United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Research Division

SRB Research Report
Number SRB-92-05

September 1992

EVALUATION OF THE CRO
PROCEDURES FOR THE QU
AGRICULTURAL SURVEY

Fatu Wesley

[illegible] THE CROP IMPUTATION PROCEDURES FOR THE QUARTERLY [illegible]AL SURVEY, by [illegible] Wesley, Research Division, National Agricultural Statistics [illegible]ment of Agriculture, Washington, D.C. 20250, September 1992. Research [illegible]

[illegible]STRACT

[illegible] Agricultural Statistics Service used an item imputation procedure for nonrespondents [illegible] Quarterly Agricultural Surveys. This study evaluates the accuracy of the item imputation [illegible] corn, soybeans and total cropland. The study used data from the 1990 June [illegible] Survey conducted in Indiana and Ohio in which a subsample of list frame [illegible] to provide their crop acreage. The difference between the values imputed [illegible] and the values they provided during the reinterview were used to estimate [illegible] survey indications for total cropland and corn were biased upward 3-4 [illegible] by state ($\alpha = .10$). Biases for soybeans were not significant. At the two [illegible] of nonrespondents with unknown agricultural operation status and [illegible] has the largest bias, which is significant ($\alpha = .10$) for cropland and corn.

****

[illegible] paper is prepared for limited distribution to the research *
[illegible] community outside the U.S. Department of Agriculture. The *
[illegible] views expressed herein are not necessarily those of NASS or USDA. *
*

****

ACKNOWLEDGEMENTS

[illegible] to thank staff and field enumerators of the Indiana and Ohio State Statistical [illegible] project possible. [illegible] provided by George Hanuschak and William Iwig [illegible] also thanks [illegible] for contributing to the nonresponse domain [illegible] for providing information on the operational imputation procedures.

September 1992

TABLE OF CONTENTS

SUMMARY iii

INTRODUCTION. 1

IMPUTATION PROCEDURE 2

General Overview 2

DAF Imputation Procedure 3

Total Cropland Imputation Procedure 5

Individual Crop Acreage Imputation Procedure 6

EVALUATION METHODS 7

RESULTS 9

Evaluation of Nonresponse Bias 9

Partitioning the Bias 11

CONCLUSIONS 13

RECOMMENDATIONS 13

REFERENCES 14

SUMMARY

The National Agricultural Statistics Service uses an item imputation procedure for nonrespondents in the Quarterly Agricultural Surveys. This research study evaluates the accuracy of the item imputation procedures used for corn, soybeans and total cropland. The study used data from the 1990 June Reinterview Survey conducted in Indiana and Ohio in which a subsample of list frame nonrespondents were asked to provide their crop acreage. Very large operations (strata 81-98) were excluded, in order to minimize an already high response burden for this group. The difference between the values imputed for the nonrespondents and the values they provided during the reinterview were used to estimate the imputation bias. The reinterview values were assumed to represent the true values.

The analysis showed that the cropland and corn estimates for the sampled list strata has an upward bias due to imputation of 3-4 percent for each state and for the states combined. All three bias estimates for cropland and corn were significant at $\alpha=.10$. Soybean biases were not significant in any category.

The bias was partitioned to nonrespondent subdomains, at the two state level, in order to identify the cause or source of the bias. Analyses indicated that 50% of the cropland bias, 43% of the corn bias, and 77% of the soybean bias was from nonrespondents with unknown agricultural operation and unknown cropland status. The bias for this subdomain was significant for cropland and corn ($\alpha=.10$), but not for soybeans.

EVALUATION OF THE CROP IMPUTATION PROCEDURES FOR THE QUARTERLY AGRICULTURAL SURVEY

Fatu Wesley

INTRODUCTION

The National Agricultural Statistics Service (NASS) conducts a Quarterly Agricultural Survey designed to provide indications of crop acreage, amount of grain stored on farms and total hogs at the state and U.S. levels (USDA 1990). Estimates of the number of corn and soybeans acres planted are important because of the relative sizes of the crops in the United States. An important item collected in the surveys is the total cropland contained in each selected farming operation. Although official estimates on total cropland are not published by NASS, this item is used to evaluate individual crop acreage indications and to impute individual crop acreage values for nonrespondents. Also, in most nonresponse situations, total cropland must be imputed before individual crop acreage values can be imputed. This paper reports results of research conducted to evaluate the accuracy of the item imputation procedure used by NASS for corn, soybeans and total cropland. "Truth" data were collected in the 1990 June Reinterview Survey (JRS) for a subsample of the sample units that were nonrespondents in June Agricultural Survey (JAS) in Indiana and Ohio. These data allowed examination of the bias due to imputation under the assumption that the reinterview values represent the true values.

The Quarterly Agricultural Survey is a multiple frame (MF) survey which consists of sample units from the list frame and from an area frame. The area frame provides complete coverage of the farm population, but does not provide the required precision, so the MF survey relies primarily on the sample units from the list frame and uses area frame units to account for the incompleteness of the list. A stratified list sample is selected for each quarterly survey where the stratification is based on historic control data for items of interest such as cropland, grain storage capacity and total hogs.

The June Reinterview Survey was conducted in Indiana and Ohio and included only sample units (farms) in the list frame. The June Agricultural Survey was conducted primarily by telephone and the reinterview was conducted in person. For reinterview purposes, units in the list frame of each state were divided into three domains. Domain 1 consisted of units that responded to the original survey. Domain 2 consisted of units that refused to cooperate, and domain 3 were units that could not be reached by an enumerator for an interview. Strata for very large operators (81-98) were excluded from the reinterview study because of the desire to minimize an already high response burden for this group. Domain 1 reinterview responses were used to evaluate the effect of questionnaire changes on the response (Warren 1991, Warren 1992). Data collected from domains 2 and 3 were used for this analysis to evaluate the crop imputation procedures.

IMPUTATION PROCEDURE

General Overview

NASS's imputation procedure (for total cropland, individual crop acreage and grain stocks) was implemented in 1987. The procedure was chosen because of its 1) generality, 2) maximum use of available information, 3) affordability and 4) availability for immediate implementation (Atkinson 1987). The imputation method for crop acreage uses a ratio estimator based on control (historical) and current cropland information. The method also uses two types of supplementary information that are collected about a nonrespondent. This information is the nonrespondent's agricultural operation (ag-op) status and cropland status. Possible agricultural operation and cropland status categories are presented in Tables 1 and 2. How the categories affect the imputation will be discussed in the next section. Any sampled unit that does not have cropland is treated as a valid useable zero record for all crop acreage items, even if the unit is a nonrespondent for grain stocks and livestock. Consequently, there is not a "zero" cropland status code.

Table 1. Agricultural Operation Status Codes

Type	**Description**
F, Type K	Know operation is a farm (F) and type of operation (partnership, etc.) is also known (K).
F, Type UK	Know operation is a farm (F) but type of operation is unknown (UK).
UK	It is not known if operation is a farm (UK).

Table 2. Cropland Status Codes

Type	Description
Positive	Operation is known to have cropland.
Unknown	It is unknown if the operation has cropland.

The primary imputation cell is the stratum/Agricultural Statistics District (ASD). A stratum for NASS agricultural surveys identifies farms within a state that are similar in size for some given commodity. ASD's refer to geographic areas within a state made up of multiple counties. Agricultural practices within an ASD are usually more homogenous than between ASDs. If a cell is empty, cells are collapsed according to a set priority scheme for the imputation.

DAF Imputation Procedure

The NASS estimation method uses a data adjustment factor (DAF) to adjust reported and imputed values. This adjustment for imputed records is:

$$Y_{(I)ci} = DAF_{(I)ci} * X_{(I)ci} \tag{1}$$

where:

$Y_{(I)ci}$ = the imputed summary crop acreage value for cell c, unit i,

$DAF_{(I)ci}$ = the data adjustment factor for cell c, unit i, and

$X_{(I)ci}$ = the imputed crop acreage value (either total cropland or a specific crop acreage) for cell c, unit i.

The imputation algorithm imputes $DAF_{(I)ci}$ and $X_{(I)ci}$ separately. The summary crop acreage value, $Y_{(I)ci}$, represents the imputed value that is summarized.

For respondents, the DAF is used to 1) adjust for duplicity in the list frame and 2) assign a value of zero to records of list frame units that do not have an agricultural operation. In the first case, the DAF adjusts for differences between the type of unit that is selected from the list frame and the type that actually exists. For example, if a sample unit is listed on the frame as an individual operation but it is found to be a partnership during the survey, then the DAF would be assigned a value less than 1 if the partner(s) also had a chance to report for the partnership. This same strategy is also used to adjust for actual duplication that is detected on the list frame. For all situations where the reporting unit is a farm and is the same as the selected unit, and no duplication exists, the DAF is 1. If a unit is coded as no longer farming, the DAF is assigned a value of zero for the unit.

In the case of nonresponse, the DAF is imputed for a nonrespondent with ag-op status that is partially (F,Type UK) or completely (UK) unknown. The actual DAF (DAF_{ci}) is used for nonrespondents whc are known to farm and whose type of operation is also known (F,Type K). If the nonrespondent's ag-op status is partially unknown (F,Type UK) then:

$$\mathrm{DAF}_{(I)ci} = \sum_{i \in f(c)} DAF_{ci}/n_{(f)c} \qquad (2)$$

$$= \mathrm{DAF}_{(f)c}$$

where:

(f)c = set of all respondents in cell c that have an agricultural operation.

When the nonrespondent's ag-op status is completely unknown, (UK), then:

$$\mathrm{DAF}_{(I)ci} = \sum_{i \in r(c)} DAF_{ci}/n_{(r)c} \qquad (3)$$

$$= \mathrm{DAF}_{(r)c}$$

where:

(r)c = set of all respondents in cell c.

In (2), since these units are known to have an agricultural operation, the DAF reflects the average duplicity adjustment that should be applied. In (3), the imputed DAF reflects both average duplicity and the average agricultural operation status.

Total Cropland Imputation Procedure

The form of the estimator used to impute the number of total crop acres depends on whether or not the unit's cropland status is known. If it is believed that nonrespondent i has cropland, then the imputed total cropland value is:

$$TC_{(I)ci} \quad = \quad C_{ci} * (TC_{(r+)c}/C_{(r+)c}) \qquad (4)$$

otherwise:

$$TC_{(I)ci} \quad = \quad C_{ci} * (TC_{(r)c}/C_{(r)c}) \qquad (5)$$

where:

C_{ci} = the control total crop acres for nonrespondent i in cell c,

$C_{(r+)c}$ = sum of control total crop acres for positive respondents in cell c,

$C_{(r)c}$ = sum of control total crop acres for all respondents in cell c,

$TC_{(r+)c}$ = sum of reported total crop acres for positive respondents in cell c, and

$TC_{(r)c}$ = sum of reported total crop acres for all respondents in cell c.

The imputed summary value, $Y_{(I)ci}$ in (1) for total cropland, is determined by the nonrespondent's category in Table 3. DAF_{ci} refers to the actual (not imputed) DAF. The bottom left cell in the table is empty since a unit with positive cropland status cannot have unknown ag-op status; that is, if the nonrespondent's cropland status is positive the operation must be agricultural.

Table 3. Total Cropland Imputation Formulas by Nonrespondent's Agricultural Operation and Cropland Status

Agricultural Status	**Cropland Status**	
	Positive	**Unknown**
F, Type K	$DAF_{ci}(C_{ci}*(TC_{(r+)c}/C_{(r+)c}))$	$DAF_{ci}(C_{ci}*(TC_{(r)c}/C_{(r)c}))$
F, Type UK	$DAF_{(f)c}(C_{ci}*(TC_{(r+)c}/C_{(r+)c}))$	$DAF_{(f)c}(C_{ci}*(TC_{(r)c}/C_{(r)c}))$
UK		$DAF_{(r)c}(C_{ci}*(TC_{(r)c}/C_{(r)c}))$

Total cropland is used in the imputation of specific crop acreage. If total cropland is not available, it is imputed and is then used in the formula that imputes the individual crop acreage.

Individual Crop Acreage Imputation Procedure

If the section in the questionnaire that has specific crop questions is coded as not being usable, then the individual crop acres are imputed. If the section is usable, then the acreage entered for each crop has to be positive or zero, and cannot be missing; therefore imputation is not performed when the section is usable. The imputation formula for nonrespondents i is:

$$SC_{(I)ci} = TC_{(\bullet)ci} * (SC_{(r)c}/TC_{(r)c}) \qquad (6)$$

where:

$TC_{(\bullet)ci}$ = the number of total cropland acres if known or $TC_{(I)ci}$ if not, for unit i in cell c,

$SC_{(r)c}$ = sum of the specific crop acres for all respondents in cell c, and

$TC_{(r)c}$ = sum of total cropland acres for all respondents in cell c.

Table 4. Specific Crop Imputation Formulas by Nonrespondent's Agricultural Operation and Cropland Status

Agricultural Status	Cropland Status	
	Positive	**Unknown**
F, Type K	$DAF_{ci}(TC_{(r+)ci} * (SC_{(r)c}/TC_{(r)c}))$	$DAF_{ci}(TC_{(r)ci} * (SC_{(r)c}/TC_{(r)c}))$
F, Type UK	$DAF_{(f)c}(TC_{(r+)ci} * (SC_{(r)c}/TC_{(r)c}))$	$DAF_{(f)c}(TC_{(r)ci} * (SC_{(r)c}/TC_{(r)c}))$
UK		$DAF_{(r)c}(TC_{(r)ci} * (SC_{(r)c}/TC_{(r)c}))$

EVALUATION METHODS

To evaluate the imputation procedure, the nonresponse bias, $\hat{B}$, and variance, $V(\hat{B})$, were estimated using a domain estimation technique. This method provides inferences regarding the size and significance of the nonresponse bias for the strata sample. Nonresponse bias refers to the amount of bias due to the imputation procedures. Domain estimation procedures were used since the objective is to estimate the amount of nonresponse bias in the total estimate due to the domain of nonrespondents in the population.

The nonresponse bias for crop acreage in stratum h is defined as:

$$B_h = N'_h \bar{D}_h$$

where:

N'_h = the population of nonrespondents (includes nonsampled units who would have been nonrespondents), and

$\bar{D}_h$ = the mean difference between the imputed and reinterview values for the population of nonrespondents.

The value B_h is estimated by:

$$\hat{B}_h = N_h(n'_h/n_h)\bar{d}_h \qquad (7)$$

where:

N_h = the population size (respondents and nonrespondents); N_h is known,

n_h = original sample size,

n'_h = set of nonrespondents from the original survey, and

$\bar{d}_h$ = the mean difference between the reinterview and imputed values for nonrespondents who responded to the reinterview survey.

$\hat{B}_h$ is shown to be unbiased with the following assumption:

$$\begin{aligned} E(\hat{B}_h) &= E(N_h(n'_h/n_h)\bar{d}_h) \\ &= N_h\ E(n'_h/n_h)E(\bar{d}_h \,|\, (n'_h/n_h)) \\ &= N_h P_h \bar{D}_h, \text{ with the assumption that } E(\bar{d}_h \,|\, (n'_h/n_h)) \approx \bar{D}_h) \\ &= N'_h \bar{D}_h. \end{aligned}$$

The assumption that $E(\bar{d}_h \,|\, (n'_h/n_h)) \approx \bar{D}_h$ was made because $\bar{d}_h$ is not expected to vary with the nonresponse rate in a specific survey.

Since $\hat{B}_h$ involves the product of two random variables ($\hat{P}_h = n'_h/n_h$, the estimated proportion of nonrespondents, and $\bar{d}_h$), the variance is expressed as follows:

$$\begin{aligned} V(\hat{B}_h) &= N_h^2((\hat{P}_h)^2 V(\bar{d}_h) + \ \bar{d}_h^2 V(\hat{P}_h)) \\ &= N_h^2((n'_h/n_h)^2 V(\bar{d}_h) + \ \bar{d}_h^2 V(n'_h/n_h)). \qquad (8) \end{aligned}$$

Nonresponse bias and variance for the sampled strata are obtained by summing $\hat{B}_h$ and $V(\hat{B}_h)$ across strata. $\hat{B}_h$ was estimated both within state and across states.

Besides comparing the results by state, subdomains of nonrespondents were also examined to determine the source of the bias. The subdomains used correspond to the nonrespondent's ag-op and cropland statuses given in Tables 3 and 4.

Estimates of the nonresponse bias for stratum h, subdomain j, ($\hat{B}_{hj}$), were obtained as follows:

$$\hat{B}_{hj} = \hat{N}'_h \sum_j d_{hji} / n''_h \qquad (9)$$

where:

$\hat{N}'_h$ = the estimated population number of nonrespondents

= $N_h\hat{P}_h$,

d_{hji} = the difference between the reinterview and imputed value if unit i is in stratum h, subdomain j (0 otherwise), and

n''_h = number of nonrespondents who responded to the reinterview survey.

The variance of $\hat{B}_{hj}$ was estimated using (8) with domain parameters replaced by subdomain parameters.

RESULTS

Evaluation of Nonresponse Bias

The imputation procedure was evaluated by examining the total nonresponse bias obtained from the domain and subdomain estimation methods (7)-(9). In the comparisons,

d_{hi} = $Y_{(I)hi} - R_{hi}$

where:

R_{hi} = Reinterview DAF_i * reinterview crop $acreage_i$, and

$Y_{(I)hi}$ = imputed value as specified in Table 3 or Table 4.

The summary values, $Y_{(I)hi}$ are imputed values based on the original JAS reported data. It was not feasible to impute based on reinterview data due to the small sample size. About 50% of the original nonrespondents who were contacted for the reinterview survey responded; their counts are given in Table 5. Values of $\hat{B}$, the total nonresponse bias of the sampled list strata, are also given in Table 5. The table shows that the biases for corn and total cropland range from three to four percent of the acreage estimated from the June Agricultural Survey's list frame for those strata, and are significant ($\alpha = .10$). None of the biases estimated for soybeans are significant.

Table 5. Nonresponse Bias (for Sample Strata) by State and Crop

State	Crop	# Obs.	$\hat{B}$ (acres)	% List Estimate	P Value
Indiana		162			
	Corn		*168,550	3.12	.077
	Soybeans		87,754	2.22	.270
	Total Cropland		*468,913	4.07	.010
Ohio		98			
	Corn		*144,252	4.08	.064
	Soybeans		13,220	.43	.843
	Total Cropland		*295,409	3.00	.080
Both		260			
	Corn		*312,802	3.61	.011
	Soybeans		100,974	1.43	.330
	Total Cropland		*764,321	3.57	.002

* Indicates bias is significant.

The nonsignificant bias for soybeans could be due to the late planting of soybeans because of the wet weather in 1990, and possibly the Ohio farmers did not fully report their intentions as of June 1, but did include later plantings in their reinterview report. The Ohio JAS expanded acreage was about 6 percent lower than the reinterview acreage (Warren, 1991). These low initial reported soybean acres could put some "pressure" on the imputed values to be lower, offsetting any upward imputation bias that would be expected based on the cropland and corn results.

Partitioning the Bias

Preceding sections showed how the ag-op status and the cropland status of a nonrespondent determines the imputed values. Partitioning the total bias to these subdomains (given in Tables 3 and 4) provides direction in identifying the true cause or source of the bias.

Table 6 shows that, for total cropland, the bias in the two subdomains where the ag-op status is unknown (partially or completely) and where the cropland status is also unknown is 3% (1.22% + 1.78%) of the list estimate for the strata included. The total bias is 3.57% of the list estimate (Table 5), so these two subdomains account for 84% of the total bias. Separate analysis indicates they include 70% of the nonrespondents in the list strata included in the study. So the size of the bias is explained to a large degree by the number of records. The subdomain where the ag-op status is partially unknown and the cropland status unknown accounts for 34 percent of the bias. Fifty percent of the bias is accounted for by the subdomain where both the ag-op and cropland status are unknown (hereafter referred to as the "UK" subdomain). Bias for this subdomain is significant (p=.03).

Table 6. Nonresponse Bias for Total Cropland by Nonrespondent's Agricultural Operation and Cropland Status

Agricultural Status	**Cropland Status**			
	Positive		**Unknown**	
	$\hat{B}$ (acres)	**% List Estimates**	**$\hat{B}$ (acres)**	**% List Estimates**
IB, Type K	-2,692	-.01	54,489	.32
IB, Type UK	73,746	.34	61,115	1.22
UK			*380,663	1.78

* Indicates that the bias is significant at α=.05 (p value=.03).

Table 7. Nonresponse Bias for Corn Acreage by Nonrespondent's Agricultural Operation and Cropland Status

Agricultural Status	Cropland Status			
	Positive		Unknown	
	$\hat{B}$ (acres)	% List Estimate	$\hat{B}$ (acres)	% List Estimate
F, Type K	4,529	.05	27,625	.32
F, Type UK	29,156	.34	116,909	1.35
UK			*134,581	1.55

*Indicates that the bias is significant at $\alpha = .10$ (p value=.097).

Table 8. Nonresponse Bias for Soybean Acreage Based by Nonrespondent's Agricultural Operation and Cropland Status

Agricultural Status	Cropland Status			
	Positive		Unknown	
	$\hat{B}$ (acres)	% List Estimate	$\hat{B}$ (acres)	% List Estimate
F, Type K	-2,266	-.03	13,580	.19
F, Type UK	2,209	.03	9,759	.14
UK			77,691	1.10

Tables 7 and 8 show the results of partitioning the bias for corn and soybean acreage imputations, respectively. The bias was partitioned by the nonrespondent's ag-op and cropland status, as given in Table 4. The tables show that the bias is not significant for soybeans in any subdomain. The bias is significant for corn in the "UK" subdomain. Although the soybean biases are not significantly different from zero, the "UK" subdomain contains almost all of the soybean bias (77 percent). Again, we see the "unknowns" as being a major factor in the total nonresponse bias.

CONCLUSIONS

Estimates of nonreponse bias in total cropland, corn and soybeans were obtained for the 1990 June Agricultural Survey in Indiana and Ohio. The total cropland and corn bias estimates were between 3-4 percent for each state and for the states combined. The cropland and corn bias estimates were significant at $\alpha=.10$. Soybean biases were not significant in any category.

Each bias was partitioned to nonrespondent subdomains at the two state level in order to identify the cause or source of the bias. Analyses indicated that 50% of the total cropland bias, 43% of the corn bias, and 77% of the soybean bias was from the "UK" subdomain.

A major conclusion of the study is that nonrespondents with unknown agricultural operation status and unknown cropland status are the significant contributors to the nonresponse bias. The bias may be due to DAF imputation, cropland imputation, and/or to an interaction between the two factors. Future study will identify the portion of the bias due to each factor. In addition the causes of the bias will be examined. The imputed DAF may be biased upwards due to a greater percentage of the nonrespondents being out of business than of the respondents. This situation may also affect the imputed cropland and interaction values. Poor quality control data may also contribute to the cropland imputation bias.

RECOMMENDATIONS

The results of analyses of the 1990 June Reinterview Survey indicate that the crop imputation procedures for total cropland and corn are biased. Further analytic study needs to be conducted to determine the cause or causes of the potential bias and if the bias can be reduced or eliminated. Other alternative imputation procedures, including the approach suggested by Kott (1990), which modifies the current procedure, should be evaluated based on cost efficiency, ease of implementation, and bias.

REFERENCES

Atkinson, Dale (1987), The Scope and Effect of Imputation in Quarterly Surveys, U.S. Department of Agriculture, National Agricultural Statistics Service, Washington, D.C.

Kott, Philip (1990), Nonresponse Adjustments in NASS Agricultural Surveys, U.S. Department of Agriculture, National Agricultural Statistics Service, Washington, D.C.

U.S. Department of Agriculture (1990), June Agricultural Survey: Enumerator's Manual, National Agricultural and Statistics Service, Washington, D.C.

Warren, Fred (1991), June 1990 Reinterview Survey: Part I, Effect of Alternative Acreage Operated Questions on Reported Acreage and Number of Hogs on Farms, U.S Department of Agriculture, National Agricultural Statistics Service, Washington, D.C.

Warren, Fred (1992), June 1990 Reinterview Survey: Part II, Effect of Parcel and Field Level Acreage Questions, U.S. Department of Agriculture, National Agricultural Statistics Service, Washington, D.C.

U.S. Government Printing Office [illegible] 60770