensation is being realized through the use of the current substitution procedures. This is the thrust of a separate research activity currently being addressed in NASS.

BIBLIOGRAPHY

Dillard, Dave. (1993) "Design Specifications to Standardize Survey Procedures." National Agricultural Statistics Service, United States Department of Agricultural, February 1993.

Jones, Ned. (1988) "QAS Substitution Analysis." Estimates Division, National Agricultural Statistics Service, United States Department of Agricultural, NASS Staff Report Number SMB-88-04.

UNBIASED ESTIMATION IN THE PRESENCE OF FRAME DUPLICATION

Orrin Musser, National Agricultural Statistics Service, USDA
Research Division, Room 305, 3251 Old Lee Hwy, Fairfax, VA 22030

INTRODUCTION

Survey organizations which conduct surveys on an ongoing basis devote much effort and expense to the maintenance of their sampling frame. Estimation of population parameters may suffer from two main types of frame deficiency: incomplete population coverage and duplication. In this paper we will focus on the problem of duplication, with emphasis on the computation of correct inclusion probabilities as a means to achieve unbiased estimation.

Duplication in the sampling frame is a serious problem which undermines the assumption of known inclusion probabilities for each population element. For a large sampling frame, while it may be too costly to determine all duplication in the frame, it may be reasonable to assume that for a given population element it may be possible to determine all duplicates in the frame. If so, then for many sampling designs, unbiased estimation is possible.

A sampling frame is a device which associates a collection or list of sampling units with a finite population of elements. It is helpful to formally describe the relationship between the sampling frame and the population. Suppose we have a population U = $\{E_1, E_2, \ldots, E_k \ldots, E_N\}$, a collection of N elements E_k and a sampling frame F = $\{F_1, F_2, \ldots, F_i \ \ldots, F_M\}$, a collection of M sampling units F_i. For each unit F_i and each element E_k, let the indicator variable δ_{ik} be defined:

$$\delta_{ik} = \begin{cases} 1 \ \textit{if} \ F_i \ \textit{represents} \ E_k \\ 0 \ \textit{otherwise.} \end{cases}$$

And let $M_k = \sum_i \delta_{ik}$ be the count of frame units which represent or "link to" population element k. We will call the collection or set of frame units linked to population element k, link-group k. Duplication exists in the frame when there are some population elements which are linked to more than one frame unit, that is $M_k > 1$ for some k. We assume that while M_k's are unknown (difficult and/or expensive to determine for a large entire population) M_k can be determined exactly for a particular unit k and thus for a sample. Assume a simple random sample of size m without replacement. We will denote this sample of frame units by s. For each sampled unit, we obtain a listing of all frame units that link to it. This set of frame units represents a single unique population element k. Since other members of this linkage group could have been sampled, it is possible to sample a population element k more than once. If we think of our sampling as sampling without replacement of population elements, we also obtain a sample s_p which contains $n(\leq m)$ distinct population elements. While each frame unit in our original sample s had equal probability of selection, each population element in our sample s_p did not have equal probability of selection, and thus estimators which assume we are sampling population elements with equal probability will be biased if there is duplication in the frame.

2. CORRECT INCLUSION PROBABILITIES

We know that the Horwitz-Thompson estimator:

$$\hat{Y}=\sum_{i=1}^{n} \frac{y_k}{\pi_k}$$

is an unbiased estimator of the population total, Y. Thus if we can compute π_k for each sampled element k, we can get unbiased estimates for population size and in general for any variable of interest, even in the presence of duplication. The inclusion probabilities, π_k, are straightforward to calculate if we know M_k for each sampled unit. If we are interested in the probability of selection for population unit k, or equivalently linkage group k, of size M_k, we use the fact that this is equivalent to 1 minus the probability of selecting a sample of size m from the frame such that no units of linkage group k were selected. This is just the ratio of the number of possible samples of size m chosen from the $(M - M_k)$ frame units which do not include any member of linkage group k over the number of possible samples:

$$\pi_k=1-\frac{\binom{M-M_k}{m}}{\binom{M}{m}}$$

Example: If we take a sample of m=5 frame unit from a frame of size M=100 in which there is duplication, what is the probability that our sample s_P of distinct population elements contains a particular population unit k for which $M_k = 1, 2$?

For $M_k = 1$(Population element k is represented only once on the frame):

$$\begin{aligned}\pi_k &= 1-\frac{\binom{99}{5}}{\binom{100}{5}}\\ &= 1-\frac{99!}{94!5!}\,\frac{95!5!}{100!}\\ &= 1-\frac{95}{100}\\ &= \frac{5}{100}=\frac{m}{M}.\end{aligned}$$

This is true in general, for each population unit for which there is no duplication, i.e. $M_k=1$, the probability of selection is just m/M, or f, to be denoted as π^*. (Recall that since we may sample a given population element more than once, n, the number of distinct population elements sampled, is a random variable and $\pi_k \neq n/N$.)

$M_k = 2$: (Population element k is represented twice in the frame)

$$\begin{aligned}\pi_k &= 1-\frac{\binom{98}{5}}{\binom{100}{5}}\\ &= 1-\frac{98!}{93!5!}\,\frac{95!5!}{100!}\\ &= 1-\frac{95}{100}\,\frac{94}{99}\\ &= .0979797\end{aligned}$$

Note that this probability is not double the selection probability for a population unit without duplication. It is interesting to look at this ratio of selection probabilities in general. Looking at the general formula for the selection probability when $M_k=2$, we may express π_k approximately as a function of the sampling fraction, f = m/M:

$$\begin{aligned}\pi_k &= 1-\frac{\binom{M-M_k}{m}}{\binom{M}{m}}\\ &= 1-\frac{(M-M_k)!}{(M-M_k-m)!m!}\,\frac{(M-m)!m!}{M!}\\ &= 1-\frac{M-m}{M}\,\frac{M-m-1}{M-1}\\ &\approx 1-(1-f)^2 = 2f(1-\frac{f}{2}).\end{aligned}$$

Thus the ratio $r_k = \pi_k/\pi^*$ where $M_k = 2$, and π^* is the probability of selection for any population element for which there is no duplication, may be expressed:

$$r_k = \frac{\pi_k}{\pi^*} \approx \frac{2f(1-\frac{f}{2})}{f} = 2(1-\frac{f}{2}).$$

Thus, as the sampling fraction approaches 0, r_k approaches two. As f gets large and approaches 1, r_k approaches 1. Thus as the likelihood of selection gets smaller, the bias due to incorrect assumptions of equal probabilities is increased. This ratio r_k is interesting because it expresses the degree to which the data y_k is "over expanded" due to the assumption of known equal selection probability. In our example where f = .05, the "pi estimator" would over expand y_k by a factor of .09797/.05 = 1.96. If N were 10(f= 1/2), then r_k will be approximately 1.5 and thus estimates which ignore duplication will over-expand data for elements with one duplicate by a factor of 1.5. If N were 10,000 then r_k would be essentially 2.

Since we assume that we may determine M_k for any population element k, then clearly we may compute π_k for each sampled element and thus use the Horwitz-Thompson estimator to obtain unbiased estimates for population totals and means. If we define $y_k=1$ for each population element k, then we could obtain an unbiased estimate of N, the true population size. This is just the sum of the reciprocals of the inclusion probabilities for the n distinct population elements in s_P. This estimator is unbiased for N:

$$\begin{aligned}\hat{N}&=\sum_{i=1}^{n}\frac{1}{\pi_k}\\ E[\hat{N}]&=E[\sum_{i=1}^{n}\frac{1}{\pi_k}]\\ &=\sum_{i=1}^{N}E[\frac{1}{\pi_k}I_k]\\ &\textit{where } I_k=1 \textit{ if } k\in s_p\\ &=\sum_{i=1}^{N}\frac{1}{\pi_k}E[I_k]\\ &=\sum_{i=1}^{N}\frac{1}{\pi_k}\pi_k=\sum_{i=1}^{N}1=N\ .\end{aligned}$$

The variance of the Horvitz-Thompson estimator of a population total Y is given by

$$V(\hat{Y}_\pi)=\sum\sum\nolimits_U(\pi_{kl}-\pi_k\pi_l)\frac{y_ky_l}{\pi_k\pi_l}$$

and an unbiased estimator of the variance is given by

$$\hat{V}(\hat{Y}_\pi)=\sum\sum\nolimits_s\frac{(\pi_{kl}-\pi_k\pi_l)}{\pi_{kl}}\frac{y_ky_l}{\pi_k\pi_l}$$

(Sarndal, Swensson, and Wretman, section 2.8). These general formulae are very useful for this situation where duplication results in unequal selection probabilities.

The second order inclusion probabilities needed for these formulae may be determined for the sample by the following formula which uses analogous reasoning to that for the first order inclusion probabilities.

$$\begin{aligned}\pi_{kl}&=P((k\in s)\cap(l\in s))\\ &=1-P((k\notin s)\cup(l\notin s))\\ &=1-\frac{\binom{M-M_k}{m}}{\binom{M}{m}}-\frac{\binom{M-M_l}{m}}{\binom{M}{m}}+\frac{\binom{M-(M_k+M_l)}{m}}{\binom{M}{m}}\end{aligned}$$

3. ALTERNATIVE STRATEGIES

One alternative "adjustment" for list duplication, and one that is currently used by NASS surveys, is the common survey practice of using a weight or data adjustment factor to account for the effect of duplication. If a population element k appears on the sampling frame M_k times, then when sampled the data is multiplied by $1/M_k$. Even if the same population element appears multiple times in the sample, every sampled unit reports. Cox(1993) describes this procedure as an adjustment of the weight "associated with sampled frame units to reflect the multiple selection opportunities for the desired population unit." This adjustment obtained by multiplying the sampling weight M/m, and the adjustment $1/M_k$ results in an overall weight of $M/(m*M_k)$. This new weight is not, in general, equal to the reciprocal of the probability of selection. As shown earlier the ratio r_k depends on the sampling fraction. Nonetheless, this procedure does result in unbiased estimation.

Suppose x is a data item with x_k being the data for each true population element k. Then for each frame unit l which is linked to element k, we define $y_{kl} = x_k/M_k$, $l = 1 .. M_k$. Thus we are letting each frame unit account for the proportion, $1/M_k$, of the data for population element k. Clearly the total of the y's is equal to the total of the x's:

$$\begin{aligned}\sum_{i=1}^{M}y_i&=\sum_{k=1}^{N}\sum_{l=1}^{M_k}y_{kl}\\ &=\sum_{k=1}^{N}\sum_{l=1}^{M_k}\frac{x_k}{M_k}\\ &=\sum_{k=1}^{N}M_k\frac{x_k}{M_k}\\ &=\sum_{k=1}^{N}x_k\end{aligned}$$

Thus a reasonable estimate for X would be

$$\hat{Y}=\sum_{i=1}^{m}\frac{M}{m}y_i\ .$$

Note again that this sum is over the entire sample of frame units. This is clearly unbiased for X, since this approach is equivalent to a simple random sample with the frame being the population. Thus $\hat{Y}$ is unbiased for Y = X.

This technique really obtains unbiased estimation by redefining the relationship of the frame to the population. If a population element appears M_k times on the frame, then each of those M_k records accounts only for the proportion $1/M_k$ of the data for population element k. This eliminates the duplication of the data.

Another approach used to obtain unbiased estimation in the presence of frame duplication is to define a "unique counting rule" which links each population element to a single frame unit. An example would be to link each population element k to the frame unit in M_k with the largest frame id, etc. In this case, population element k is sampled only if this particular frame unit is selected. Thus, if duplication were detected after data collection, there could be loss of data.

REFERENCES

Cox, B. (1993), "Weighting Class Adjustments for Nonresponse in Integrated Surveys: Framework for Hog Estimation," NASS Research Report SRB-93-03.

Lessler, J., Kalsbeek, W. (1992) *Nonsampling Error in Surveys*, New York: John Wiley.

Sarndal, C., Swensson, B., Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

ESTIMATING AGRICULTURAL LABOR TOTALS FROM AN INCOMPLETE LIST FRAME

D. Scot Rumburg, Charles R. Perry, Raj S. Chhikara*, and William C. Iwig, USDA/NASS
Charles R. Perry, Research Division, 3251 Old Lee Hwy., Room 305, Fairfax, VA 22030

KEY WORDS: Post-stratification, Multiple Frame, ratio estimate

ABSTRACT

The National Agricultural Statistics Service previously has conducted monthly labor surveys to estimate the number of total agricultural laborers. It employs a multiple-frame approach, using both a list and area frame. The list frame is highly efficient in sampling the target population of agricultural operations but does not have complete coverage of that population. The area frame covers all agricultural operations but is relatively inefficient in sampling those operations. An approach utilizing population count estimates from an initial area sample and post-stratified estimates from the monthly list sample has been investigated as a method for improving the precision of the survey estimate while reducing area frame respondent burden. Preliminary results indicate that survey to survey ratios of post-stratified list-only estimates can produce estimates which are comparable to current multiple frame estimates in both level and variance.

1. INTRODUCTION

A multiple frame approach, employing both a list and an area frame, has long been a cornerstone for many of the agricultural surveys which are conducted by the National Agricultural Statistics Service (NASS). Area frame responses often account for a majority of the total variance for multiple frame estimates but only a small part of the total indication. For this reason and others, it was recommended that a study be initiated to investigate alternatives to the current multiple-frame approach for administering surveys. A post-stratification approach whereby list respondents could be used to represent the entire target population, was recommended for consideration (Vogel, 1990a, 1990b and 1991). Kott (1990a and 1990b) elaborated on the proposal and outlined the two model-based estimators, their variance and potential bias. Perry, et al. (1993) provide an estimation method for the variance of a generalized post-stratification estimator based on its linear approximation using a Taylor Series expansion. Survey data from the California Agricultural Labor Survey series from July 1991 through June 1992 were used to investigate the alternative estimators.

2. METHODOLOGY

2.1 NASS Survey Methodology

NASS conducts numerous surveys with regard to agricultural commodities and related subjects. The majority of these surveys employ a multiple-frame (MF) methodology using both a list frame and an area frame. The list frame is stratified based on known data about agricultural operations with regard to the survey item(s) of interest. The list frame is not a complete listing of all agricultural operations. For the 1992 survey year beginning in June, the entire NASS list frame is estimated to contain 56% of all agricultural operations (often referred to simply as farms) and 81% of all land in farms. The area frame is stratified based on the agricultural intensity of a region. Unlike the list frame it has complete coverage of all agricultural operations in the U.S.

All reporting units (agricultural operations) in the June area survey (JAS:A) are classified as either overlap (OL) or as non-overlap (NOL) with the list frame. All operations found to be NOL are divided into several sampling pools to be used in follow-on surveys for the year. The list frame takes precedence over all OL operations when a MF estimate is calculated. A MF estimate is obtained by summing the list frame sample component estimate with the area frame's NOL sample component estimate. In most cases, the list frame provides about 75% of the total MF estimate while the NOL component adds only the remaining 25%. However, the NOL estimate is often a major contributor to the overall variance of the MF estimate, due to both the high variability of sampled units for many commodities and the sizable sample weights associated with small sampling fractions. The post-stratification approach investigated in this paper is an attempt to improve the reliability of the NOL component of MF estimates.

2.2 Post-Stratification Methodology

The proposed list-only estimator based on modeling of the NOL population represents a departure from the present NASS survey design and estimation methodology. Three factors motivate use of list only estimators: (1) the NOL sample units are highly burdened, (2) the current NOL estimates are often

[*] Raj S. Chhikara is Professor, Division of Computing and Mathematics, University of Houston - Clear Lake, Houston, Texas 77058

unreliable, and (3) the presence of NOL sample units increases the complexity of a survey.

Post-stratification for the Agricultural Labor Survey (ALS) was based on three classification variables: (1) The peak number of agricultural workers an operation expected to have over the course of a year (Peak), (2) the annual farm value of sales for agricultural goods (FVS), and (3) the type of farm operation (FType). These classification variables were selected based on their ability to describe distinct post-stratum populations and to correlate with the number of hired agricultural workers, which is the variable of interest. Basic strategy to obtain homogeneous post-strata populations involved selecting class boundary values for the two numerical classification variables (Peak and FVS), and creating combinations of the third categorical variable (FType). No more than twelve total post-strata could be created in order to maintain adequate sample counts for all post-strata across all surveys. Depending on cutoff values and FType groups selected, fewer post-strata could be constructed. An attempt was made to maintain a minimum of 20 respondents per post-stratum for all post-strata, though this was not always possible.

2.2.1 Post-Stratified Estimators

Post-stratification is often used as a variance reduction tool in a design unbiased survey. It can also compensate for the undercoverage of a target population by a particular selected sample. Both uses are employed for the approach explored in this paper. First it is hoped more homogeneous populations are produced with post-strata, resulting in variance reduction. Second, the list frame is used exclusively as the selected sample for follow-on surveys, resulting in undercoverage (actually non-coverage) of the NOL.

Once selected, the list sample is post-stratified to obtain post-stratum estimates. In the case of unweighted list responses, the estimator of the characteristic of interest Y is of the form:

$$\hat{Y}^{PS}_{(unex)} = \sum_{\substack{all\,k\\ post\text{-}strata}} (\hat{N}_{k(J)}) \cdot \left(\frac{1}{n_k} \sum_{i \in U_k} y_i \right) \qquad (Eq.1)$$

where

$\hat{N}_{k(J)}$ = k^{th} *post-stratum population size estimate from the June survey (JAS:A),*

n_k = k^{th} *post-stratum sample size, and*

U_k = *the set of all useable sample reporting units in the* k^{th} *post-stratum*

Similarly, a weighted estimator of Y is of the form:

$$\hat{Y}^{PS}_{(expd)} = \sum_{\substack{all\,k\\ post\text{-}strata}} (\hat{N}_{k(J)}) \cdot \left(\frac{\sum_{i \in U_k} w_i y_i}{\sum_{i \in U_k} w_i} \right) \qquad (Eq.2)$$

where

w_i = i^{th} sample *reporting unit weight, and* other *variables are defined as in Equation* 1

For each choice of $\hat{Y}^{PS}$ one can compute a ratio and ratio expansion based on a combined ratio.

3. RESULTS

3.1 Preliminary Research - Simulation Studies

Simulation studies provided a theoretical perspective on several aspects of the post-stratification methodology. Approximate variance estimates were derived and evaluated. These numerical evaluations showed that the performance of a post-stratified estimator is largely a function of the sample size used to estimate the post-stratum sizes, the sample size used to estimate the post-stratum means of the variable of interest, and the ratio of these two sample sizes. The relative efficiency of the post-stratified estimators all increased as the ratio of the two sample sizes increased. Given the sample size for the follow-on survey, the sample size for the base survey should be at least twice as large for gains in efficiency. Moreover, for post-stratification to be effective, the entire sample size in the follow-on survey should be at least 50 (preferably much larger) with the sample size in all post-strata at least 10 (preferably 20 or more).

3.2 Comparison of List and NOL Respondents

Table 1 shows that the NOL has a lower average estimate within nearly all post-strata for California, whether one compares weighted or unweighted responses. Particularly troubling are the large FVS post-strata with open-ended peak workers (5 or more) and specifically the fruit, nut and vegetable post-stratum. The few NOL respondents which fell into this category had many fewer hired workers than did their list counterparts. The high FVS post-stratum, with Peak 5+ and FType Crop & Misc produced a larger NOL average hired workers than did the list and was due to one large NOL respondent reporting 391 hired workers.

Table 1 also characterizes the difference between weighted and unweighted averages. Unweighted averages are consistently higher than weighted averages for both list and NOL respondents for nearly all post-strata. Since operations with larger numbers of hired workers are sampled at a higher rate, and because operations with larger numbers of workers tend to represent fewer number of farms, the sampling weights are negatively correlated with the number of hired workers, the variable of interest. This situation occurs even within post-strata. The negative correlation of weights and number of hired workers within post-strata suggests that the unweighted average will tend to overestimate the number of hired workers per farm for both frames.

TABLE 1. Counts and Mean Number of Hired Workers Within Post-Strata For the California July 1991 Agriculture Labor Survey

Survey Post-strata Definitions			Counts		Weighted Mean		Unweighted Mean	
FVS	FType	Peak	List	NOL	List	NOL	List	NOL
$1-50K	Crops&Misc	0-4	49	70	0.28	0.12	0.24	0.24
$1-50K	Crops&Misc	5+	1	1	0.00	0.00	0.00	0.00
$1-50K	Veg,Frt&Nut	0-4	70	79	0.21	0.11	0.17	0.13
$1-50K	Veg,Frt&Nut	5+	28	15	2.03	0.27	6.89	0.40
$1-50K	Dairy,Poultry, GrnHse&Nursry	0-4	4	0	1.36	.	0.75	.
$1-50K	Dairy,Poultry, GrnHse&Nursry	5+	0	0	.	.	.	.
$50K+	Crops&Misc	0-4	56	35	1.11	0.39	0.93	0.69
$50K+	Crops&Misc	5+	59	17	7.90	15.10	13.80	35.10
$50K+	Veg,Frt&Nut	0-4	57	15	0.70	0.48	0.77	0.67
$50K+	Veg,Frt&Nut	5+	249	38	15.30	5.29	38.80	16.30
$50K+	Dairy,Poultry, GrnHse&Nursry	0-4	30	0	1.15	.	1.30	.
$50K+	Dairy,Poultry, GrnHse&Nursry	5+	63	5	22.90	16.30	33.70	19.40

Cell counts and means for the weighted and unweighted response values by frame. Note that the NOL cell averages tend to be smaller than the list averages and that the weighted cell averages tend to be smaller than the unweighted averages.

3.3 Overall Performance of the Estimators

3.3.1 Post-Stratified Estimators

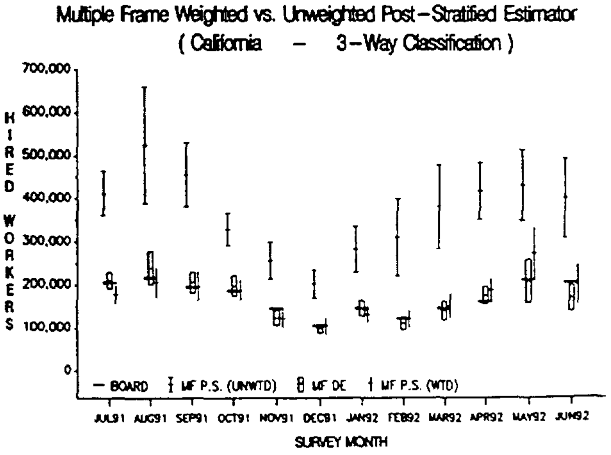
The combinations provided by selecting unweighted or weighted averages and an ability to select for list-only, NOL-only or both respondent types, produced six possible post-stratification estimators to study and evaluate. The NOL-only estimators were used only in conjunction with list-only estimators to provide comparative differences between the two frames on a state level basis. The MF post-stratified estimators were used to evaluate changes in variance due to list-only post-stratification.

Not surprisingly, it was found that the unweighted estimator consistently overestimated the actual labor force by a large margin (recall Table 1). The estimators using unweighted survey values produced the largest biases of all the estimators. Use of weighted survey values produced adequate, though somewhat more variable, estimates when compared to MF survey design direct expansion (MF DE) estimates. Since much of the post-stratification information is included in the list survey design (FVS and FType) and because the bulk of the ALS estimate comes from the list, it is not surprising the weighted MF post-stratified and the MF DE estimates are comparable.

Figure 1 depicts the level of bias produced by using a strictly unweighted post-stratified estimator and compares survey estimates across the 1991 ALS series year. For this and all succeeding graphs of this type, the vertical length of each estimate represents one standard error from the survey estimate in either direction. In some extreme cases the length in one or both directions has been truncated.

Also not surprising, given the post-stratum mean differences as shown in Table 1, it was found that the list-only estimator consistently overestimated the actual number of laborers while the NOL-only underestimated the actual labor force number. Figure 2 illustrates graphically the problems inherent in the weighted list and NOL-only post-stratified estimators, again comparing survey estimates to the Agricultural

Figure 1.



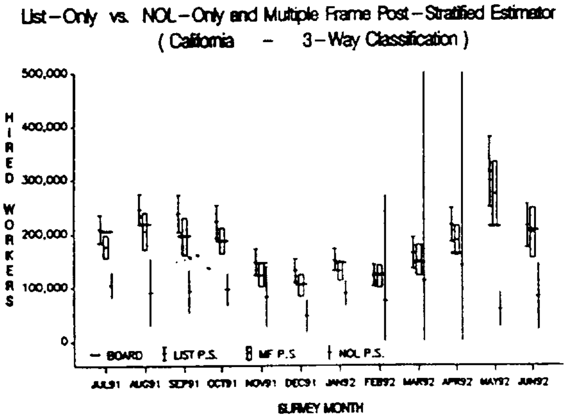
(Vertical Symbol Length Represents Two Standard Errors)

Comparison of Weighted versus Unweighted Post-Stratified Estimators.

Statistics Board as well as to the combined MF estimate across the 1991 ALS year.

Overall, list-only post-stratification CVs for California were mostly comparable with original MF DE CVs. This occurs for the most part because list-only post-stratified estimates generally are larger than the survey indication and have more variance introduced through the use of estimated June population counts. This leaves the overall percentage error of the total (CV) roughly equal to the MF DE CV. One must remember however, that the computed variance underestimates actual variance by as much as 10% resulting in a CV increase of approximately 5% since the $\sqrt{1.1} = 1.049$. For purposes of this report however, all CVs displayed will be the actual value computed with no compensation for bias. For California, the average CV for the weighted list-only post-stratification estimate for the survey year 1991 averaged 15.6%. This compares to an average MF DE CV for California of 14.2%.

Figure 2.



Comparison of the List-Only, NOL-Only and Multiple Frame Post-Stratified Estimators.

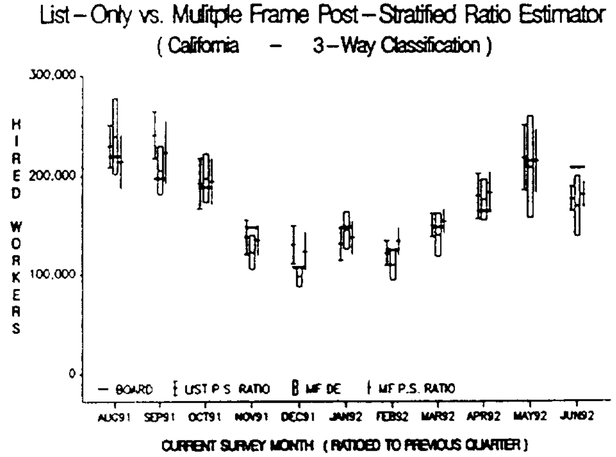
3.3.2 Ratio Estimators

Post-stratified combined ratio expansion estimates were calculated using MF and list-only data. In addition, a combined survey design ratio expansion estimate was computed using list-only data. Eleven monthly estimates were produced over the survey year for each estimator since a ratio estimate for July 1991 was not feasible. The ratio estimators were produced using only matched useable reports from both surveys.

3.3.2.1 Post-Stratified Ratio

Ratio expansion estimates were obtained using a combined post-stratified ratio estimator and the three-way post-strata classification scheme. Post-stratified survey total estimates using either list-only or MF respondents were constructed, and the results are shown in Figure 3 alongside the actual MF DE and the Board number for that month. The weighted list-only post-stratified survey total ratio tracks well with the Board estimate and, in fact, seven of the eleven ratio expansion estimates obtained for California were closer to the Board estimate than the MF DE indication. The average CV for California was 11.3% for the list-only ratio expansion estimate, which was less than the MF DE average CV of 14.6% over the same eleven surveys.

Figure 3.

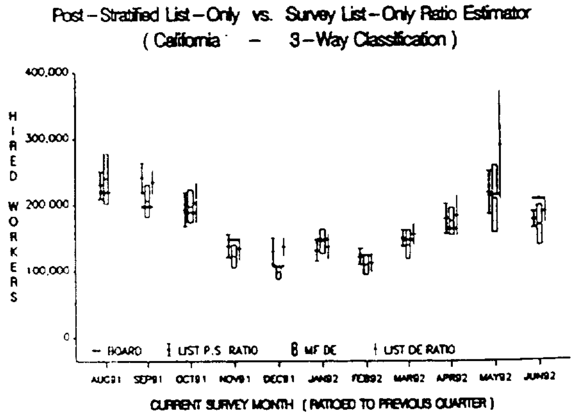


Comparison of the List-Only and Multiple Frame Post-Stratified Ratio Estimators.

3.3.2.2 Survey Design List-Only Ratio

Figure 4 compares the three-way post-stratified list-only combined ratio expansion with the survey design list-only combined ratio expansion. The post-stratified ratio estimator uses a weighted ratio, accounting for differences in farm numbers across post-strata. It is this difference which makes the post-stratified combined ratio estimator a more accurate estimator than the survey design combined

Figure 4.



Comparison of List-Only Post-Stratified and List-Only Survey Design Ratio Estimators.

ratio estimator. The two estimators have about the same precision.

3.4 Summary of Results

The combined three-way post-stratified list-only ratio estimator seemed to provide a viable estimate for the total number of hired workers indication. Though the post-stratification model is somewhat complex and would have to be optimized for each state or region, it does fulfill the objective of using a sample which ignores a subgroup, specifically the NOL.

4. CONCLUSIONS

For the Agricultural Labor Survey, there appear to be differences in mean values of list and NOL respondents within post-strata. Also, the sample design produces negative correlations between the sample weight and the response within post-strata. These two factors make the unweighted post-stratified estimator biased. Though bias is reduced in the case of the weighted post-stratified estimator, differences between weighted list and NOL respondents still exist within post-strata. Ratio expansion estimators, however, appear to avoid these problems and may have potential within the NASS framework.

The list-only combined ratio expansion estimator using three-way post-stratification appears to model the NOL adequately, while reducing variances on average. However, development of post-strata for individual states and regions would be a time consuming job and.would involve reworking of the current survey summary system. Additionally, an estimator that uses only list respondents will probably be biased and must be cautiously approached and monitored if any list-only estimator were to become operational.

One problem with the post-stratified estimators investigated here is the estimated farm counts from the JAS:A. These counts are estimated using the area weighted estimator and tend to be quite variable. The inaccuracies can be corrected to some degree by using the MF population estimate. Any variability in the counts translates to higher overall variances of the post-stratified estimates. The post-stratified ratio estimators reduce the magnitude of this problem, but more accurate population estimates would surely help these estimators also.

5. REFERENCES

Flores-Cervantes, Ismael; 1990a; **Simulation Results: A Preliminary Study of the "Strawman" Estimator and Derived Formulas**; November 21, 1991; NASS Internal Correspondence.

Flores-Cervantes, Ismael; 1990b; **Effect of an Incomplete List on the Design Based "Strawman" Estimator**; November 26, 1991; NASS Internal Correspondence.

Flores-Cervantes, Ismael; 1990c; **Estimating the Number of Farms in a Post-Stratum using the June Agricultural Survey**; December 2, 1991; NASS Internal Correspondence.

Kott, Phil; 1990a; **Some Mathematical Comments on Modified "Strawman" Estimators**; NASS Internal Discussion Paper.

Kott, Phil; 1990b; **Comparing "Strawman" and "Weighted Tract" Estimators in June**; NASS Internal Discussion Paper.

Nealon, John Patrick; **Review of the Multiple and Area Frame Estimators**; SF&SRB Staff Report No. 80; Wash., D.C. March 1984.

Perry, Charles R., Chhikara, Raj S., Deng, Lih-Yuan, Iwig, William C., and Rumburg, D. Scot; 1993; **Generalized Post-Stratification Estimators in the Agricultural Labor Survey**; SRB Research Report Number SRB-93-04, Washington, D.C., July 1993.

Rumburg, D. Scot, Perry, Charles R., Chhikara, Raj S., and Iwig, William C.; 1993; **Analysis of a Generalized Post-Stratification Approach for the Agricultural Labor Survey**; SRB Research Report Number SRB-93-05; Washington, D.C., July 1993.

NASS Program Planning Committee, 1992; **Minutes of the Program Planning Committee of the National Agricultural Statistics Service**; July 14-16, 1992.

Turner, Cheryl L.; **Evaluation of Estimation Options for the Monthly Farm Labor Survey**; SRB Research Report Number SRB-91-12; Washington., D.C., December 1991.

Vogel, Frederic A; 1990a; **"Strawman" Proposal for Multiple Frame Sampling**; October 3, 1990; NASS Internal Memo.

Vogel, Frederic A; 1990b; **"Strawman" Proposal**; November 20, 1990; NASS Internal Memo.

Vogel, Frederic A; 1991; **A New Approach to Multiple Frame Sampling and Estimation**; January 16, 1991; NASS Internal Discussion Paper.

AN ANALYSIS OF THE SAMPLING FRAME FOR THE CHEMICAL USE AND FARM FINANCE SURVEY

Susan Cowles and Susan Hicks, USDA/NASS
Susan Cowles, 200 N. High, Room 608, Columbus, OH 43215

I. Abstract

The National Agricultural Statistics Service (NASS) has begun a pilot study of the Chemical Use and Farm Finance Survey (CUFFS). The CUFFS combines parts of Form H of the Objective Yield Survey and the Cost of Production Survey (COPS) versions of the Farm Cost and Returns Survey (FCRS). Most NASS surveys utilize a multiple frame design -- a combination of list and area frames. The purpose of this analysis is to evaluate whether the multiple frame design is necessary for CUFFS. If not, then the sample will be selected from the list frame only.

II. Overview of the Chemical Use and Farm Finance Survey (CUFFS)

The CUFFS design consists of three phases. In the first phase, operations are contacted to determine if they have the commodity of interest. In the second phase, pesticide and fertilizer use information is collected in the Fall from operations that reported having the commodity. Finally, those same operations are recontacted the following Spring to obtain economic data.

The CUFFS was developed by NASS in an effort to:

- reduce respondent burden,
- improve data quality, and
- improve response rates.

The data that CUFFS would collect is currently collected through two other surveys: the Objective Yield Cropping Practices Survey (Form H) and the Farm Costs and Returns Survey's Cost of Production Survey (FCRS-COPS). When, or if CUFFS becomes operational, it would totally replace Form H and the COPS questionnaire would be shortened for crops targetted by CUFFS. The shorter interviews for Objective Yield and FCRS-COPS will reduce respondent burden for these two surveys.

Data quality is expected to improve for the COPS version of CUFFS as a result of collecting information closely following harvest. Currently, the COPS versions of the FCRS collects data six months after harvest. A better response rate is expected for the economic data due to its association with chemical use data. Farmers are more willing to provide data on chemical use due to widespread public concern for farming's effect on water quality and the environment.

III. Why Select a List Only Sample for CUFFS?

The list frame consists of known farm operators while the area frame consists of all land segments. The area frame is a complete frame and thus is used to measure undercoverage in the list frame. Farm operators found in the area frame that are not represented on the list comprise the NonOverlap Sample or NOL. There are three major advantages to a list only sample:

1) reduction in respondent burden for the NOL,

2) cost savings, and

3) reduction in variances.

Respondent burden is a major advantage of a list only sample. The NOL domain is relatively small due to small area frame sample sizes and more complete list frames. However, the relatively small population of NOL operators must be spread across all surveys, with the result that some NOL operators must be interviewed for multiple surveys.

The cost savings due to a list only sample are small compared to total survey costs. However, for less common commodities the NOL produces few if any positive operations. Thus, the cost per positive record is high. If the NOL domain is included, this cost could be reduced by screening operations by telephone for the commodity of interest prior to interview. See table 1.

Table 1

	CV%			June Planted Acres (in Mil)		
Commodity	NOL	LO	MF	NOL	List	MF
Corn	3.1	.6	.7	12.1	68.9	81.0
Soybeans	3.4	.9	.9	9.6	52.0	61.6
Spring Wheat	2.4	1.7	1.9	2.4	16.6	19.0

LO=List Only MF=Multiple Frame

At the U.S. level, the NOL contributes about 15% to total planted acres for major commodities, but contributes about 40% to the total variance.

For most commodities, the CV for a list only sample would be smaller than the multiple frame CV. However, the decrease in variance comes at a cost and that cost is bias. A list only sample introduces an inherent bias into the estimate by excluding some members of the population from the sample universe. However, if farm operators in the NOL domain are similar to farm operators on the list frame, then the bias may be minimal.

IV. Analysis Study to Compare List vs NOL Estimates

The goal of the research was to compare chemical use data from the list and NOL domains to see if there were significant differences. Ideally we would like to compare list to NOL estimates from the CUFFS questionnaire. However, the CUFFS pilot survey, which was conducted in Minnesota, used the proposed list only sample design, thus the NOL component was not available. To obtain a proxy for CUFFS chemical use data, we obtained Minnesota Form H data for corn, soybeans and spring wheat. Form H data is area frame only. The data was divided into overlap (OL) and nonoverlap (NOL) domains to allow comparisons between the two domains.

The OL and NOL domains were determined by classifying operations as OL or NOL to FCRS for 1991. The OL to FCRS group was further divided into groups determined by whether they were in a strata being sampled for CUFFS. If an operation was OL to FCRS and in a CUFFS strata, it was OL to CUFFS. All others were considered NOL to CUFFS.

The Form H summary system was used to obtain the mean rate of application per treatment and mean percent of acres treated for each active ingredient by domain. We then compared the estimates obtained between OL and NOL domains for the twelve most common commodity/chemical combinations.

V. Methodology

Percent acres treated is estimated as:

$$\hat{p}_d = \frac{n_d}{u_d}$$

where:

d = *OL or NOL domain*

n_d = *number of positive responses in domain d*

u_d = *number of usable responses in domain d*

The variances were calculated using the usual formulas for the variance of a proportion when the data are obtained by a simple random sample. In fact, the sample design was more complicated than a SRS, but the effect of this approximation should be a slight overestimate of the variance, which we were willing to accept.

Mean rate of application is estimated as:

$$\hat{R}_d = \frac{\bar{z}_d}{\bar{y}_d}$$

where:

$\bar{z}_d$ = *average rate of applicationfor each commodity/ chemical combination in domain d*

$\bar{y}_d$ = *average number of treatmentsfor each commodity/chemical combination in domain d*

For mean rate of application per treatment, we calculated bootstrap-t confidence intervals instead of the usual t-test because of concerns about normality of the statistic being tested. Using the bootstrap methodology we constructed histograms of the distribution of the statistic mean rate of application per treatment. The histograms suggested severe departures from normality for some statistics. See Rao and Wu (1988).

We selected 10,000 bootstrap samples from the combined sample -- OL and NOL domains combined. For each bootstrap sample we calculated the usual t-statistic for a difference. Then based on the distribution of the t-statistics we estimated the 5th and 95th percentiles of the t-distribution for each commodity/chemical combination. The confidence interval is defined as:

$$\{\ \hat{D} - t_{.95}\hat{\sigma}(\hat{D}), \hat{D} - t_{.05}\hat{\sigma}(\hat{D})\ \}$$

where:

$\hat{D} = \hat{R}_{OL} - \hat{R}_{NOL}$ -- *from full sample*

$\hat{\sigma}(\hat{D})$ = *standard error of difference*

$t_{.05}, t_{.95}$ = *percentiles of the bootstrap t distribution*

VI. Results for Mean Rates of Application per Treatment

Table 2 shows the mean rates of application per treatment, the normal t-test for the difference and the bootstrap-t confidence interval.

The normal t-tests are included for comparison purposes. The bootstrap confidence interval reflects the skewness in the distributions of the differences while the

Table 2

Commodity	Active Ingredient	OL Rate per Treatment	CV(R) (%)	NOL Rate per Treatment	CV(R) (%)	Normal test t(diff)	Bootstrap CI LL	Bootstrap CI UL
Corn	Nitrogen	67.25	2.50	63.07	5.09	1.151	-2.05	9.90
	Dicamba	0.32	2.79	0.25	5.21	4.598*	0.05	0.10*
	Atrazine	0.80	4.35	0.82	12.23	-0.215	-0.23	0.14
	Alachlor	2.24	3.70	2.63	12.18	-1.174	-0.91	0.20
	Metolachlor	2.14	3.18	2.31	5.18	-1.243	-0.38	0.07
Soybeans	Trifluralin	0.77	3.15	0.81	5.09	-0.955	-0.13	0.03
	Imazethapyr	0.05	2.06	0.06	2.30	-1.820*	-0.01	0.00
	Alachlor	2.60	3.26	2.54	6.89	0.310	-0.24	0.42
	Bentazon	0.69	5.44	0.76	8.10	-1.071	-0.19	0.06
Spring Wheat	MCPA	0.29	4.91	0.30	7.76	-0.403	-0.07	0.03
	2,4-D	0.26	16.78	0.31	14.45	-0.761	-0.15	0.06
	Bromoxynil	0.24	5.09	0.19	18.97	1.148	-0.01	0.21

normal t-test relies on the assumed bell-shaped distribution. Therefore, the bootstrap-t confidence interval more accurately reflects the true differences between the OL and NOL groups.

Using the normal t-test, two significant differences would have been found: Dicamba used on corn and Imazethapyr applied to soybeans. Their respective t-values are 4.598 and -1.820 which, in absolute value, are greater than the 1.645 critical value for a 90% confidence test. The bootstrap intervals show only one clear difference between the OL and NOL mean rates of application per treatment of Dicamba on corn. However, we could expect to find one or two significant differences, by chance, even if no true difference exists based on a 90% confidence interval. We conclude that the data do not suggest a difference in mean rate of application between the two domains.

VII. Results for Percent Acres Treated

Table 3 shows the results for a difference between the OL and NOL domains for percent acres treated.

Of the twelve commodity/chemical combinations tested, four showed a significant difference. Also, the difference for spring wheat treated with MCPA had a t-statistic of 1.639 which is quite near the critical value of 1.645. The absolute difference for MCPA was 16.1 percentage points. If this difference is considered significant, all three of the spring wheat/chemical combinations show significant differences. Alachlor applied to corn and Trifluralin applied to soybeans also showed significant differences.

As with rate of application per treatment, based on a 90% confidence interval one or two significant differences could be expected, by chance, when no true difference exists. However, because at least four significant differences were found, we conclude there appear to be differences in percent of acres treated between the OL and NOL domains.

Table 3

Commodity	Active Ingredient	Percent Acres Treated OL	Percent Acres Treated NOL	t
Corn	Nitrogen	97.0	98.3	-0.905
	Dicamba	30.8	26.7	0.898
	Atrazine	32.3	29.3	0.644
	Alachlor	25.8	19.0	1.654*
	Metolach..	25.0	26.7	-0.382
Soybeans	Triflur..	46.6	37.0	1.959*
	Imazeth..	55.5	51.3	0.848
	Alachlor	10.0	12.3	-0.735
	Bentazon	12.4	10.4	0.647
Spring Wheat	MCPA	67.6	51.5	1.639
	2,4-D	27.6	51.5	-2.455*
	Bromoxynil	37.1	21.2	1.866*

IV. Conclusions

We looked at rate per treatment and percent acres treated for twelve commodity/chemical combinations. For rate of application per treatment, the data do not show a consistent statistical difference. However, several differences were found between OL and NOL percent acres treated. While the data suggest some differences exist, we are reluctant to draw conclusions for the nation as a whole based on results for one state, for the following reasons:

- Cropping practices vary by state.
- Commodities vary by state.
- Applications of chemicals vary by commodity.

The next phase of the research will examine 1992 Form H data from Minnesota and Louisiana to determine whether these results are consistent over time and across states. We recommend delaying the decision about whether or not to proceed with a list only sample for CUFFS until that research is complete.

References

Cowles, Susan and Hicks, Susan (forthcoming), "An Analysis of the Sampling Frame for the Chemical Use and Farm Finance Survey," NASS Staff Report.

Rao, J.N.K. and Wu, C.F.J., (1988) "Resampling Inference with Complex Survey Data," JASA, 83, 231-241.

U.S. Department of Agriculture (1983): Scope and Methods of Statistical Reporting Service, Publication No.1308, Washington, D.C.

AGRICULTURE DATA SYSTEMS - A U.S./CANADA COMPARISON

Robin O. Roark
USDA/NASS, International Programs Office
So. Agriculture Bldg Room 4132; Washington, D.C. 20250-2000

Key Words: NASS and AgDiv

The Canada/U.S. Free Trade Agreement has opened the border for more agricultural trade between Canada and the United States. It will also increase the need for agricultural data and comparison of statistics from each country. The Agriculture Division (AgDiv) of Statistics Canada (STC) provides a wide array of agriculture statistics for Canada just as the National Agricultural Statistics Service (NASS) provides for the United States. However, procedures for sampling, data collection, analysis, and compiling data can be quite different. Even the structure of the agriculture industry, the structure of the two governments and of the two agencies plays a role in how data are collected, summarized, and published.

NASS, the statistical agency for the U.S. Department of Agriculture (USDA), is responsible for agriculture production and inventory statistics and some of the economic statistics. Economic Research Service, another economic agency within USDA, is responsible for compiling the statistical data into farm income and economic projections. World Agriculture Outlook Board, also one of the USDA economic agencies, uses the U.S. agriculture statistics from NASS, along with data from other countries, to estimate the world agriculture supply and demand.

The Agriculture Division of Statistics Canada is responsible for agriculture production, inventory and economic statistics and also compiles data for farm income and economic projections. Some economic data analysis work for outlook projections is done in conjunction with Agriculture Canada. Agriculture Canada is the agriculture policy ministry (department) of the Canadian Federal Government.

The primary difference in the structure of NASS and AgDiv is that AgDiv is part of a centralized statistical system known as Statistics Canada. Within STC there are several program divisions, including AgDiv, that have the responsibility of preparing statistical data related to their division. Other divisions have specific supporting functions to all STC program divisions, such as research, survey design, computer programming, dissemination, etc. The de-centralized U.S. statistical system has several statistical agencies. Each statistical agency is responsible for a particular area of data but must also support their own research, survey design, programming, dissemination, etc. NASS is the statistical agency for agriculture within USDA.

Both organizations have field offices to facilitate data collection, but the functions of these offices are different. For NASS, State Statistical Offices (SSO) are responsible for list maintenance, data collection, data entry, summarization, analysis and administrative work. The SSO's are a significant part of the NASS structure and have a great deal of input into the analysis and estimation of the data. They receive guidance and support from the main office in Washington, D.C. and concentrate primarily on agriculture related statistics. The State offices also have the freedom to be involved in state funded projects that are not part of the National program.

For AgDiv, data collection is done at the Regional Offices (RO). The ROs are part of STC and are primarily data collection and report dissemination centers. The ROs are not directly tied to AgDiv and they collect all types of statistical data. They generally do not get involved in the analysis, summarization or estimation of the data. However, the AgDiv does have agreements with each of the Provincial Governments. These agreements state that the Provincial Statisticians will review the data and estimates prior to the publication of the data. These statisticians are allowed some input into the level of the published estimates.

Both organizations use sales of agricultural products, without regard for acreage, as the defining factor for establishing an operation as a farm. However, the cut-off for the sales value is different. The definition of a farm for the U.S. is any operation that has $1,000.00 in agriculture sales or expected sales. For Canada the definition of a farm is any operation that produces agricultural product(s) for sale.

Census of Agriculture

The Agriculture Census for Canada is conducted every five years by AgDiv. The enumeration coincides with the Census of Population. Therefore, a question is asked on the population census questionnaire about farming interests. If the response is positive, then an Agriculture Census questionnaire is filled out by the respondent. The questionnaires are delivered by a STC enumerator and are mailed back. The enumerators do follow-up of the non-response. Analysis of data gives an expected under-coverage of about 1.5% to 3%, depending on the estimate.

Agriculture Census data are reviewed by AgDiv analysts at the provincial, and sub-provincial level, concentrating on the top contributors with most of the manual editing done on a macro-level. Only if severe problems are detected, or in the review of extremely large operators, are individual records reviewed by analysts. After Census data have been reviewed and published, AgDiv completes a 5 year historic review of all acreage, production, inventory, and economic estimates. Generally, AgDiv estimates, for the year of the census, are revised to match Ag. Census estimates, with minor adjustments due to differences in reference dates. The impact of the under coverage or duplication is assumed to be negligible.

The U.S. Census of Agriculture is conducted every 5 years by the Agriculture Division of the Census Bureau, part of the Department of Commerce. The U.S. Agriculture Census is a stand alone collection, not tied to the U.S. Population Census. The U.S. Agriculture Census is a mail out survey with telephone follow-up of the non-response. The Agriculture Division of the Census Bureau maintains a list of farms. The list is updated from responses to their surveys and from outside sources, such as NASS, income tax records, etc. Under-coverage from the Agriculture Census is about 13% of the farms. Under coverage from the Census is primarily with the smaller farms. The Agriculture Census uses the NASS area frame to estimate potential under coverage. However, Census published numbers are totals from the survey and are not adjusted for the under coverage. Duplication also causes some problems, especially with producer contract arrangements. Census data are reviewed by NASS statisticians at the county, State, and National level. Like AgDiv, data are reviewed on a macro-level with review of individual records limited to severe problems and extremely large operators. After Census data have been reviewed and published, NASS completes a 5 year historic review of all acreage, production, and inventory estimates. However, estimates from the Agriculture Census are not used as official NASS estimates. NASS makes adjustments to Census data to account for duplication, under coverage and differences in reference dates.

The linkage between the Canadian Census of Agriculture and the Population Census, allows for more Census estimates of the social characteristics of agriculture. However, the Agriculture Division of the U.S. Bureau of Census conducts follow-on surveys to establish estimates for most of the same statistics.

Sampling Frames

List - The farm register (list frame) for AgDiv is based primarily on names received during the Ag. Census. The Ag. Census is used as the basis for the list frame for a 5-year period. The majority of updates are based on changes found during surveys conducted during the 5 year period. The list frame, for most probability surveys, is frozen between Census years. The samples for these surveys are selected shortly after the current Census. New names are not generally added to the frame, except names of operators that are new to agriculture and take over an existing operation are allowed to replace an existing name. Some surveys, such as the fruit and vegetable survey, do use producer organization lists to update new names in between Census occasions. These samples are re-drawn every year. Coverage at the time of the Census is estimated to be about 97% for most samples, but this percentage drops at the rate of about 1-2% per year after the Census, depending on the commodity being measured.

The list frame for NASS is continually updated and samples are redrawn every year. Since updates to control data are based on information received during surveys and on information from producer organizations and government program participation lists, etc., not all names and control data are updated each year. NASS does not receive names or control data from the U.S. Agriculture Census. The NASS list frame was built many years ago from outside organization lists. Coverage runs at about 55% for the number of all farms, but varies significantly by State. Coverage is concentrated on larger farms with coverage of farm land at about 80%.

Area - The AgDiv area frame is designed to produce a weighted segment indicator to be used in conjunction with list frame surveys. When screening is done, the primary data collected are name and address information, total acres, acres in the segment (a piece of land with identifiable boundaries used as a sampling unit in area-frame sampling), and some general information about the type of farm. Agricultural operations are then determined to be either overlap (included on the list frame) or non-overlap (not included on the list frame). The non-overlap operations are then included in subsequent multi-frame surveys. Virtually no data indicators are produced from the area frame alone.

The NASS area frame survey is designed to produce closed segment indicators, open segment indicators and weighted segment indicators. Indicators from the June Area Frame Survey are used both as independent indications and also used in conjunction with list surveys to produce multi-frame indications. Area screening includes collection of tract (the area of land located within a segment that is under a single operating arrangement) data for all agriculture operations found in the area frame sampled segments and entire farm data for all operations that are not known to be overlap with the list frame. Non-overlap operations are also included in Agriculture Survey program for the rest of the survey year. The NASS area frame plays a much more significant role in the estimation program for NASS since the list coverage is lower.

AgDiv has begun using telephone enumeration to identify area frame operators in some areas of the Prairie Provinces (Alberta, Saskatchewan, and Manitoba). Segments in this region are drawn to follow the range and township boundaries. Since only limited data are collected at the time of the screening, preliminary results have been very favorable. Due to the complexity of segment boundaries in the other regions, the segments are personally enumerated. All NASS segments are personally enumerated due to both the precise segment boundaries and the amount and type of data that must be collected. The design of the AgDiv area frame specifically excludes areas not considered to be involved in agriculture, such as rangeland and urban areas. The NASS design includes these areas but samples them at a proportionally lower rate.

Sample Design

Both organizations use probability sample designs on all major surveys. Crop estimates for AgDiv are based on a series of surveys called the Crop Panel surveys. The Crop Panel surveys begin with a Seeding Intentions and Grain Stocks Survey in early April. The panel surveys continue with the June Acreage Survey, July 31 Yield and Grain Stocks Survey, September 15 Yield Survey, November Acreage & Production Survey, and December 31 Production & Stocks Survey.

Crop estimates for NASS are based on a series of integrated surveys called the Agriculture Survey program and the monthly Agriculture Yield Surveys. The March 1 Agriculture Survey is used to estimate seeding intentions. The June 1 Agriculture Survey is used to establish the planted acres and preliminary harvested acres. The September 1 Agriculture Survey is the end-of-season indicator for production of small grains and the December 1 Agriculture Survey is the end-of-season indicator for production of other field crops and hay. The Agriculture Surveys are also used to collect quarterly on-farm grain storage data. The monthly Agriculture Yield surveys collect yield and production data on crops during the growing season. The crops included in the survey will vary from month to month depending on the growing season of each crop and the program for that crop. The first small grain yield survey is conducted in May and the first row crop/hay yield survey is in August.

The actual sample design is fairly similar for the two organizations with both designs stratified by acreage of cropland. The NASS sample is stratified by State on cropland and grain storage, with some strata for specialty crops such as tobacco. The AgDiv Crop Panel is stratified on cropland by sub-provincial regions.

The Agriculture Survey sample for NASS is also stratified to collect quarterly hog & pig inventory and farrowing data. Reference dates for hog inventory estimates are the same as the dates for the four Agriculture Surveys. The cattle & sheep estimates reference dates are January 1 for both cattle and sheep and July 1 for cattle only. Therefore, cattle and sheep data are collected via a separate survey with separate stratification and sampling. The AgDiv livestock surveys are done twice each year and include cattle, hogs, and sheep. The reference dates for the livestock surveys are January 1 and July 1.

There are also numerous probability and non-probability surveys conducted by both organizations to obtain statistics on commodities such as fruit, vegetables, poultry, prices paid and received by farmers, farm income and expenses, etc. Due to the structure of the farm programs and marketing boards, there are more administrative data available to the AgDiv than are available to NASS. The supply managed commodities, which are milk, eggs and poultry meat, have extensive administrative data available that are used by the AgDiv in lieu of survey data. Farm expense data for AgDiv are obtained through a sample of income tax records rather than an additional survey of farmers. Both organizations use administrative data whenever available to help relieve respondent burden and lower data collection costs.

Data Analysis

NASS questionnaires that are not collected using Computer Assisted Telephone Interview (CATI) procedures are put through a complete manual review prior to data entry. Discrepancies are reviewed and corrected. Within AgDiv, the manual editing of data prior to data entry is virtually non-existent. Data that are not collected with CATI are briefly reviewed by the data entry division for clarity. However, data are not edited as being "correct" or "incorrect".

Computer editing of data, after collection, is used in both organizations. NASS's computer editing is designed to review data and flag errors with only a small number of the errors corrected by the editing program. The remaining errors are then reviewed and corrected by statisticians. The computer editing program for AgDiv is designed to make corrections or perform imputations for most of the errors. Statistician review and correction of the remaining micro-level errors is not as prevalent.

Expansion (or raising) factor adjustment is used by both organizations in most probability surveys to account for missing data and refusals. Some surveys in both organizations use automated imputation procedures. Response rates for the production surveys are very similar between the two agencies.

Estimation

The estimation program in NASS requires that most data and estimates be reviewed, analyzed, and approved by the Agriculture Statistics Board (ASB). The ASB is made up of the ASB Chairperson, the Estimates Division Director, the respective commodity statistician(s) and their Branch Chief, and 1 or more statisticians from 1 or more SSOs. For selected estimates, the Secretary of Agriculture, or a representative from the Secretaries office is briefed about the estimates prior to the release of the data. Within AgDiv, generally only the statisticians (Federal and Provincial statisticians) and their immediate supervisor review the data and estimates prior to the release.

The concept of livestock inventory estimates are nearly identical. The weight groups, age groups, and livestock definitions are virtually the same. However, the reference dates and estimation procedures are different. AgDiv produces sheep inventory estimates twice a year while NASS produces these estimates only for January 1 each year. Both organizations produce January 1 and July 1 cattle inventory estimates. For hog estimates, the reference dates are off by one month. NASS's reference dates are December 1, March 1, June 1, and September 1 with AgDiv dates being January 1, April 1, July 1, and October 1. NASS conducts a survey for each of the 4 quarterly estimates while AgDiv makes the estimates for April 1 and October 1 without the use of a survey. The inventory estimates are based on administrative data and previous survey data.

Both organizations, in spite of the differences in structure, make use of market trends and information provided by field experts. The primary estimation tool for both organizations is the balance sheet. AgDiv estimates are made so the balance sheet residual is zero where NASS will generally allow small residuals to remain.

Crop estimates for seeding intentions and for preliminary yield surveys have a different base concept between NASS and AgDiv. The seeding intentions estimates from NASS are designed to forecast what the actual planted acres will be, thus requiring the statistician to make a forecast of the seeded acres. For AgDiv, seeding intentions are designed to be a point estimate showing current seeding plans of farmers, without making a projection of what planted acres will actually be.

The same concept holds true for yield surveys. With NASS, data analysis done with the monthly yield surveys is designed to forecast the final yield and production of the each commodity. NASS uses objective yield surveys for the wheat (winter, spring & durum), corn, soybeans, cotton and potatoes. The objective yield surveys are conducted in the major producing states and usually account for over 80% of the production. The objective yield models are designed to compare current conditions with historic conditions and compare to final yields. For AgDiv, data analysis is designed to estimate current yields, with no adjustments made for historic trends or comparisons of survey indications to final yields. The only objective yield survey for AgDiv is for potato production.

Conclusion

Most of the differences mentioned above have advantages and disadvantages when compared together. Despite the differences in structures and external controls, both organizations produce a wide array of agriculture statistics that are used to establish agriculture policy for the respective countries. However, the background and even the definition of what the data represent are often quite different. Therefore, when comparing the data from each organization, the data concepts are equally as important as the numbers themselves.

ANALYSIS OF RESPONSE BIAS IN THE JANUARY 1992 CATTLE ON FEED REINTERVIEW PILOT STUDY AND THE JULY 1992 CATTLE ON FEED REINTERVIEW SURVEY

Robert Hood, USDA/NASS
Research Division/3251 Old Lee Hwy., Fairfax, VA 22030

Abstract. To assess the accuracy of reported cattle on feed (COF) inventory, the National Agricultural Statistics Service (NASS) developed a series of reinterview surveys to study response bias and to identify specific reasons for reporting errors in order to improve the survey instruments, training and estimation for COF inventory. A three-phase plan, including a pilot study in January 1992, a semi-operational survey in July 1992 and a fully operational survey in January 1993, was designed to meet these objectives. This paper discusses the results of the January 1992 and July 1992 COF reinterview studies.

For each study, a subsample of respondents reporting for the parent survey was recontacted for face-to-face reinterviews in which a subset of the original questions was re-asked. Differences between the reinterview response and the original parent survey response were reconciled to determine a final "proxy to the true value", which was used to measure response bias.

Although no bias estimates were possible for the January pilot study, useful cognitive information was collected. For July, response bias estimates were generated for several survey items. Although differences were observed between the reinterview responses and the original parent survey responses, the net bias was not significantly different from zero for total COF inventory. The contribution to the bias due to reasons for differences between the responses was also examined to detect any underlying relationships.

I. INTRODUCTION

Over the years, the National Agricultural Statistics Service (NASS) has conducted a variety of reinterview surveys to evaluate the quality of its Agricultural Surveys (AS). The purpose of these reinterview surveys has been to study response bias (as opposed to response variance) and to determine reasons for reporting errors. To assess the accuracy of reported cattle on feed (COF) inventories, a new series of reinterview surveys were developed to study response bias. Specific reasons for reporting errors were obtained to guide efforts to improve the survey instruments, training and estimation for cattle on feed inventory. The main focus of this reinterview program was cattle on feed reporting by smaller farmer-feeder operations, as opposed to larger commercial feedlots. A three-phase plan was designed to implement a reinterview program for COF at NASS. This plan included a one-state pilot study in January 1992, a two-state semi-operational survey in July 1992 and a fully operational five-state survey in January 1993. This paper discusses the setup and results of the first two steps.

In estimating response bias, a "proxy to the true value" must first be obtained. In this study, as in previous reinterview studies at NASS, the reconciled value was considered to be the "true" or final value. Considerable cost and effort was expended to ensure that the value obtained during reconciliation was the best proxy to the true value, as reinterviews were done face-to-face and conducted by supervisory and experienced enumerators. When the original and reinterview responses differed, the enumerators were instructed to determine the "correct" response during the reconciliation process. If there was no difference, i.e. the same response was given during both interviews, this common response was considered the final value. If the respondent could not determine which response was correct, or if a difference was not reconciled by the enumerator, the final value was missing and the observation was not used for that item. If the respondent indicated that either response could be correct, then the average of the two responses was used as the final value. A third response, different from both the original and reinterview responses, was also possible if the reinterview respondent said that neither the original nor the reinterview response was correct.

The formulas used to calculate response bias and variance estimates were based on a stratified random sample design. For the ith observation in stratum h, response bias was measured as: $B_{hi} = O_{hi} - F_{hi}$ for stratum $h = 1,\ldots,L$ and unit $i = 1,\ldots,n_h$ where

O_{hi} = original response
F_{hi} = final or reconciled value.

A negative bias indicates underreporting of a survey item, whereas a positive bias indicates overreporting.

II. REINTERVIEW PROCEDURES

For both January and July 1992, a subsample of respondents reporting for the respective parent Agricultural Survey was recontacted by supervisory and experienced enumerators for face-to-face reinterviews. To get the most accurate data possible, enumerators were instructed to contact the person most knowledgeable about the operation, even if that person was not the same as the parent survey respondent. Reinterviews were to be conducted within ten days of the initial survey in order to minimize recall bias.

Responses to the parent survey were provided to the enumerators in a sealed envelope on a reconciliation form. The reconciliation form contained the questions that appeared on both the parent survey and the reinterview survey; the parent survey responses; and spaces to record the reinterview response, the reconciled "correct" response, and a written explanation in the event that a difference between responses occurred. To maintain the independence between the two responses, the envelopes containing the original parent survey responses were not to be opened until after the reinterview was completed. Having two independent responses and asking the respondent to resolve any discrepancies enabled us to obtain the best possible data.

Immediately after conducting the reinterview, the enumerator would open up the reconciliation form and explain to the respondent that he/she had the information obtained from the initial survey and would like to compare the responses for the few items that appeared on both interviews. Each difference (no matter how small) would then be reconciled to obtain the "correct" response, and a written explanation for why the difference occurred would be recorded on the reconciliation form.

The reinterview questionnaire, used to collect a second independent response for comparison to the original response, was similar to but shorter than the parent survey for both January and July. Reinterview questionnaires for January and July were almost identical. Questions that were common to both the parent survey and reinterview survey included questions pertaining to basic operation description, total cattle on feed inventory and total cattle inventory. Some questions were shortened by dropping "include" and/or "exclude" phrases, while others were reworded in order to ensure that the reinterview/reconciliation process obtained the best "proxy to truth". If a cognitive problem exists with the current operational wording of a particular question, then simply re-asking the question the same way may not uncover an underlying response bias. Since questionnaire wording was to be studied, enumerators were instructed to ask the reinterview questions exactly as worded on the questionnaire. The reinterview questionnaires for both January and July contained additional "cognitive" questions as well as a section on terminology (in which the respondent was asked to give his/her definition of some terms currently being used in our surveys) to be used in evaluating survey definitions and concepts, as well as questionnaire wording. "Probing" questions were also asked to determine if all cattle on feed were being reported and being reported accurately.

II. January 1992 Pilot Study

In January 1992, a reinterview pilot study was conducted in Iowa during the NASS January Agricultural Survey. The objective of this study was to work out the logistics of conducting a reinterview survey for cattle on feed and to field test the reinterview and reconciliation forms. A small non-random subsample of respondents to the January Agricultural Survey who were initially contacted by Computer Assisted Telephone Interviewing (CATI) were selected for face-to-face reinterviews. The subsample was concentrated roughly within a hundred mile radius of the State Statistical Office located in Des Moines. Samples eligible for reinterview were those that reported positive cattle on feed capacity on the initial CATI interview. Of the thirty-two completed reinterviews, twenty-six reported both positive cattle on feed capacity and cattle on feed inventory, while six reported positive capacity but no inventory.

Although no response bias estimates or other statistics were possible for this small non-random sample, the logistics of conducting a reinterview for cattle on feed were worked out and information on the problems of reporting cattle and cattle on feed data were obtained. Some general results from the January pilot study are listed below.

- Cattle were often misclassified. The reference to **heifers** in three of the six breakdowns seemed to confuse the CATI respondent as to which category should be used, often resulting in some animals being counted twice.

- Collecting data by phone can be difficult, especially when a question contains multiple categories, such as the cattle breakdowns consisting of six possible categories. The respondent cannot see all the possible choices at one time, thus he does not know what his options are and may include animals in one category that should be included in a later category. Several respondents said they would not have had to adjust their numbers as often if they had known all the choices beforehand.

- Placing animals into the correct weight categories was difficult for both CATI and reinterview respondents. There was a lot of guessing as to whether or not cattle were over 500 pounds. Animals less than 500 pounds are considered to be calves by NASS.

- Total cattle inventories were often misreported due to incorrect classification of animals and by the placement of animals into more than one category.

- There was great variability in the definition of a calf among the respondents for this survey. Some respondents used weight as a criterion, while others specified age.

- Reported feedlot capacity for cattle on feed probably indicates the maximum number an operation **could ever** hold, not the maximum number that **would normally** be fed for the slaughter market.

III. July 1992 Reinterview Survey

The July 1992 Cattle on Feed Reinterview Survey was designed as a semi-operational survey to facilitate the transition from a research activity to an operational program in January 1993. The primary objectives were to provide real-time response bias estimates for agency use, to expand the domain of samples eligible for reinterview beyond CATI, and to continue collecting cognitive information to improve both the reinterview and operational survey instruments.

Reinterviews were conducted on a subsample of the July Agricultural Survey (AS) respondents originally contacted by CATI in Iowa and Minnesota. A small subsample of non-CATI respondents were also selected for reinterviews in Iowa. Making non-CATI samples eligible for reinterview was an innovation for reinterview studies at NASS. The non-CATI domain was included because it continues to represent a significant amount of our AS data collection, particularly during the January AS for which the reinterview program is designed. A stratified random sample with stratum sampling rates similar to the parent survey stratum rates was allocated for reinterview. There was a total of 440 samples selected for reinterview, with 220 in each state. Of these, only completed parent survey samples, including those coded out-of-business, were eligible for reinterview. Parent survey refusals and inaccessibles were ineligible for reinterview. Out of the 440 units selected for reinterview, 303 units were eligible for reinterview and 266 had both usable reinterview and parent survey data. The reinterview non-response rate (for the 303 eligible units) was only 9.2%.

For July, response bias estimates for total cattle on feed and total cattle and calves were generated at both the state and the two-state combined levels. Response bias estimates were calculated for original response minus final response and for edited data minus final response. Original and edited data produced similar results with respect to statistical significance for the two states. Response bias estimates for edited minus final values are shown in Table 1. No significant response bias was detected for total cattle or cattle on feed at either level. There was wide variability in the response bias estimates in both magnitude and direction (i.e., positive or negative) between the two states for total cattle on feed. Iowa reporting showed negative biases of 2.8% compared to positive biases of 13.4% for Minnesota. Although no significant response bias was detected, differences between the initial and reinterview surveys did occur. Nearly half (48%) of the responses differed between the two surveys for total cattle and about one quarter (24%) of the responses differed for total cattle on feed. The differences simply tended to cancel each other out.

The precision of the bias estimates was very low, as indicated by the large standard errors, relative to the bias estimates. The small sample size was not the only factor influencing the bias estimates and the significance tests. The actual number of non-zero differences played an important role also. Although there were 266 usable observations overall, the actual number of differences was far less for each item. There were 52 non-zero differences for cattle on feed and 112 for total cattle and calves. These few differences were spread over 10 strata in Iowa and 8 strata in Minnesota. With such a structure, the small number of non-zero differences, the large number of zero differences, and the large expansion factors resulted in extreme variances which resulted in low precision for the response bias estimates. This lack of precision of response bias estimates is a problem that continues to plague us with reinterview surveys. Work continues on sample design and estimation improvements to increase our response bias estimation precision.

Table 1. Response Bias Estimates for the July 1992 COF Reinterview Survey.

	Edited Value - Final Value			
Item/State	Bias	% of Edited	Standard Error	95% CI
TOTAL COF				
Iowa	-25,912	-2.8	3.7	(-10.0, 4.4)
Minnesota	60,117	13.4	10.5	(-7.3, 34.0)
Total	34,205	2.5	4.0	(-5.3, 10.3)
TOTAL CATTLE				
Iowa	-74,411	-1.8	2.8	(-7.4, 3.7)
Minnesota	-48,080	-1.9	2.4	(-6.7, 2.9)
Total	-122,491	-1.9	2.0	(-5.8, 2.0)

IV. REASONS

One of the goals of the July reinterview survey was to identify the reasons for discrepancies between the original and reinterview responses in order to evaluate the questionnaires and to determine how much of the bias may be fixable. During the reconciliation process, explanations were recorded by enumerators for each difference that occurred between an original and reinterview response. These reasons were then grouped into three general categories, "estimation or rounding", "definition or interpretation" and "other" (i.e., reasons that could not be attributed to the first two categories). In general, differences due to "definitional" reasons can be viewed as being potentially fixable by changes in the survey instruments, procedures or training. Differences due to "estimation" or "other" reasons probably are not as correctable, if correctable at all.

Since response biases can be positive or negative and therefore cancel each other out, using the net bias could be misleading when analyzing biases. Therefore, the absolute value of each non-zero difference was expanded to obtain the total absolute response error for each reason category. Table 2 shows the frequency of differences by reason category and the percentage of the total absolute response error attributable to each category. While "estimation" reasons accounted for 38.5% and 20.5% of the differences for COF and total cattle, respectively, these reasons contributed the least to the total absolute response error (8.6% for COF and 4.9% for total cattle). "Definitional" reasons were responsible for the majority of the total absolute response error for COF, accounting for 66.4% of the bias, while "other" reasons, responsible for 60.7% of the bias, contributed the most for total cattle. Table 2 shows that there is opportunity for improvement in the survey procedures, instructions and questionnaires. Recall that reasons due to "definitional" problems are considered fixable. "Definitional" reasons accounted for almost two-thirds of the total absolute response error for COF and over one-third for total cattle.

Table 2. Percentage of Total Absolute Response Error by Reason Category for Original Minus Reconciled Values. Frequencies of Response Errors are Shown in Parenthesis.

	Reason Category			
Item	Estimation	Definition	Other	Total
Total Cattle on Feed	8.6% (20)	66.4% (19)	25.0% (13)	100% (52)
Total Cattle & Calves	4.9% (23)	34.4% (19)	60.7% (70)	100% (112)

Table 3. Frequency Table of Relative Bias by Reason Category (Two States Combined).[1]

	Reason Category					
	Estimation		Definition		Other	
Item/Relative Bias[2]	# of Obs	% of Bias	# of Obs	% of Bias	# of Obs	% of Bias
Total Cattle on Feed						
Bias ≤ 20%	17	(85%)	3	(16%)	5	(38%)
Bias > 20%	3	(15%)	16	(84%)	8	(62%)
Total	20	(100%)	19	(100%)	13	(100%)
Total Cattle & Calves						
Bias ≤ 20%	21	(91%)	9	(47%)	52	(74%)
Bias > 20%	2	(9%)	10	(53%)	18	(26%)
Total	23	(100%)	19	(100%)	70	(100%)

[1] Includes only observations with a bias

[2] Relative Bias = 100 * (Original value - Reconciled value)/Reconciled value

In order to study the relationship between the magnitude of the bias and the reason categories, a relative (percentage) bias was calculated for each observation with a non-zero difference between the original value and reconciled values. Two levels of relative bias were used - less than or equal to 20% in magnitude and greater than 20% in magnitude. Table 3 shows the relationship between the magnitude of the relative bias and the reason categories. The results indicate that there is a significant relationship between the magnitude of the relative bias and the reason categories. "Estimation" reasons tended to be associated with smaller biases for both items. "Definition" reasons were associated with larger biases for cattle on feed but were more evenly distributed for total cattle. "Other" reasons were associated with larger biases for cattle on feed but with smaller biases for total cattle.

A Closer Look at Total Cattle on Feed

The primary focus of this series of reinterview surveys (i.e., the January 1992, July 1992 and January 1993 surveys) was cattle on feed inventory. The reinterview program for cattle on feed grew out of the concern that inventories were being overreported in the farm feeder states. Thus, the results of the July 1992 reinterview study may have been somewhat surprising. No statistically significant bias at the individual or combined state levels was detected. In fact, the results indicated only a slight overreporting of 2.5% at the combined level. Iowa reporting indicated a slight underreporting of 2.8%. Minnesota overreporting was estimated 13.8%, but the variance was large enough for the result to be insignificant. Do these results then indicate that there is no problem? Not necessarily! What must be remembered when looking at the results from July is the sample size was very small. With such a small sample size (recall that there were only 266 usable samples), the results are very volatile and mistakes on just a few reports can have an enormous impact on the final bias estimate.

Table 2 showed the percent of the total absolute response error accounted for by each of the three reason categories. "Definitional" reasons were the major contributor, accounting for 66% of the total absolute response error. "Other" reasons were responsible for 25% and "estimation" reasons for about 9%. The differences attributable to "definitional" reasons are listed below in Table 4. Also shown are their individual percent contribution to the "definitional" absolute error and the number of times each reason was reported.

For each of the five reported "misunderstandings", the reinterview response was determined to be the correct response during reconciliation. The source of the reporting error for these five samples was attributed to either the initial respondent, the initial enumerator or both. The same person responded for two of the five reports. For the five cases of "did not understand question", the reinterview response was also determined

Table 4. Definitional Reasons Reported for Total Cattle on Feed (Two States Combined).

Reason for Difference	% of Definitional Absolute Response Error	Number of Times Reported
Included cattle/calves from another operation	0.5	1
Did not report as of the reference date	4.9	2
Respondent did not figure death loss in total	7.2	2
Respondent did not understand the question	9.6	5
Respondent forgot to include some cattle or calves	13.7	4
Misunderstanding between enumerator & respondent	64.1	5
Total	100.0	19

to be the correct response. The source of error was attributed to the initial respondent in four cases and to both the initial respondent and the initial enumerator in the other case. The same person responded to four of these five cases.

For cattle on feed inventory, there was a total of 52 non-zero differences between the original and reinterview responses (excluding one outlier); 34 in Iowa and 18 in Minnesota. There was variability in the composition of the differences between and within the two states. Iowa had about four times as many negative differences as Minnesota (21 vs. 5). Minnesota had more positive differences than negative (13 vs. 5), while the opposite was true for Iowa (21 negative vs. 13 positive).

Of the 13 negative differences for Iowa, 4 were due to a "misunderstanding between the enumerator and respondent", accounting for 46% of the total negative bias for cattle on feed in Iowa. Two cases in which the "respondent forgot to include some cattle or calves" accounted for almost 21% of the total negative bias. Eight "estimation" reasons accounted for only 12% of the total negative bias. As for the positive differences, the major contributor was one case in which the "respondent had not made a decision on marketings", accounting for almost half of the total positive bias for Iowa.

Whereas the reason "misunderstanding between enumerator and respondent" accounted for 46% of the total negative bias for Iowa, one difference due to this reason was responsible for 70% of the total positive bias in Minnesota. For Minnesota, the five negative differences contributed very little to the overall bias. In all, there were seven "estimation", eight "definitional" and three "other" reasons for Minnesota. To demonstrate just how volatile the bias estimates were, without the one difference due to a "misunderstanding", the percent bias in Minnesota would have dropped from 13.4% to only 3.7%.

In order to reduce response bias and improve data collection, enumerator training should emphasize the reason why a reinterview is being conducted, why it is important to read the questionnaires exactly as worded and the importance of a positive attitude when conducting a reinterview. With the relatively small sample size, data quality is very important. As was seen in the July 1992 reinterview survey, one observation can completely change both the magnitude and direction of the bias estimates for a survey item, so taking the time to collect good data must be stressed.

CONCLUSION

Although the January and July 1992 reinterview studies did not detect any significant overall response bias for cattle on feed and total cattle inventories, useful information on problems associated with reporting cattle on feed, as well as cattle, was obtained. The two studies showed a substantial number of differences between original and reinterview responses which resulted in great variability. However, the differences were nearly offsetting, resulting in non-significant response bias. The results also indicated that there may be room for improvement in the current survey procedures (including questionnaire design and wording) used to collect COF data. "Definition or interpretation" problems were found to account for nearly two-thirds of the total absolute response error for COF. This can be looked upon as being both good and bad. It is bad in the sense that so much "definitional" bias indicates that there may be a problem with the operational survey. However, it is good in the sense that "definitional" problems are considered more "fixable" than "estimation" or "other" problems. In our efforts to reduce response bias and to improve the survey instruments, high priority ought to be given to reducing the errors attributed to "definitional" reasons.

AN EVALUATION OF ROBUST ESTIMATION TECHNIQUES FOR IMPROVING ESTIMATES OF TOTAL HOGS

Susan Hicks and Matt Fetter, USDA/NASS
Susan Hicks, Research Division, 3251 Old Lee Hwy., Room 305, Fairfax, VA 22030

I. Abstract

Outliers are a recurring problem in agricultural surveys. While the best approach is to attack outliers in the design stage, eradicating sources of outliers if possible, large scale surveys are often designed to meet multiple, conflicting needs. Thus the survey practitioner is often faced with outliers in the estimation stage. Winsorization at an order statistic and Winsorization at a cutoff are two procedures for dealing with outliers. The purpose of this paper is to evaluate the efficiency, in terms of true MSE, of Winsorization for improving estimates of total hogs at the state level and to evaluate the efficiency of a data-driven technique for determining the optimal cutoff.

KEY WORDS: Outlier, Winsorization, Minimum Estimated MSE Trimming

II. Design of the Quarterly Agricultural Surveys

The National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture (USDA) provides quarterly estimates of total hogs at the state and national level through its Quarterly Agricultural Surveys (QAS). The QAS uses both a list and an area frame in a multiple frame (MF) approach to provide estimates for a variety of commodities in addition to hogs. Known farm operators are included on the list.

The area frame sampling is based on land use stratification. All land in the contiguous 48 states has a positive probability of selection in the area frame. Thus, the area frame is a complete frame and can be used to measure undercoverage in the list frame. Tract operations found in the area sample are matched to the list frame. Operators not on the list comprise the Non Overlap sample, or NOL.

III. Why are Estimates of Total Hogs so Variable?

Providing reliable estimates of total hogs through the multiple frame approach has always been difficult. The sampling variability in hog estimates is closely associated with the sampling variability in the NOL. While the NOL typically contributes only 25% to the total estimate, its contribution to the total variance is around 75%.

Outliers in the NOL can severely distort the estimates. Rumburg (1992) studied the causes and characteristics of NOL outlier records in five states. He cited three major contributors to outliers in the NOL:

- increased weights due to subsampling,
- the transitory and variable nature of hog production, and
- the location of hog operations on land with little agriculture.

The area frame stratification, based on land use strata, is more efficient for field crops than for livestock items, which tend to be less correlated with land use. Basically the variability in the NOL domain, which is a subset of the area frame, can be attributed to two factors:

1) the population within each strata is highly skewed to the right, and

2) the sample size is small.

IV. What is a Hog Outlier?

Most of the literature on truncation estimators for survey sampling describes its application to the problem of variability in weights. For household surveys, where we're frequently estimating Bernoulli characteristics, outliers are indeed caused by extreme sampling weights. For agricultural surveys extreme observations are caused by a combination of moderate to large weights and moderate to large values.

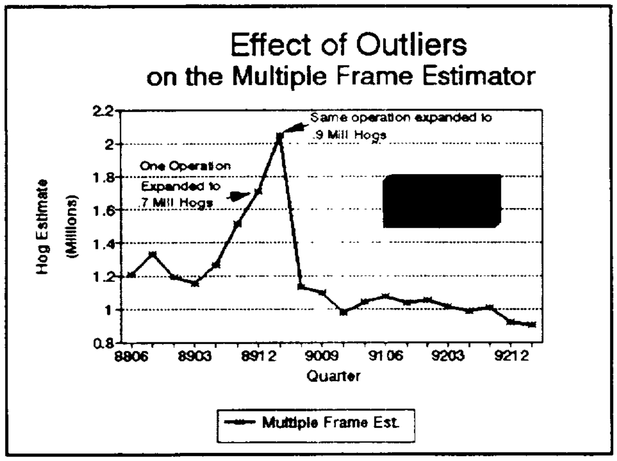
Lee (1991) addressed this problem by differentiating between outliers defined by classical statistics and influential observations. In classical statistics, outliers are unweighted values situated far away from the bulk of the data. Influential observations are valid reported values that may have a large influence on the estimate. Influential observations may involve outliers, but more frequently are a combination of relatively large sampling weights and relatively large data values. For our purposes the term outlier will refer to influential weighted survey values, not unweighted values.

V. Current Procedures for Handling Outliers

Currently each state reviews potential outliers during the editing stage. Typically the 20 largest weighted values are listed through the Potential Outlier Prints System (POPS). The state commodity statisticians review these outputs and questionable data are verified. If the weighted data are correct, no changes are made.

However, the preliminary state hog recommendations are adjusted for outliers. Figure 1 shows the effect of extreme outliers in Georgia. In December of 1989, the operation that expanded to .7 million hogs comprised approximately 42.4% of the total multiple frame estimate. This is exactly the kind of situation we want to correct for.

Figure 1



Currently, outliers are adjusted in a somewhat ad hoc fashion at the state level. Treatment of outliers could include truncating the weight to 1.0, truncating the weight to some other value, or not truncating the weight at all. Although the effect of outliers is compensated for in the state recommendation, the records are never changed. This avoids the potential for biasing the national indication. Outliers at the state level are rarely outliers at the national level.

VI. Description of Two Winsorization Estimators

We evaluated two types of robust estimators for improving state level hog indications: Winsorization at a cutoff, t, and Winsorization at r order statistics. The form of the estimator which adjusts to a fixed cutoff is:

$$\hat{Y}_t = \sum_{j=1}^{n} adj\ w_j^t\ y_j \qquad (1)$$

where:

t = *truncation level*

y_j = *reported hogs for j^{th} unit*

$$adj = \frac{\sum^{n} w_j}{\sum^{n} w_j^t}$$

$$w_j^t = \begin{cases} w_j & \text{if } w_j y_j \leq t \\ t/y_j & \text{if } w_j y_j > t \end{cases}$$

w_j = *design weight for j^{th} unit*

In this version of the standard truncation estimator, we truncate the weights of those observations whose weighted value expands larger than t so that the expanded value now equals t. The truncated portions are then "smoothed" over all observations.

We also evaluated estimators which adjust for the r largest values. The form of the estimator is:

$$\hat{Y}_r = \sum_{j=1}^{n} adj\ w_j^r\ y_j \qquad (2)$$

where:

$$w_j^r = \begin{cases} w_j & \text{for } j=1,\ldots,n-r \\ \dfrac{w_{n-r} y_{n-r}}{y_j} & \text{for } j=n-r+1,\ldots,n \end{cases}$$

$$adj = \frac{\sum^{n} w_j}{\sum^{n} w_j^r}$$

To evaluate the efficiency of each estimator for improving estimates of total hogs at the state level we developed a monte carlo simulation.

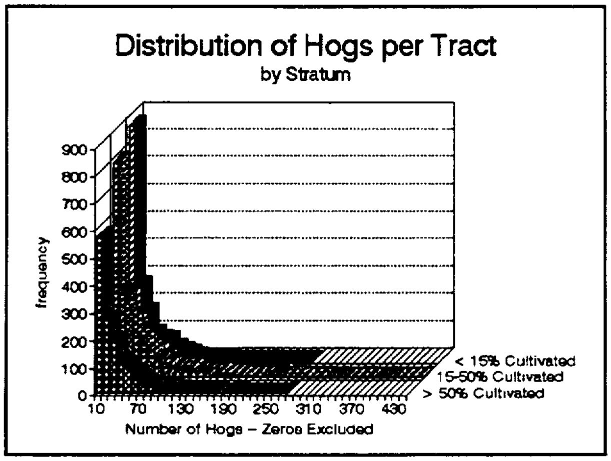
VII. Description of the Monte Carlo Simulation

We built our simulation around one state, Georgia. Because of the complexities of the multiple frame design and because the major source of outliers and sampling variability is from the NOL, we restricted the simulation to the NOL domain of the area frame.

A positive NOL segment is defined as any sampled segment that contains at least one NOL hog operation. Separate farm operations are divided into tracts. The number of sample segments in the area frame is fixed while the number of sample segments in the NOL domain is variable. For the NOL domain the sample unit is the segment, but the reporting unit is the tract. To simplify the simulation we modelled the positive NOL tracts and assumed a fixed NOL sample size.

The tract level weight is a product of the stratum sampling weight and a tract adjustment factor. The adjustment factor prorates an operation's reported value back to the tract level for operations that only partially reside within the sample segment. Based on historical June data from '91 and '92 for Georgia, we developed parametric models of the weighted tract level hog data for positive NOL tracts.

Figure 2



The three models -- one for each strata -- were all gamma density functions. Due to the sparseness of the data it was difficult to validate the model. However, our main interest was in developing a reasonable, highly skewed distribution rather than developing highly accurate models for Georgia NOL tracts.

We created a fixed sample universe for each stratum based on the estimated gamma distribution and the estimated number of positive NOL segments. We also estimated the proportion of positive NOL segments to total NOL segments. With the zero segments included, the result is a highly skewed population with a large spike at zero and a long right tail. See Figure 2.

From the fixed universe, we drew 1000 stratified simple random samples with replacement. Table 1 was created using a SAS program based on 1000 samples. Some of the other graphs were created from a Fortran program using the same data based on 10,000 samples. To compare the performance of the estimators for different sample sizes we chose a sample of size 360 to mimic the June sample and a sample of size 216 to mimic a follow-on sample.

The efficiency of the estimators was estimated as the ratio of the MSE of the unbiased estimator to the MSE of the new estimator. See Table 1 in next section.

VIII. Evaluation of Winsorization at a Cutoff and Winsorization at an Order Statistic

Ernst (1980) compared seven estimators of the sample mean which adjust for large observations. Four of the estimators were modifications of Winsorization at a cutoff, t. The other three estimators were modifications of Winsorization at an order statistic. Ernst showed that for the optimal t, the estimator which substitutes t for the sample values greater than t has minimum mean squared error. Earlier work by Searls (1966) showed that gains are achieved for wide choices of t when the data originate from an exponential distribution. The results from our monte carlo study are consistent with those studies.

Table 1

Truncation Level	MSE Ratio	Number Truncated	MSE Ratio
June			
14000	1.243	1	1.018
12000	1.299	2	.980
10000	**1.321**	3	.918
8000	1.236		
Follow-on			
18000	1.396	1	1.034
15000	**1.440**	2	.993
12000	1.402	3	.908
9000	1.175		
6000	.739		

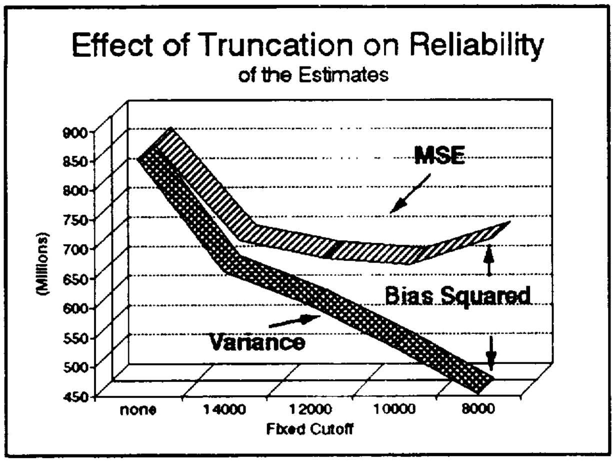
For both the June and follow-on samples, Winsorization at a cutoff is more efficient than Winsorization at an order statistic. Further for smaller cutoffs, more samples are truncated and more observations are truncated per sample. Thus, the bias in the estimator increases. Figure 3 shows the decomposition of variance and bias at each level of trimming evaluated for the June sample size.

In choosing a cutoff for truncation we'd like to minimize the number of samples truncated. In general, we don't want a cutoff that truncates every sample, but rather a cutoff that corrects for rare extreme observation like that depicted in Figure 1.

IX. Minimum Estimated MSE Trimming

In practice, we do not know the underlying distribution of the data and likewise the optimal value for t. The survey practitioner has to weigh the benefits of trimming -- decrease in variance -- with the

Figure 3



costs -- increase in bias.

The obvious criterion for evaluating the effect of different trimming levels is the estimated MSE. Potter (1988) documented a procedure called Minimum Estimated MSE Trimming. Because we do not know the true parameter, Y, we are limited to evaluating the bias based on the unbiased estimator $\hat{Y}$. The estimate of MSE ($\hat{Y}_t$) is derived from the relation:

$$\begin{aligned} E(\hat{Y}_t-\hat{Y})^2 &= Var(\hat{Y}_t)+Var(\hat{Y})-2Cov(\hat{Y}_t,\hat{Y}) \\ &\quad +[E(\hat{Y}_t)-E(\hat{Y})]^2 \\ &= MSE(\hat{Y}_t)+Var(\hat{Y})-2Cov(\hat{Y}_t,\hat{Y}) \end{aligned}$$

where:

$\hat{Y}_t$ = *the trimmed estimate*

$\hat{Y}$ = *the unbiased estimate*

Thus, an unbiased estimate of MSE($\hat{Y}_t$) is:

$$M\hat{S}E(\hat{Y}_t)=(\hat{Y}_t-\hat{Y})^2-\hat{V}(\hat{Y})+2C\hat{O}V(\hat{Y}_t,\hat{Y})$$

If the correlation between the truncated estimate and the untruncated estimate is approximately 1.0, then this reduces to:

$$M\hat{S}E(\hat{Y}_t) = (\hat{Y}_t-\hat{Y})^2 - \hat{V}(\hat{Y}) + 2\ [\hat{V}(\hat{Y}_t)\hat{V}(\hat{Y})]^{1/2} \quad (3)$$

In this procedure, the estimated MSE is computed for various trimming levels and the trimming level with the minimum MSE is selected for implementation. This procedure can be used to suggest optimal trimming levels in (1) or number of observations to trim in (2). While the minimum MSE technique should identify the optimal trimming level over many samples, for any particular sample it could identify a trimming level far from the optimum. This occurs because our estimate of MSE is conditional on the sample we have drawn.

The efficiency of this estimator, in estimating the true MSE, depends on the efficiency of the variance estimators and the validity of the correlation assumption. In general, for a simple design and ignoring the effects of editing and nonresponse adjustments, we have unbiased estimators for V($\hat{Y}$). However, obtaining an unbiased estimator for the variance of a truncation estimator is less straightforward. One approximation that is often made is to estimate the sampling variance of $\hat{Y}_t$ by treating the trimmed weights as if they represented the untrimmed weights in the usual variance formulae. We have used this approximation in (3).

X. Evaluation of Minimum Estimated MSE Trimming as a Data-driven Estimator

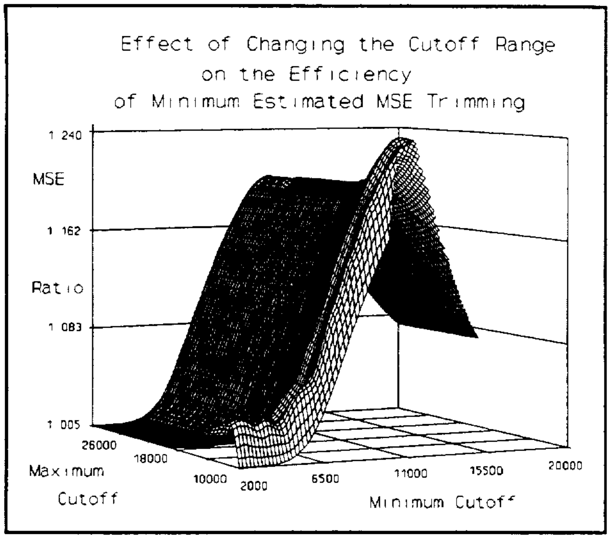
We wanted to evaluate the efficiency of the minimum MSE technique as a data-driven estimator. With this estimator each sample would be truncated at different levels to determine the level that minimized (3). Thus, the cutoff varies from sample to sample. Some preliminary runs showed that the efficiency of this data-driven estimator was highly dependent on the range of trimming levels evaluated. Thus, we evaluated this estimator over the set of possible trimming levels. The minimum trimming levels ranged from 2000 to 20,000 and the maximum trimming levels ranged from 10,000 to 28,000. We evaluated this estimator based on 10,000 monte carlo samples. For each monte carlo sample, the MSE estimator in (3) was used to determine the optimal cutoff for that sample for each range of trimming levels. The minimum and maximum trimming levels were each incremented by 1000 covering all combinations within the ranges specified above.

The level of the estimate, $\hat{Y}_t$, that minimized (3) was retained for each cutoff range and sample. The true MSE of the estimator for each combination of minimum and maximum cutoff was calculated in the usual fashion based on 10,000 $\hat{Y}_t$ estimates. And the MSE ratio is as defined before.

Recall from Table 1 that the optimal cutoff is around 10,000 to 11,000. As Figure 4 shows, this estimator is most efficient when the range of trimming levels evaluated is close to the optimum trimming level. As the minimum cutoff is reduced the efficiency of the estimator drops off rather dramatically. Whereas increasing the maximum cutoff has a minimal effect on the estimator.

For any particular sample the estimated MSE technique could identify a trimming level far from the optimum. This occurs because the estimated MSE is conditional on the sample we have drawn. Thus as Figure 4 shows, the efficiency of this technique as an estimator depends on the range of cutoffs we choose

Figure 4



to evaluate and how close that range is to the true optimum, similar to Winsorization at a cutoff.

Figure 5

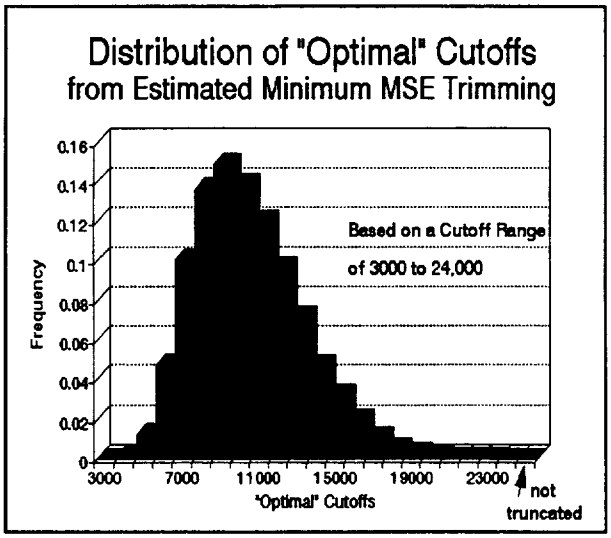


Figure 5 shows the distribution of estimated "optimal" cutoffs in increments of 1000 when this estimator is evaluated over the range 3000 to 24000.

Again, we see that the estimated "optimal" cutoff is data-dependent.

XI. Recommendations

As has been proven theoretically by Ernst (1980), Winsorization at the optimal cutoff is preferable to Winsorization at an order statistic. We believe this estimator holds promise for improving NASS state level hog indications. Further, while the data shows that Winsorization at a cutoff performs well for a wide range of cutoffs, the best efficiencies are obtained for cutoffs greater than or equal to the optimum. However, if we adopt the fixed cutoff estimator, we need to be careful about choosing the cutoff.

Minimum Estimated MSE Trimming provides an alternative to Winsorization at a cutoff when the optimal cutoff is unknown as is frequently the case. However, the efficiency of this data-driven estimator is dependent on how close the range of cutoffs evaluated is to the true optimum. And this technique is computationally intensive.

Minimum Estimated MSE Trimming could be used to suggest optimal cutoffs at the state level, but as Figure 5 shows the estimated cutoff is still highly dependent on the sample data. In the end, determining the optimal cutoff for complex sample designs may remain more art than science.

REFERENCES

Ernst, L. R. (1980), "Comparison of Estimators of the Mean Which Adjust for Large Observations," Sankhyā: The Indian Journal of Statistics, 42, 1-16.

Lee, H. (1991) "Outliers in Survey Sampling," prepared for the Fourteenth Meeting of the Advisory Committee on Statistical Methods.

Potter, F. (1988) "Survey of Procedures to Control Extreme Sampling Weights," Proceedings of the Survey Research Methods Section of the ASA.

Rumburg, S. (1990) "Characteristics of Directly Expanded Hog Data Outliers," NASS Staff Report, No. SRB-92-02, U.S. Department of Agriculture.

Searls, D. T. (1966) "An Estimator for a Population Mean Which Reduces the Effect of Large True Observation," Journal of the American Statistical Association, 61, 1200-1204.

AN ANALYSIS OF VARIANCE APPROACH TO DEFINING IMPUTATION CELLS FOR A COMPLEX AGRICULTURAL SURVEY

John Amrhein, National Agricultural Statistics Service
NASS/RD, 3251 Old Lee Hwy, Fairfax, VA 22030

KEYWORDS: Analysis of Variance, Wald Statistic, Complex Survey, Imputation

Abstract

Conventional analysis of variance techniques, such as F tests, that are used to determine if subpopulation means are significantly different, rely on the assumption that the observations are independent and identically distributed. It is often the case that survey data, collected using a complex sampling design, violate this assumption. This paper demonstrates the use of an analysis of variance model, adjusted for a complex sampling design, as an effective tool to define imputation cells when adjusting for nonresponse in a complex agricultural survey. A solution to the normal equations is derived in the usual manner. However, a resampling technique is used to obtain a consistent estimate of the covariance matrix. This estimate is then used to calculate a Wald statistic for conducting F tests on testable hypotheses. Examples for several survey items are discussed.

Introduction

The use of models in the analysis of sample survey data continues to be an important area of study. Skinner et al. consolidate much of the work that has been accomplished concerning this issue. This paper presents an application of the analysis of variance model to complex survey data. Although the discussion focuses on employing analysis of variance to define imputation cells, the general method described is appropriate for many survey applications involving an analysis of variance.

The first section of this paper discusses two broad approaches to modelling complex survey data and introduces a general linear model to be considered. The Wald statistic is presented as the conventional test statistic for an analysis of variance. An adjustment to the conventional technique that allows the application of an analysis of variance to non-iid observations is discussed. The next section describes, through an application to a complex survey, how an analysis of variance model can be a useful tool to test the effectiveness of imputation cells that are often used when imputing for survey nonresponse.

The Model and the Test Statistic

Complex survey designs are implemented to increase the precision of estimates when there is knowledge about the underlying structure of the population of interest or to facilitate data collection. In such cases the assumption of independent and identically distributed (iid) observations underlying conventional data analysis techniques, found in many computer analysis packages, is invalid. Ignoring the complex sampling design in favor of the iid assumption during data analysis results in biased estimation of sampling variances. Therefore, an improved analysis can be realized by accounting for the complex design.

Skinner et al. discuss aggregated and disaggregated approaches to modelling complex survey data. The aggregated approach models the survey variables of interest at the population level and accounts for the survey design through adjustments to standard analysis procedures under iid assumptions. The disaggregated approach includes the survey design in the specification of the model. For example, columns of binary variables defining membership in strata or clusters can be included in the matrix of independent regressor variables. A solution to the normal equations can be obtained and linear combinations of the coefficients can be used to test for significant differences of the cluster (or subpopulation) means.

Consider the fixed effects analysis of variance model

$$y = X\beta + e \qquad (1)$$

where y is a vector of values for a measured or observed survey item; X is a known design matrix; $\beta' = (\mu, \alpha_1, \dots \alpha_d)$, where d is the number of effects, is a vector of coefficients of unknown value; and e is a vector of residuals such that $E(e)=0$ and $V(e)=V$. Assuming that the residuals are iid Normal random variables, the standard analysis of variance can be conducted. Under these assumptions, the Wald statistic:

$$Q_k=(K'\hat{\beta}-m)'[K'\hat{V}_{\hat{\beta}}K]^{-1}(K'\hat{\beta}-m) \qquad (2)$$

where $\hat{\beta}$ represents a solution to the normal equations, has a central chi-squared distribution with degrees of freedom equal to the rank of K′ under the null hypothesis H_o: $K'\beta = m$, where K´ is a contrast matrix whose rows represent testable linear combinations of the coefficients. By correctly specifying K´ and setting m=0, the significance of the effects in partitioning the variance of the dependent variable can be tested using conventional F tests. Koch et al. present an approach to adjusting conventional analysis of variance techniques for data from complex survey designs; that is, for cases in which the iid assumption is not valid. A further review and example are presented by Freeman where he labels this method the KFF method. In these examples the method of weighted least squares is used to estimate the vector β and a consistent estimator (such as a resampling technique) is used to estimate the covariance matrix for β.

Similarly, Skinner et al. discuss the following procedure, which is followed in this study. Let Π be the diagonal matrix of inclusion probabilities for the sampled units and consider again the model in (1). The weighted least squares estimator for the parameter vector, β, for a model of full rank is

$$\hat{\beta} = (X'\Pi^{-1}X)^{-1}X'\Pi^{-1}Y. \qquad (3)$$

This is the product of the design unbiased, Horvitz-Thompson estimators for X´X and X´Y (Shah et al.). The estimator in (3) is model unbiased under the model in (1) (Skinner et al.) and has true variance

$$V(\hat{\beta}\,|X) = V_{\hat{\beta}} = (X'\Pi^{-1}X)^{-1}X'\Pi^{-1}V\Pi^{-1}X(X'\Pi^{-1}X)^{-1}. \qquad (4)$$

If the estimator for the covariance matrix given in (4) is consistent, then the techniques described for the iid case can be used for tests of significance. In this case, the Wald statistic in (2) is approximately distributed chi-squared. Standard statistical software can be used to conduct the necessary regression, and various resampling techniques are available for consistent estimation of (4). Rao et al. present some recent work concerning the use of the jackknife, balanced repeated replication and the bootstrap methods for inference with complex survey data. A discussion concerning which method is most appropriate is beyond the scope of this paper. This is one of several areas requiring further investigation.

Application to Mean Imputation

Kalton and Kasprzyk explain how most models underlying imputation or reweighting adjustments for nonresponse fit the form of the general linear model in (1) where the design matrix X defines the inclusion of an observation in a reweighting or imputation cell and may also include auxiliary variables, and β is an effects vector. One method of adjusting for nonresponse involves defining population cells in which nonresponse is assumed or known to be ignorable; that is, we want to form cells in which the nonrespondents are considered to be a simple random sample of the original sampled units in that cell and the within-cell variance is small. By conducting an analysis of variance, the contribution of a variable in defining homogeneous groups can be measured by testing if its coefficient is significantly different from zero.

The National Agricultural Statistics Service (NASS) of the United States Department of Agriculture collects crop, livestock and grain stock inventory data through a series of sample surveys. NASS draws samples from a list frame and an area frame and conducts concurrent surveys each June. NASS' area frame consists of the land area of the United States. Each county within each state is stratified based on land use and, in the event of agricultural land, percent cultivation. Substratification is performed based on the type of agricultural activity (crop, livestock, etc.). A two (or sometimes three) stage sampling process selects segments of land for enumeration in June. A segment is a cluster of tracts. Tracts are defined as areas of land within a segment under one operating arrangement. Each tract is associated with an operator so that it can be matched against the list frame. Tracts from the area frame sample that were found to be not eligible for sampling from the list frame, labeled non-overlap (NOL) tracts, are identified and used as a measure of the incompleteness of the list. The expanded (weighted) data from NOL tracts are combined with expanded (weighted) data from the list sample to derive multiple frame totals. NOL tracts are restratified based on data collected in the June survey. Based on the new strata, a stratified simple random subsample of the NOL tracts is drawn for each subsequent, or "follow-on", survey in the twelve month survey cycle following the June survey. As in June, the totals from the area subsample are added to totals from the list sample for full population multiple frame totals. Thus, the overall sample design for selecting NOL tracts for follow-on surveys is a two phase stratified design.

NASS uses agricultural statistics districts (geographical delineations) within each state to define imputation cells to adjust for item nonresponse in the NOL sample of the follow-on surveys. This study was initiated to determine the appropriateness of these cell definitions. Because of data abundance and the one phase design, June NOL observations were used in this study rather than follow-on survey observations. The one phase design is easier to mimic when resampling.

The model analyzed in this study is the following:

$$\mathbf{y}_{ij} = \mu + \alpha_i + \mathbf{e}_{ij} \qquad (5)$$

where: y_{ij}=value of survey item for unit ij
$i=1,\ldots,d$ agricultural statistics districts
where $1<d<9$ for a given state
$j=1,\ldots,n_i$ observations
μ=the population mean for the state
α_i=the effect of the i^{th} ASD
e_{ij}=the residual for the ij^{th} observation
where $E(e_{ij})=0$ and $V(e_{ij})=\sigma^2_i$

Although the above model groups the observations into subpopulations (districts), these groups are aggregates of survey design clusters or strata. Therefore, the aggregated approach as described by Skinner et al. was adopted for this study. The objective was to determine if separating the responses into districts aids in partitioning the variance. Results indicating that districts significantly partition the variance of reported values for the survey item (e.g. cropland acres, total number of hogs, etc.) would provide support for the assumption that the cells are homogeneous groups of iid observations and that nonrespondents are a simple random sample within each cell. It is not an objective to derive a predictive model for the dependent variable. Indeed, the analysis of variance models in this study are not full rank. Therefore, the solutions to the normal equations are not unique and cannot be used for prediction.

The bootstrap technique described by Rao et al. was used, with 250 iterations, to estimate the variance given in (4). The null hypothesis was H_o: $K'\beta=m$, where $K' = [\mathbf{0}_{d-1}\ \mathbf{1}_{d-1}\ \mathbf{-1 \cdot I}_{d-1}]$ and $\mathbf{m=0}$. Index d is the number of districts in the state; i.e. for the fixed effects, α_i, i ranges from 1 to d. Defining K′ in this manner tests that all α_i's are equal or, effectually, that all α_i's equal zero. For a good explanation of why this is so, the reader is referred to Searle.

The cell means were tested for significant differences by calculating a p-value which is defined as the probability that a random variable that is distributed $F_{d-1,\, n-(d+1)}$ is greater than the observed value of the Wald statistic divided by its degrees of freedom. That is:

$$\textbf{p-value} = \textbf{Prob}[\ \mathbf{F}_{d-1,\, n-(d+1)} > \mathbf{Q}_k/(\mathbf{d-1})\ ].$$

When the denominator degrees of freedom, n-(d+1), is low, using the F distribution as described above offers a refinement over comparing Q_k to a χ^2_{d-1} value (Skinner et al. p. 79). For the hypothesis described above, a small p-value would indicate that the district means are significantly different.

Results obtained in this study indicate that districts aid in the separation of the variance of the agricultural survey items in this study. The analysis was conducted using only positive data. The objective was to test if any difference existed between district means for those farms that had the item of interest. Therefore, for example, when analyzing cropland acres, only those farms with reported cropland acres greater than zero were included. The four survey items that were tested are as follows, with the number of observed p-values that were less than .1 over the number of states included in the analysis given in parentheses: cropland acres (36/48), number of hogs (13/30), on-farm grain storage capacity (29/36) and winter wheat harvested acres (6/15). States were excluded from a given analysis if, as in the cases of Alaska and Hawaii, they are not included in the survey program, or there were too few observations for the survey item. For example, Rhode Island has a small number of hog operations and, therefore, was not included in the analysis of total number of hogs. Also, NOL sampled units are used as a measure of list incompleteness. Therefore, we are dealing with fewer observations than if data from the list frame sample were used.

From these findings, it can be concluded that defining imputation cells based on agricultural statistics districts partitions the population into more homogeneous groups than if the cells were defined at the state level. Cell means were found to be significantly different in enough states that NASS should not collapse imputation cells to the state level. It cannot, however, be concluded from this study that districts partition the population into the most homogeneous groups. There may be other auxiliary variables available that partition the population into more homogeneous groups.

Discussion

A natural question to ask is if any improvement in analysis was realized by accounting for the sample design. Therefore, an analysis of variance was also conducted ignoring the sample design and assuming iid observations. The number of states with an observed p-value of less than .1 over the number of states tested, given in parentheses, is as follows: for cropland acres (21/48), number of hogs (3/30), on-farm grain storage capacity (14/36) and winter wheat harvested acres (8/15). The lower occurrence of significance, at a .1 level, (except for winter wheat harvested acres) suggests that the stratification resulted in more precise estimates. The analysis of variance that accounted for the stratification detected differences that the analysis assuming independent observations did not. The difference in results is most apparent with the number of hogs survey item. With only three of thirty states showing significance, one would conclude that districts do not aid in partitioning the variance. The estimated design effects in these cases would be less than one.

In conclusion, the strategy outlined here for conducting an analysis of variance is an effective tool that can be used to determine the effectiveness of the imputation cells rather than relying solely on expert opinion, as is often done.

References

Freeman, D. H. Jr. (1988), "Sample Survey Analysis: Analysis of Variance and Contingency Tables," Handbook of Statistics 6, eds. P.R. Krishnaiah and C.R. Rao, New York: North Holland, 415-426.

Kalton, G. and Kasprzyk, D. (1986), "The Treatment of Missing Data," Survey Methodology, 12:1, 1-16.

Koch, G. G., Freeman, D. H. Jr. and Freeman, J. L. (1975), "Strategies in the Multivariate Analysis of Data from Complex Surveys," Int. Stat. Rev., 43:1, 59-78.

Rao, J. N. K., Wu, C. F. J. and Yue, K. (1992), "Some Recent Work on Resampling Methods for Complex Surveys," Survey Methodology, 18:2, 209-217.

Searle, S. R. (1971), Linear Models, New York: John Wiley and Sons, p. 240.

Shah, Babubhai V., Holt, M. M. and Folsom, R. E. (1977), "Inference About Regression Models From Sample Survey Data," Bulletin of the International Statistical Institute, 41:3, 43-57.

Skinner, C. J., Holt, D. and Smith, T. M. F. (1989), Analysis of Complex Surveys, New York: John Wiley and Sons.

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:

<u>Region</u>

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.