EVALUATION OF AN ITEM IMPUTATION PROCEDURE FOR AGRICULTURAL SURVEYS

Fatu Wesley
U.S. Department of Agriculture
Room 4801 S. Bldg, 14th & Indep., S.W., Washington, D.C. 20250

KEY WORDS: Cropland, nonresponse, bias

1. INTRODUCTION

The National Agricultural Statistics Service (NASS) conducts a Quarterly Agricultural Survey designed to provide indications of crop acreages, amount of grain stored on farms and total hogs at the state and U.S. levels [2,3]. An important item collected on the surveys is the total cropland contained in each selected farming operation. Although official estimates on 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. However, 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 total cropland. 'Truth' data were collected in the 1990 June Reinterview Survey for a subsample of the sample units that were nonrespondents in the June Agricultural Survey 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 is a multiple frame 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 (JRS) was conducted in Indiana and Ohio and included only sample units(farms) in the list frame. The June Agricultural Survey (JAS) 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 were excluded from the reinterview study because of the desire to mininmize an already high response burden for this group. Domain 1 reinterview responses were used to evaluate the accuracy of the original response. Data collected from domains 2 and 3 were used for this analysis to evaluate the cropland imputation procedure.

2. IMPUTATION METHOD

2.1 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 [1]. The imputation method 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 business status and cropland status. Possible business 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. Business status codes

Type	Description
1. IB, Type K	Operation is known to be in business (IB) and type of of operation (partnership, etc.) is also known (K).
2. IB, Type UK	Operation is known to be in business (IB) but type of operation is unknown (UK).
3. UK	It is not known whether the in business (UK).

Table 2. Cropland Status Codes

Type	Description
1. Positive	Operation is known to have cropland.
2. UK	It is unknown whether the the operation has cropland.

2.2 Cropland Imputation Procedure

The NASS estimation methodology uses a list adjustment factor (LAF) to adjust reported values. This adjustment, in terms of cropland, is as follows.

Let $Y_i = LAF_i * X_i$ (1)

Where,

Y_i = the survey cropland value for unit i
LAF_i= the list adjustment factor for unit i
X_i= the actual reported cropland value for unit i

For **respondents**, the LAF 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 are out of business. In the first case, the LAF 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 LAF 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 in business and is the same as the selected unit, the LAF is 1. If a unit is coded as being out of business, the LAF is assigned a value of zero for the unit.

In the case of nonresponse, the LAF is imputed for a nonrespondent with business status that is partially (IB, Type UK) or completely (UK) unknown. The actual LAF is used for nonrespondents who are known to be in business and whose type of business operation is also known(IB, Type K).

If the nonrespondent's business status is partially unknown, (IB, Type UK), then:

$$LAF_{IB} = \Sigma_i \; LAF_{i \epsilon IB} / n_{IB} \qquad (2)$$

where,

IB = set of all "in business" respondents

When the nonrespondent's business status is completely unknown,(UK), then:

$$LAF_R = \Sigma_i \; LAF_{i \epsilon R} / n_R \qquad (3)$$

where,
R = set of all respondents

In (2), since these units are known to be in business, the imputed LAF reflects the average duplicity adjustment that should be applied. In (3), the imputed LAF reflects both average duplicity and the average business status.

The form of the estimator used to impute the number of cropland acres depends on whether or not the unit's cropland status is known. If it is believed that the nonrespon-espondent has cropland then:

$$Y_{R+} = C_i * (S_{R+}/C_{R+}) \qquad (4)$$

otherwise,

$$Y_R = (Ci *(S_R/C_R) \qquad (5)$$

where,

C_i = the control cropland acres for the ith nonrespondent
C_{R+}= sum of control cropland acres for positive respondents
C_R= sum of control cropland acres for all respondents
S_{R+}= sum of reported cropland acres for positive respondents
S_R= sum of reported cropland acres for all respondents

The primary imputation cell is the stratum/agricultural statistic 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 collasped according to a set priority scheme for the imputation.

In summary, the imputed survey value for cropland is determined by the nonrespondent's category in Table 3. The bottom left cell in the table is empty since a unit with positive cropland status cannot have unknown business status; that is, if the nonrespondent's cropland status is positive the operation must be in business (IB).

Table 3. Imputation formulas by type of norespondent.

Business Status	Cropland Status	
	Positive	Unknown
IB, Type K	$LAF_{i*}(C_i*(S_{R+}/C_{R+}))$	$LAF_{i*}(C_i*(S_R/C_R))$
IB, Type UK	$LAF_{IB}(C_i*(S_{R+}/C_{R+}))$	$LAF_{IB}(C_i*(S_R/C_R))$
UK		$LAF_R(C_i*(S_R/C_R))$

3. 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 sampled. 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.

The nonresponse bias for total cropland in stratum h is defined as:

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

where,

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

In this study, B_h is estimated by:

$$\hat{B}_h = N_h(n'_h/n_h)\bar{d}_h \qquad (6)$$

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
$\bar{d}_h$ = 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:

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

The assumption that $E(\bar{d}_h|(\underline{n}'_h/n_h)) \approx \bar{D}_h)$ was made because 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 d_h), the variance is expressed as follows:

$$V(\hat{B}_h) = N_h^2((\hat{P}_h)^2V(\bar{d}_h) + \bar{d}\,^2V(\hat{P}_h))$$
$$= N_h^2((n'_h/n_h)^2V(\bar{d}_h) + \bar{d}\,^2V(n'_h/n_h))$$

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 nonresrespondents were also examined to evaluate the source of the bias. The subdomains were based on nonrespondent cropland and business status categories given in Table 3.

Estimates of the nonresponse bias for subdomain j, $(\hat{B}_{hj})$, were

obtained as follows:
Let

$$\hat{B}_{hj} = \hat{N}'_h \sum_i (d_{hij})/n''_h \qquad (7)$$

where,

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

d_{hij} = d_{hi}, if unit i is in subdomain j
0, otherwise

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

4. RESULTS

4.1 Evaluation of Nonresponse Bias

The imputation procedure was evaluated by examining the total nonresponse bias obtained from the domain and subdomain estimation methods (6)-(7). In the comparisons,

let $d_{hi} = x_{hi} - r_{hi}$

Where,
r_{hi} = reinterview LAF_i * reinterview cropland_i;
x_{hi} = imputed value as specified in Table 3.

About 50% of the original nonrespondents who were contacted for the reinterview survey responded. The imputed values (x_{hi}) for these units were obtained from NASS's Estimates Division and were based on data from the June Agricultural Survey (original survey). Values of $\hat{B}$, the total nonresponse bias of the sampled list strata, are given in Table 4. The table shows that the biases range from three to four percent of the total cropland estimated from the June Agricultural Survey's list frame for those strata. The p values for the bias estimates are .08 (Ohio), .01(Indiana) and .002 (both states). Note that the three estimates are significant at α=.10 and that Indiana and the states combined are significant at α=.05.

Table 4. Total nonresponse bias (for sampled strata) by state and across states

State	# Obs	$\hat{B}$ (acres)	% List Estimate	P Value
Indiana	162	468,913	4.07	.01
Ohio	98	295,409	3.00	.08
Both	260	764,321	3.57	.002

4.2 Partitioning of The Bias

Preceding sections showed how the "business,cropland" category of a nonrespondent determines the imputed values. Partioning the total bias to these sub-domain categories (given in Table 3) provides direction in indentifying the true cause or source of the bias. The results of the subdomain analysis are discussed below.

Table 5 shows that most of the nonresponse bias is due to the two subdomains where the business status is unknown (partially or completely) and where the cropland status is also unknown. The subdomain where the business status is partially unknown and the cropland status is unknown accounts for 34 percent of the bias. Fifty percent of the bias is accounted for by the subdomain where both the business and cropland status are unknown. Bias for the latter subdomain is significant(p=.03). This bias may be due to LAF 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. Separate analysis indicates that the imputed LAF is biased upwards probably 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.

Table 5. Nonresponse bias based on nonrespondent business, cropland) status.

	Cropland Status			
	Positive		Unknown	
Business Status	$\hat{B}$ (acres)	% List Estimate	$\hat{B}$ (acres)	% List Estimate
IB, Type K	-2,692	-.01	51,489	.24
IB, Type UK	73,746	.34	261,115	1.22
UK			380,663*	1.78

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

5.0 SUMMARY

Estimates of nonreponse bias in total cropland were obtained for the 1990 June Agricultural Survey in Indiana and Ohio. The bias estimates were 4 percent for Indiana, 3 percent for Ohio and 3.5 percent for the states combined. All three bias estimates were significant at α=.10.

The bias was partitioned to nonrespondent subdomain categories in order to identify the cause or source of the bias. Analyses indicated that 50% of the bias was from the "UK business status, UK cropland status" category and that the bias was significant at α=.05.

Future study will examine the portion of the bias due to LAF imputation, cropland imputation and to the interaction. Causes of the bias will also be examined.

REFERENCES

1. Atkinson, Dale, The Scope and Effect of Imputation in Quarterly Surveys, U.S. Department of Agriculture, National Agricultural and Statisitics Service, Washington, D.C., 1988

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

Acknowledgements: The author thanks Ron Fecso of the National Agricultural Statistics Service for consultation on the nonresponse domain estimator and Don Allen, also of NASS, for reviewing the paper.