United States Department of Agriculture

National Agricultural Statistics Service

Research and Applications Division

SRB Staff Report Number SRB-90-11

June 1990

NONRESPONSE ADJUSTMENTS IN NASS AGRICULTURAL SURVEYS

Phillip S. Kott

NONRESPONSE ADJUSTMENTS IN NASS AGRICULTURAL SURVEYS, by Phillip S. Kott, Research and Applications Division, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, DC 20250, June 1990. NASS Staff Report, Number SRB-90-11.

ABSTRACT

The National Agricultural Statistics Service (NASS) currently uses two distinct adjustment methods to account for nonresponse on the list side of its Agricultural Surveys. A complex imputation program is employed for crop (acreage, production, and stock) items, while an adjusted form of reweighting is used for livestock items. This paper examines the subtle differences in the two approaches. A unified imputation/reweighting methodology for crop acreages and livestock is proposed and alternative specifications of the methodology are critiqued. Extensions to production, stocks, and the area side are also discussed.

KEY WORDS

Imputation, reweighting, model, stratum, agricultural statistics district.

ACKNOWLEDGMENTS

The author would like to thank Jack Nealon, John Witzig, Ron Bosecker, Bob Hale, Ned Jones, and Jerry Thorson for their helpful comments and advice. He would also like to thank Fred Warren for providing and explaining the data set analyzed in this report. Finally, Mike Bellow uncovered a number of minor errors in an earlier draft of this report. Any remaining errors are the responsibility of the author alone.

June 1990

Washington, DC

TABLE OF CONTENTS

SUMMARY

This paper addresses some of the issues involved in adjusting for nonresponse in NASS's Agricultural Surveys. The main focus is on list expansions for cropland and total hogs from the June 1989 survey since these topics are the simplest to address analytically. Nonresponse adjustments for other items and for area expansions are also discussed but in a less thorough manner.

Particular findings are:

1) The differences in the Summary 1 and Summary 2 list frame indications of total cropland in 11 states are analyzed. Note: Summary 1 essentially reweights for nonresponse at the stratum level, while Summary 2 uses a complex imputation program.

2) A unified imputation methodology for total cropland and total hogs on the list side is proposed. The current adjusted estimator for livestock is shown to be very close to a particular specification of this unified methodology.

3) The relative advantages and disadvantages of using control data and agricultural statistics district are discussed. It is demonstrated that using control data may be biasing cropland imputation slightly upward. On the other hand, the difference between imputed and real values appears to be minimized when both control data and agricultural statistics district are used.

The paper ultimately favors one of the following three alternatives:

1) The current imputation method should be modified in several subtle ways for crop acreage items on the list side. Adjusted livestock reweighting should be modified in conformal ways.

2) The imputation program should be modified to essentially impute missing values (except for production values) with the appropriate imputation cell means. In most cases, an imputation cell will be the intersection of a stratum and an agricultural statistics district. All missing livestock and crop items (except production) on the list side could be imputed this way.

3) A reweighting scheme very close to the current method of nonresponse adjustment for hogs should be incorporated into the SPS Summary System for all livestock and crop items on the list side except production. NASS should then begin using agricultural statistics district information in stratification.

There is also a brief discussion of stocks and a proposal for area side reweighting.

INTRODUCTION

The adjusted reweighting algorithm for handling nonresponse in livestock survey items can be shown to be equivalent to a particular method of imputation. It is not true, however, that the livestock imputation methodology contained within the current Bosecker-Atkinson Imputation Program (BAIP) is virtually identical to that method (note: the BAIP is used operationally by NASS to impute crop and stock items but not livestock items). This is because the BAIP does more than impute for missing survey values. It also imputes for missing list adjustment factors (LAF's). This tends to make livestock estimates based on BAIP imputation higher than livestock estimates calculated with the adjusted reweighting algorithm.

A unified method of imputation is proposed here that removes LAF imputation. Another revision is suggested that eliminates the distinction between record-by-record imputation (as performed by the BAIP) and selected-unit-by-selected-unit imputation (as implicitly "performed" by adjusted reweighting).

Currently, crop imputation treats the intersection of an agricultural statistics district and a list stratum as an imputation cell, while the livestock nonresponse adjustment inherently treats the stratum as the imputation cell (the unused BAIP does likewise for livestock items). Crops imputation makes use of control data; livestock reweighting does not. The advantages and disadvantages of these two different approaches are discussed in terms of bias, fit, and convenience.

There is a brief discussion of total land imputation, stock imputation, and nonresponse adjustments on the area side. Although it is dangerous to form general conclusions based on a limited empirical study, this paper suggests that it may not be unreasonable for NASS to adopt an imputation scheme very close to the current method of livestock nonresponse adjustment for all crop list frame expansions except production. An analogous scheme is also favored for area frame expansions. Alternatives are discussed, in particular modifying the imputation program so that it uses agricultural statistics district information but not control data.

SUMMARY 1 VERSUS SUMMARY 2

Let us briefly review how the Enumerative Summary System (ESS) handled missing survey values for list frame reporting units in the June 1989 Agricultural Survey. Before proceeding, a note on terminology is in order: a record is termed valid for a particular survey item if the item value (which is usually provided by the reporting unit but can be hand imputed by NASS) passes the Edit Program [1]; otherwise, the reporting unit is termed an item nonrespondent and the record termed missing for the item. A reporting unit can also be a survey nonrespondent if it totally refuses to participate in the survey or is inaccessible (item code 910 equals 6 or greater). Note, however, if NASS hand imputes values for such a reporting unit (item code 910 equals 4), it is deemed a survey respondent.

The ESS produced two direct list expansions for every crop item. The Summary 1 expansion essentially reweighted for nonresponse at the stratum level. The expansion factor for a valid record from a given list stratum was set equal to the total number of names in the stratum population divided by the number of names corresponding to valid records (Technical note: Summary 1 used item code 138 exclusively to determine the vality of records; when a particular survey item value was missing for a reporting unit with an item code 138 equaling 0 or 3, the missing value was imputed as described below.)

The situation was more complicated when a selected name resulted in two records, one valid and one not, but we will not discuss that situation in detail here.

The Summary 2 direct expansion made use of imputed values for missing survey items. The expansion factor for such values was the original list stratum expansion factor (the total number of stratum names divided by the number of sampled names). Naturally, the same original expansion factor was also applied to valid records from the stratum.

The key to understanding Summary 2, of course, is the Bosecker-Atkinson Imputation Program (BAIP). A survey item value is the product of two factors - a raw item value and a list adjustment factor (LAF). It is possible for either value to be missing from an Edited Data File (if the LAF is missing so will be the raw item value but the converse is not necessarily true). The BAIP imputes for both missing LAF's and missing raw item values.

For survey respondents, the vast majority of LAF's on the June 1989 Edited Data File were either 0 or 1: 1 when the reporting unit (record) was coded in business; 0 when not. Missing LAF's for survey nonrespondents were imputed by a complicated process documented in [2]. For our purposes, it suffices to recognize that a survey nonrespondent coded as being in business (item code 921 equals 11) had an imputed LAF virtually equal to 1, while a nonrespondent with unknown business status (item code 921 equals 12) had an imputed LAF equal to what may be viewed as its probability of being in business.

Raw item values were also imputed by a complicated process documented in [2]. In this section, we focus on cropland imputation (item code 802). First, all reporting units were grouped into imputation cells. An imputation cell was the intersection of a list stratum and an agricultural statistics district, hereafter called a <u>stratum/district</u>. Most reporting units, whether cropland respondents or not, had MLSFCRP3 values (past computed cropland from an external source) on the June Edited Data File. For simplicity, let us assume all had such values. If a reporting unit did not respond to the cropland question but was known to have cropland (because item code 802 equaled -1 or item code 138 equaled 1), then a raw cropland value was imputed for it by multiplying its MLSFCRP3 by Σ CROPLAND / Σ MLSFCRP3, where both summations were over all reporting units <u>with positive CROPLAND values</u> in the the same imputation cell as the nonrespondent. If it was not known whether a cropland nonrespondent had cropland (item code 802 equaled 0 and item code 138 equaled 2), then a raw cropland value was imputed for it by multiplying its MLSFCRP3 by Σ CROPLAND / Σ MLSFCRP3, where both summations were now over all reporting units in the cell with known cropland values (either item code 802 was positive or code 802 equaled 0 while 138 equaled 3).

Imputation was more complicated when a cropland nonrespondent had a missing MLSFCRP3 value or when there were fewer than two item respondents in an imputation cell with positive cropland values. We ignore both these possibilities here.

If we also ignore the distinction between selected names from the list frame and reporting units for the June survey (and in the great majority of cases there is no distinction), then the Summary 1 approach to nonresponse can also be looked at as a type of imputation. Missing survey values (raw values times LAF's) were implicitly imputed by the stratum mean of the valid survey values.

Obviously, Summary 2 imputation is more sophisticated than Summary 1 "imputation" because it makes use of the following pieces of information about the cropland nonrespondent (when available):

1) whether it is in business,

2) whether it has crops,

3) its agricultural statistics district, and

4) control information thought to be related to cropland (the MLSFCRP3 value).

The unambiguous effect of the first two points listed above is to increase Summary 2 indications relative to Summary 1 indications. The reason for this is that Summary 1 implicitly estimates both the probability of a nonrespondent being in business and the probability of it having cropland. These estimates are (almost) always less than 1. On the other hand, if a cropland nonrespondent is coded as in business, then Summary 2 essentially replaces the Summary 1 implicit estimate of whether this reporting unit is in business by 1. A similar phenomenon holds true when the nonrespondent has crops. The effects of points 3 and 4 above are more ambiguous.

Table 1 compares indications resulting from five different methods of imputation. One method, essentially equivalent to BAIP imputation (see the Appendix for details), is the standard. The others are compared to it. The first column shows the percent reduction in the June 1989 cropland indications for 11 states that one would obtain if the cropland survey value (LAF times raw value) were imputed directly. This method, called the modified BAIP, ignores information on business status. More details on this method are given in the following section.

The second column of Table 1 shows the percent reduction in the indications from using an imputation method that ignores survey information about both business status and whether the cropland nonrespondent had cropland. The third column shows the percent reduction from using a imputation method virtually identical to Summary 1.

Table 1 shows that, as expected, the effects of removing "in business" and "has cropland" information are to decrease the indications. That happens in all 11 states. The combined effect of removing control and agricultural statistics district information is less clear. On the whole, it decreases the indication another 0.61% (= 3.71% - 3.10%). It has the opposite effect in three of the 11 states however. Moreover, the total effect is less than removing "in business" information (0.75%) and much less than removing "has cropland information" (2.35% = 3.10 - 0.75).

In the next section, it will be argued that the modfied BAIP is part of a unified imputation/reweighting methodology. The last question Table 1 addresses is the effect of using stratum/districts as the imputation cells rather than the strata themselves. The fourth column of the table shows how the indications would change if the strata served as the imputation cells in the modified BAIP. The effect, while large in some states (1.05% for Nebraska), is negligible (-0.07%) at the 11 state aggregate level.

NASS did not use the BAIP to impute for missing hog records in the June 1989 Agricultural Survey. The BAIP, however, can impute for such values. The program would treat the list strata as imputation cells and <u>not</u> use control data. In principle, then Summary 2-like indications could be determined for total hogs and pigs. That was done for the 11 states under study. Table 2 reports the percent reduction in the total hogs and pigs indications from Summary 2 that one would obtain if, first, the survey item values were imputed directly (i.e., "in business" information was not used), and then second, "has hogs" (item code 499 equals 1) information was not used.

The details of how these later two calculations were made are given in the following section. It will be shown that they are very close to the current "adjusted" and "operational" indications for total hogs and pigs.

Table 1. -- Comparing List Expansions for Total Cropland in 11 States Using a Variety of Imputation Methodologies

State	% Reduction from Sum 2* of using the modified BAIP**	% Reduction from Sum 2 of using the modified BAIP without "has cropland" info	% Reduction from Sum 2 of using Sum 1***	% Increase from the modified BAIP of using Strata as the Imputation Cells
Illinois	0.70	1.68	0.45	-0.16
Indiana	1.60	3.97	5.88	-0.11
Iowa	0.36	3.27	4.00	-0.52
Kansas	0.75	2.46	2.70	-0.55
Michigan	0.41	1.45	0.87	0.23
Minnesota	1.34	3.99	4.86	-0.09
Mississippi	0.71	2.42	1.96	0.23
Nebraska	0.00****	4.46	6.25	1.05
Ohio	1.36	6.18	8.33	-0.05
South Dakota	0.48	2.25	3.31	-0.20
Wisconsin	1.03	2.12	2.87	0.03
11 State Aggregate	0.75	3.10	3.71	-0.07

NOTE: These calculations were performed using the Imputed Data File. Consequently, only records with item code 138 equal to 0 or 3 were treated as valid.

* See the Appendix for details on how the BAIP was simulated for the Summary 2 numbers.

** See the following section for a description of the modified BAIP. (Note: the imputation cell was primarily the stratum/district; the collapsing rule for an imputation cell in the modified BAIP mirrored that of the BAIP.)

*** Summary 1 was simulated by reweighting all valid cropland values in the following manner: the original expansion factor was multiplied by the number of records in the stratum (valid or not) and divided by the number of valid records in the stratum. This is identical to the Summary 1 methodology when there is a one-to-one correspondence between selected names and survey records (see also "NOTE" above). It is a good approximation otherwise.

**** This number is positive buts rounds to zero.

Table 2. -- Comparing List Expansions for Total Hogs and Pigs in 11 States Using Different Imputation Methodologies

State	% Reduction from Sum 2* of using the modified BAIP**	% Reduction from Sum 2 of using the modified BAIP without "has hogs information
Illinois	0.65	4.63
Indiana	1.19	5.76
Iowa	0.55	6.29
Kansas	0.69	4.89
Michigan	0.41	3.67
Minnesota	0.90	5.69
Mississippi	0.94	3.31
Nebraska	0.01	9.18
Ohio	1.58	10.57
South Dakota	0.68	5.95
Wisconsin	0.40	3.30
11 State Aggregate	0.69	5.98

* See the Appendix for details on how the BAIP was simulated for the Summary 2 numbers.

** See the following section for a description of the modified BAIP. (Note: the imputation cell was the stratum.)

A UNIFIED FORMULATION OF IMPUTATION

We begin this section by describing the modified BAIP for cropland. Let **S** be the set of reporting units for a particular imputation cell, **R** the set of cropland respondents, $\mathbf{R}^+$ the set of item respondents with positive cropland, **K** the set of cropland nonrespondents known to have cropland (item code 802 equals -1 of item code 138 equals 1).

Let y_j be the survey value for cropland respondent j in the imputation cell. Suppose reporting unit i has a missing cropland value, then y_i is imputed by

$$y_i^* = \begin{cases} x_i B^+ & \text{when i is known to have cropland} \\ x_i P B^+ & \text{otherwise,} \end{cases} \qquad (1.1)$$

where x_i is the MLSFCRP3 value for unit i (which itself may be imputed; see the Appendix),

$$B^+ = \frac{\sum_{j \in R^+} y_j}{\sum_{j \in R^+} x_j} \text{, and} \qquad (1.2)$$

$$P = \frac{\sum_{j \in (R^+ \cup K)} x_j}{\sum_{j \in (R \cup K)} x_j} . \qquad (1.3)$$

As already noted, the BAIP imputes raw item values, not survey values. There is a more subtle distinction between the BAIP and the modified BAIP. The BAIP incorrectly fails to sum over the set K (item nonrespondents known to have cropland) when implicitly computing P -- the estimated probability of i having cropland. This odd method of estimating the probability of i having cropland was chosen so that if K were empty, $B = PB^+$ would have the standard form:

$$B = \Sigma_{i \in R}(y_i)/\Sigma_{i \in R}(x_i).$$

Suppose that the imputation cell in question is a stratum and that no selected unit (i.e., sampled name) in the stratum has more than one reporting unit associated with it. Now let y_j in the above equations denote reporting unit j's survey value for total hogs and pigs (item code 300), and assume all the x_i are equal to 1. Finally, suppose R, R^+, and K are redefined around the total hogs item and the "has hogs" question (item code 499). It is not difficult to show that imputing for missing hog item records with equation (1) accomplishes the same thing as adjusted reweighting.

If all the reporting units in a cell shared the same expansion factor, E, the direct expansion for that cell using the modifed BAIP (equations (1.1) through (1.3)) could be shown to equal

$$Y = E\ (\sum_{i \epsilon R} y_i + \sum_{i \epsilon N} y_i^*)$$

$$= E\ (\sum_{i \epsilon S} x_i) PB^+, \qquad (2.1)$$

$$= \{E\ (\sum_{j \epsilon S} x_j P / \sum_{j \epsilon R^+} x_j)\} \sum_{i \epsilon R^+} y_i$$

$$= \qquad E^* \qquad \sum_{i \epsilon R^+} y_i \ . \qquad (2.2)$$

where N is the set of item nonrespondents. Equation (2.1) is the form in which the adjusted reweighting estimator for livestock items is usually presented (see [3]). Remember, all the x_i equal 1 and the imputation cell is the stratum for livestock items, so that B^+ becomes simply the mean of the positives. Equation (2.2) reveals where the name "adjusted reweighting" comes from. The term E^* can be thought of as a reweighted expansion factor adjusted with "has hogs" (or "has cropland") information.

In the following sections, we will investigate alternative <u>specifications</u> of the unified imputation methodology expressed in equations (1.1) through (1.3), which we have called the modified BAIP. By specifications, we mean how the imputation cells are defined and what values are used for the x_i. The Edited Data File contains a variable MLSFHOGS that could be used for total hogs imputation like MLSFCRP3 is used for cropland.

Let us now turn to the case where one selected unit splits into several reporting units. The modified BAIP imputes at the record level just like the original BAIP. An additional step, described below, can be taken before the Program is run to make it more like the selected unit level "imputation" of livestock adjusted reweighting.

The control value for a record from a sampling unit that splits into m reporting units would be calculated by taking the original sampling unit's control value from the Master File (currently MLSFCRP3 for cropland; 1 for total hogs) and dividing by m. This is what was done in producing Tables 1 and 2. Combining this step and the modified BAIP will reproduce the adjusted estimator for livestock items exactly when there is no item nonresponse in reporting units from split sampling units. Otherwise, there will be a difference but it is likely to be very minor. For our 11 state composite, the difference between the actual adjusted estimator and the calculated modified BAIP for total hogs was 0.02%.

A column in Table 1 refers to the "modified BAIP without cropland info." This meant imputing missing y_i values by $x_i B$, where B is defined above as $\Sigma_{i \epsilon R}(y_i)/\Sigma_{i \epsilon R}(x_i)$. Likewise, a column in Table 2 refers to the "modified BAIP without hogs info," for which an analogous approach was taken.

THE THEORETICAL MODEL BEHIND IMPUTATION

The modified BAIP assumes that the total cropland (or total hogs) survey value for a reporting unit within a particular imputation cell can be thought of as generated from the following model:

$$y_i = \pi_i(\beta^+ x_i + e_i), \qquad (3)$$

where π_i and e_i are random varables independent of each other and across units, such that

$$\text{Prob}(\pi_i = 1) = \pi, \ \text{Prob}(\pi_i = 0) = 1 - \pi,$$
$$\text{and } E(e_i) = 0, \ \text{Var}(e_i) = \sigma_i^2.$$

Furthermore,

$$E(e_i | i \epsilon R) = E(e_i | i \epsilon N), \qquad (4.1)$$

while $E(\pi_i | i \epsilon R \cup K) = E(\pi_i | i \epsilon N - K)$. (4.2)

(Note: $i \epsilon N - K$ means $i \epsilon N$ but $i \epsilon K$.)

When this last pair of assumptions is met item nonresponse is said to be <u>ignorable</u> under the model in (3).

Observe that under the assumptions of the model, $E(B^+) = \beta^+$ and $E(\pi_i) = E(P) = \pi$, so that $E(y_i - y_i^* | \pi_i = 1) = E(y_i - y_i^*) = 0$. This means that if the assumptions of the model are correct, the imputation expressed by equations (1.1) through (1.3) does not add any (model) bias to the direct expansion estimator. The trick, of course, is that the assumptions of the model - especially those concerned with ignorability - must be correct.

Observe that equation (3) can be expressed more traditionally as:

$$y_i = \beta x_i + u_i, \tag{5}$$

where $\beta = \pi\beta^+$, and $u_i = \epsilon_i + (\pi_i - \pi)\beta^+ x_i$. With

this formulation, nonresponse is said to be ignorable when

$$E(u_i | i \in R) = E(u_i | i \in N). \tag{6}$$

NASS believes, however, that respondents are often more likely to have no cropland (or hogs) than nonrespondents (the easier answer to give is "I don't have any"). This means that the left hand side of (6) is likely to be smaller than the right hand side because of the impact of π_i on u_i. As a result, nonresponse would not be ignorable.

The modified BAIP attempts to compensate for the greater tendency of nonrespondents to have cropland (or hogs) than respondents by assuming (4.2) instead of (6); that is, moving nonrespondents known to have cropland (or hogs) into the respondent category when considering the imputation of π_i. Mild empirical support for this practice is offered in a following section.

The next two sections are concerned with the specifications of the model in equation (3). That is, choosing the imputation cell and the x_i so that the model fits the data and nonresponse can reasonably be thought of as ignorable.

MEASURING THE FIT OF ALTERNATIVE SPECIFICATIONS

There is no way of using the survey itself to judge how good alternative specifications of the model are at predicting the true survey values of nonrespondents. It is possible, however, to compare how well these specifications predict <u>respondent</u> values. That is what this section does. The results discussed in this section carry the most weight when nonresponse is ignorable under each of the specifications being compared. Nevetheless, we leave the question of ignorability for the following section.

Let us focus first on cropland. Recall that the model stipulates that within an imputation cell:

$$y_i = \pi_i(\beta^+ x_i + e_i), \qquad (3)$$

where e_i is a random variable such that $E(e_i) = 0$ and $Var(e_i) = \sigma_i^2$. If all the σ_i were equal to σ_0, say, then applying a standard regression package to the members of R^+ (remember: $\pi_i = 1$ for all members of R^+) would produce an unbiased estimate of σ_0^2, call it s^2. It is not difficult to show that when the σ_i^2 are not all equal, s^2 is a consistent estimate of $\Sigma_{i \epsilon R^+} \sigma_i^2/p$, where p is the size of R^+.

The estimator s is an inverse measure of the goodness-of-fit of the model among the members of R^+ -- the smaller s is the better the fit. Thus, given several specifications involving that same imputation cell, it is possible to compare s values to see which one has the best fit.

Table 3 displays the results of a comparison of six specifications for cropland. As before, the data source is the June 1989 Agricultural Survey in 11 states. The standard specification treats the stratum as the imputation cell and sets $x_i = 1/m_i$ (where m_i is the number of reporting units generated from the same sampling unit as i; m_i is usually 1). The first alternative uses the same x_i values but treats the stratum/district as the imputation cell. An s for this alternative was calculated for each stratum/district and then divided by the stratum s value of the standard. We call this ratio the <u>relative root mean squared error</u> (RRMSE) of the alternative for the stratum/district. The state and aggregate median of the alternative's RRMSE's are found in the first column of Table 3.

All of the columns of Table 3 are medians of RRMSE's for different specifications. One interesting result concerns MLSFCROP which is past total cropland from an external source rather than computed cropland (MLSFCRP3) (see [4] for more details on the distinction between the two). The MLSFCROP/m_i appear to make better x_i values than the MLSFCRP3/m_i. This is not surprising since the definition of MLSFCROP is closer to CROPLAND than is MSLFCRP3. The argument for using MLSFCRP3 has been that it would ultimately better serve in a predictor for specific crop values. For that reason, an analysis similar to Table 3's was performed on corn (item code 530), soybeans (600), and winter wheat (540). It turned out that for these crops there is virtually no difference in the 11-state median RRMSE's between using the MLSFCRP3/m_i values as the x_i and using the MLSFCROP/m_i values.

From Table 3 we can safely conclude that there is a fairly consistent gain in fit (that is, reduction in median RRMSE) from both using MLSFCROP/m_i values as the x_i and from treating the stratum/districts as the imputation cells.

There may be a problem in treating stratum/districts as imputation cells when there are few positive respondents in them. To try to get a theoretical grasp on the issues involved, assume the model in equations (5) and (6) hold and that all $x_i = 1$ and $\sigma_i = \sigma_0$. The added model variance due to nonresponse at the imputation cell level can be shown to equal $(1 + 1/n)E(s^2)$ (see [5]), where n is the size of R. This tells us that even if there is a better fit from using stratum/districts as imputation cells, there may be no reduction in variance (or even an increase) from doing so when n is small.

The above result depends on an ignorability assumption different from the one assumed here (equation (6) rather than (4.1)). Nevertheless, it seems reasonable, based on the bottom line of Table 3, that when $x_i = 1/m_i$ a stratum/district containing four or more positive respondents is a preferable imputation cell to the stratum containing it (since $(1 + 1/4)(.87)^2 < 1$). Moreover, if the x_i equal either the MLSFCROP/m_i and MLSFCRP3/m_i, then a stratum/district containing two or more positive respondents appears to be preferable to the stratum containing it (since, for example, $(1 + 1/2)(.54/.74)^2 < 1$) when x_i = MLSFCROP/m_i).

Table 3. -- Comparing the RRMSE's of Alternative Specifications of the Model for Total Cropland in 11 States

State	Alternatives to the Standard*				
	1**	2**	3**	4**	5**
Illinois	0.89	0.71	0.59	0.89	0.77
Indiana	0.84	0.55	0.46	0.78	0.72
Iowa	0.91	0.75	0.59	0.87	0.73
Kansas	0.82	0.55	0.54	0.78	0.82
Michigan	0.92	0.69	0.40	0.86	0.51
Minnesota	0.80	0.63	0.58	0.72	0.75
Mississippi	0.89	0.75	0.77	1.03	1.03
Nebraska	0.84	0.73	0.56	0.88	0.76
Ohio	0.78	0.45	0.42	0.85	0.88
South Dakota	0.90	0.69	0.54	0.86	0.65
Wisconsin	0.90	0.60	0.46	0.75	0.56
11 State Composite	0.87	0.66	0.54	0.86	0.74

Note: Imputation cells with less than two members of R^+ were omitted from the analysis.

* The standard sets $x_i = 1/m_i$ and the stratum is the imputation cell.

** For alternative 1:
$x_i = 1/m_i$ and the stratum/district is the imputation cell.

For alternative 2:
$x_i = MLSFCRP3/m_i$ and the stratum/district is the imputation cell.

For alternative 3:
$x_i = MLSFCROP/m_i$ and the stratum/district is the imputation cell.

For alternative 4:
$x_i = MLSFCRP3/m_i$ and the stratum is the imputation cell.

For alternative 5:
$x_i = MLSFCROP/m_i$ and the stratum is the imputation cell.

Table 4. -- Comparing the RRMSE's of Alternative Specifications of the Model for Total Hogs and in 11 States*

State	Alternatives to the Standard*		
	1**	2**	3**
Illinois	0.93	0.78	1.00
Indiana	0.72	0.65	0.98
Iowa	0.76	0.76	0.98
Kansas	0.76	0.70	0.99
Michigan	0.84	0.78	0.90
Minnesota	0.75	0.73	0.99
Mississippi	0.85	0.83	0.99
Nebraska	0.79	0.72	1.01
Ohio	0.73	0.68	0.92
South Dakota	0.88	0.90	0.98
Wisconsin	0.83	0.75	0.98
11 State Composite	0.81	0.75	0.98

Note: Imputation cells with less than two members of R^+ were omitted from the analysis.

* The standard sets $x_i = 1/m_i$ and the stratum is the imputation cells.

** For alternative 1:
$x_i = 1/m_i$ and the stratum/district is the imputation cell.

For alternative 2:
$x_i = MLSFHOGS/m_i$ and the stratum/district is the imputation cell.

For alternative 3:
$x_i = MLSFHOGS/m_i$ and the stratum is the imputation cell.

Table 4 is analogous to Table 3 except that it concerns imputing for total hogs and pigs. Oddly enough, there is virtually no increase in fit (reduction in median RRMSE) from using the MLSFHOGS values as controls when the strata are the imputation cells. Using the same reasoning as above, there should be at least two positive respondents in the imputation cell whichever x_i is used (MLSFHOGS/m_i or $1/m_i$).

ASSESSING THE BIAS FROM ALTERNATIVE SPECIFICATIONS

Although one reason for using control data and agricultural statistics district information in the modified BAIP is to better predict missing survey values, another, perhaps more important, reason is the removal of potential sources of nonresponse bias. We have noted that when the model in equations (3) through (4.2) is correct, no bias is added to a direct expansion when missing values are imputed using the modified BAIP. Nevetheless, whether it really is reasonable to assume the model for particular specifications remains an open question at this time.

To better understand the issues involved in that question, consider the following example. Let y_i be a total cropland survey value and let x_i = MLSFCROP/m_i. Suppose that

$$y_i = \pi_i(\beta^+ x_i + e_i), \qquad (3)$$
$$E(e_i | i \in R) = E(e_i | i \in N), \qquad (4.1)$$
and
$$E(\pi_i | i \in R \cup K) = E(\pi_i | i \in N - K) \qquad (4.2)$$

hold for all reporting units in a stratum/district. For some reason, however, modified BAIP imputation has been based on different specifications of the model; namely that

$$y_i = \pi_i(\beta'/m_i + e_i'), \qquad (3')$$
$$E(e_i' | i \in R') = E(e_i' | i \in N'), \qquad (4.1')$$
and
$$E(\pi_i | i \in R' \cup K' = E(\pi_i | i \in N' - K') \qquad (4.2')$$

apply for all reporting units in a <u>stratum</u>, where R', N', and K' are respectively the sets of respondents, nonrespondents, and known positive nonrespondents in the stratum.

It is conceivable that the second model (equations (3') through (4.2')), while not as good as the first model ((3) through (4.2)), is nonetheless reasonable.

That is what was assumed in the previous section. We know, however, from Table 3 of that section that cropland values were related to MLSFCROP/m_i values and that they varied systematically across districts within strata (the better fit when x_i = MLSFCROP/m_i and the stratum/districts were the imputation cells tells us so). As a result, for equation (4.1') to be reasonable: (1) the propensity to respond should not be correlated with the x_i values; and (2) there should be no systematic differences among the district response rates.

Table 5 is concerned with comparing the the x_i values (where x_i = MLSFCROP/m_i or MLSFHOGS/m_i) of respondents and nonrespondents. It does this by computing the value log(r/n) for each stratum, where r is the mean of the x_i among the respondents (and later the positive respondents) and n is the analogous mean among the nonrespondents (and later the known positive nonrespondents). The statistic log(r/n) was chosen to compare r and n across strata because it has the nice properties that log(n/r) = - log(r/n) and log (kr/[kn]) = log(r/n)).

By taking the mean(s) and median(s) of these stratum values across the 11 states under review, we test whether there was a systematic tendency for nonrespondents to have systematically higher x_i values then respondents. The first row of Table 5 appears to confirm the existence of this tendency, certainly for the total hogs item and perhaps for the total cropland item as well (at least judging by the signed-rank test).

Recall, however, that the underlying question was the reasonableness of equation (4.1'). Observe that without loss of generality the e_i' in equation (3') can be arbitrarily set equal to zero when π_i = 0; i.e., when the reporting unit has no cropland (or hogs). As a result only the behavior of reporting units with positive item values is of any relevance.

The second row of Table 5 denies that that there is a systematic tendency for x_i values of positive respondents to differ from those of known positive nonrespondents for either the cropland or hogs item. This piece of information is not absolutely convincing because it is still possible that the x_i values of <u>unknown</u> positive nonrespondents are systematically different from those of known positive nonrespondents. Nevertheless, the reasonability of the second model can not be disputed based on the data relating to x_i values alone.

Table 5 also gives support to NASS's unwillingness to assume that reporting units having cropland (or hogs) were as likely to respond as those who don't. Full item response appears to be inversely related to the survey item itself, while partial response (at least answering "has cropland" or "has hogs") might not be so related. Unfortunately, a stronger statement than that can not be made from the evidence uncovered in this study.

While Table 5 serves to ease fears that failing to incorporate x_i values in the modified BAIP would bias direct expansions, Table 6 shows that, in fact, incorporating them may lead to bias. Suppose the true model were

$$y_i = \pi_i(\alpha + b^+x_i + e_i''), \qquad (3'')$$

$$E(e_i''|i \epsilon R) = E(e_i''|i \epsilon N), \qquad (4.1'')$$

and $$E(\pi_i|i \epsilon R \cup K) = E(\pi_i|i \epsilon N - K), \qquad (4.2'')$$

at the stratum/districts level, but that the modified BIAP was based on model 1 (equations (3) through (4.2)). It has been shown that this practice will bias the expansion (equation (2.1)) in the direction of the sign of α (see [6]). The absolute value of the bias is inversely related to the number of positive respondents in the cell and positively related to the variance of the x_i within the cell.

The modified BAIP based on model 1 will bias the expansion upward if the α-values had a systematic tendency to be positive across the stratum/districts. Table 6 suggests they do, certainly for total hogs and perhaps for total cropland (at least according to the signed-rank test).

There is no analogous bias from basing the modified BAIP on model 2 (equations (3') through (4.2')). As long as α is less than 0.5, however, the fit of model 2 will not be as good as that of model 1 (again, see [6]).

NASS implicitly assumed model 1 when it used the BAIP for the June 1989 Agricultural Survey. There is a 0.52% difference in the direct expansions of the 11-state aggregate based on imputation with and without control data (this new calculation compares using x_i = MLSFCRP3/m_i as in column 2 of Table 1 to

$x_i = 1/m_i$)). It is impossible to tell whether this is the result of the systematic upward bias of using the control data in imputation or the slight (and perhaps statistically insignificant) tendency for the control values of nonrespondents to exceed those of respondents (see Table 5).

Table 7 attempts to measure whether there was a tendency for one region in state to have a systematically different cropland item response rate than another region (the hog response is strongly correlated with cropland response). Agricultural statistics districts 1, 2, and 3 form one region, while districts 7, 8, and 9 form the other. The long footnote explains the statistic used.

Since there are fewer than 20 strata in all states, none of the tests are very powerful. Nevertheless, it appears that in at least some states there are systematically different response rates across regions. For example, in Indiana the first region has a significantly higher response rate (at the .002 level, either test) among known positives than the second region.

Recall that the last column of Table 1 measured the effect of treating the strata as the imputation cells rather than the stratum/districts. The effect on the 11-state aggregate was trivial, -0.07%. The effect on Indiana was also very small, -0.11%. The effect of Nebraska, however, was not: 1.05%.

Table 5. -- Comparing the Control Values of Respondents and Nonrespondents Across the 11 States

	Total Cropland		Total Hogs & Pigs	
	Mean	Median	Mean	Median
Log Ratio of Respondents to Nonrespondents*	- 0.04	- 0.04	- 0.26	- 0.04
(Significance**)	(.317)	(.019)	(.000)	(.000)
Log Ratio of Positive Respondents to Known Pos. Nonresp.*	0.03	- 0.01	- 0.01	- 0.01
(Significance**)	(.467)	(.956)	(.780)	(.752)

* Let x_i = MLSFCROP/m_i for cropland and MLSFHOGS/m_i for hogs. The "log ratio" value is the log of the stratum mean of the x_i among the (positive) respondents divided by the stratum mean among the (known positive) nonrespondents. Only well defined stratum ratios were used in calculating the means and medians.

** The significance level of a two-sided test of whether the mean (or median) is different from zero. A conventional t-test was used for the mean, while a signed-rank test was used for the median.

Table 6. -- Is Total Cropland (and Hogs) Really a Linear Function of the Control Variable Through the Origin Across the 11 States?

	Total Cropland		Total Hogs & Pigs	
	Mean	Median	Mean	Median
Stratum/District Estimate of a/m *	0.15	0.19	0.41	0.45
(Significance**)	(.223)	(.000)	(.000)	(.000)

* Let x_i = MLSFCROP/m_i for cropland and MLSFHOGS/m_i for hogs. The parameters of $y_i = \alpha + bx_i + r_i$ were estimated (separately for cropland and hogs) using the p positive reporting units in the stratum/district; $m = \Sigma y_i/p$. All stratum/districts with less than three positive reporting units have been excluded from the calculation of means and medians.

** See footnote ** for Table 5.

Table 7. -- Comparing Cropland Response Rates in Two Distinct Regions of 11 States

	Significance of Respondent to Nonrespondent Log Ratio*		Significance of Pos. Respondents to Known Pos. Nonresp. Log Ratio*	
	t-test	s-test	t-test	s-test
Illinois	.222	.520	.488	.320
Indiana	.171	.129	.002	.002
Iowa	.563	.597	.692	.404
Kansas	.284	.542	.891	.791
Michigan	.557	.688	.052	.156
Minnesota	.731	.860	.257	.391
Mississippi	.304	.049	.074	.091
Nebraska	.091	.153	.092	.194
Ohio	.041	.034	.313	.266
South Dakota	.685	.978	.798	.762
Wisconsin	.125	.193	.363	.496

* For each stratum, let z = Log (x/y), where x is the number of (positive) respondents in the stratum from agricultural statistics districts 1, 2, or 3 divided by the number of (positive) respondents in the stratum from agricultural statisitcs districts 1, 2, 3, 7, 8, and 9, and y is the number of (known positive) nonrespondents in the stratum from agricultural statistics districts 1, 2, or 3 divided by the number of (known positive) nonrespondents in the stratum from agricultural statistics districts 1, 2, 3, 7, 8, and 9. The table displays significance levels of two-sided tests of the null hypothesis that the true state mean of the z equals zero. The s-test is the signed-rank test.

DISCUSSION

In the last several sections, we have acted as if the superiority of the modified BAIP's treatment of business status over the BAIP's goes without saying. It does not. The preference for the former shown in this paper is the result of three considerations:

1) the modified BAIP's approach is more consistent with livestock adjusted reweighting,

2) Summary 2 crop expansions are thought to be too high,

3) many reporting units with unknown business status have in the past been incorrectly coded "in business" (for example, some states have inaccessibles coded as in business in the June 1989 Agricultural Survey; if a unit is inaccessible, how do we know whether it is in business?).

Although it can be dangerous to make broad generalizations based on one survey period, the following observations seem to be warranted:

1) If future nonresponse adjustments are to use cropland controls, MLSFCROP appears to be better than MLSFCRP3. In addition to MLSFCROP producing a better predictor of cropland (and an equally good predictor of corn, soybeans, and winter wheat in the 11 states studied), NASS often has the MLSFCROP value for a sampling unit but not its MLSFCRP3 value. The BAIP presently has a (needlessly) complicated fallback methodology for handling that situation.

2) Although there may be some justification for using control data when imputing crops but not hogs (a slightly better fit, less bias), there is no justification - at least none unearthed in this study - for using agricultural statistics district information only for crops.

3) There may be a bias from using control data in cropland imputation, and there does not appear to be a bias from <u>not</u> using such data. This and the desire to keep things simple suggest that the NASS may want to abandon the use of control data in crop and cropland imputation.

All hog items are like the total hogs and pigs item in June and so the analysis in this paper applies to them. By contrast, it is possible for the reporting unit in June to be a crop item nonrespondent but have a response to the total cropland question. This happens so infrequently that using the real cropland responses of those reporting units to impute for their specific crop values is a costly and mostly unproductive exercise (the Appendix offers empirical support to this argument). Crop item should be treated as cropland with "has crops" (i.e., the results of item code 138 exclusively) replacing "has cropland."

If NASS abandons the use of control data (stock items are discussed in the following section) in imputation, it will be tempting to save processing time by performing imputation during the Survey Processing System (SPS) Summary. This would be hard to do in the Summary if stratum/districts serve as the imputation cells because there may be nonrespondents in cells with less than <u>two</u> positive respondents. Moreover, the empirical work here suggests that this lower limit should be doubled for total cropland (but ironically not for total hogs). What is needed is the ability to collapse cells when necessary. That ability can be found in an external program like the BAIP. Alternatively, NASS could use district information in the stratification process, so that the strata could serve as the imputation cells, thereby improving the fit and decreasing the fear of bias. Clearly, this suggestion needs more careful study before it is implemented.

Whatever method of imputation is finally selected, variances can be (conservatively) estimated using the same formula as for the adjusted livestock estimator (see [7]).

EXTENSIONS

Extending the analysis to list expansions for livestock and crop items on follow-on Agricultural Surveys is largely straightforward. Moreover, there is no reason why total land (item 900) should not be imputed in a manner analogous to total cropland.

One problem with specific crop imputation in a follow-on survey is that compared to June it is more common to have a cropland value but not specific crop values for a reporting unit. This is because the Edit Program [1] will replace missing cropland values with previously reported ones. This practice needs to be analyzed in a separate empirical investigation. Nevertheless, it seems reasonable that if NASS chooses to do away with control data for cropland imputation, it should likewise remove total cropland from crops imputation.

Another problem in crops imputations is production. It is possible for a reporting unit to supply harvested acres of a crop but not its production (for example, supply item code 400, harvested corn, but not 401, corn production). The BAIP imputes for this situation in a reasonable way that also should be analyzed empirically in the future (the harvested acres of the crop is multiplied by
Σ production / Σ harvested acres, where both summations are over all reporting units in the stratum/district with valid production values). For the time being, the methodology should remain in effect virtually as is if a separate imputation program is continued. If not, stratum level imputation along the same lines could be incorporated into the SPS summary.

The question of stock imputations has been avoided up until this point. There are too many difficult questions concerning stocks to be answered here. We regrettably are forced to leave most of the issues of stock imputation for another time. Quickly, however, if the BAIP is modified as outlined here, the stocks sections of the program could remain largely as is. The one change, and it also affects production, is that survey values (raw values times LAF's) would be imputed directly based on respondent survey (not raw) values.

On the other hand, if imputation is moved to the SPS Summary, stock imputation will have to be rethought. One possibility is to make it exactly parallel to crop imputation. In current practice, however, it is possible to be a known positive for a specific grain (or rice) stock. There is no crop item with that property except production.

There is much that NASS could do differently in the new SPS Summary. For example, it will be easier in the future to distinquish between a raw value of 0 and a missing value. This could allow greater flexibility in all item imputation, a speculation that merits attention in the future.

Another place deserving of attention is the area (or non-overlap) side. Presently only stocks are imputed by the BAIP for June, while crops but not livestock are imputed for follow-on surveys. The raw values for crop items in June and livestock items for all Agricultural Surveys are hand imputed by NASS if necessary.

One suggestion for handling area frame nonresponse in the future for crop and livestock items would be to break up a state's area frame into 10 reweighting cells, one cell for nonagricultural subtracts (reporting units) and one for agricultural subtracts in each of the nine agricultural statistics districts. Within each cell and for each survey item, the expansion factor E_i for item respondent i would be inflated to

$$E_i^* = \frac{n(p + k)}{p(r + k)} E_i, \qquad (7)$$

where n is the number of sampled subtracts in the cell, r is the number of item responding subtracts in the cell, p is the number of positive item responding subtracts, and k is the number of known positive item nonresponding subtracts.

This is precisely the approach to nonresponse taken in the SPS Summary for labor surveys (see [7]), except that in labor surveys there is only one reweighting cell and k = 0. The theory behind (7) is that full item nonresponse (not even answering the "has crops" or "has hogs" question) is random within reweighting cells and that item response is random among known positives. Observe the similarility of equation (7) to equation (2.2) when all the $x_i = 1$.

Variance estimation under the reweighting scheme described above would be unchanged from the procedure currently in the SPS Summary. If the assumptions underlying the reweighting are correct, variance estimators will be conservative (see [7]).

Before implementing the above approach, of course, more research is called for, first into the exact formation of reweighting cells and second into the potential for bias from not using June information in follow-on survey nonresponse adjustments.

RECOMMENDATIONS

The main recommendation of this paper is that nonresponse adjustment procedures for Agricultural Surveys should be unified at least conceptually. In particular:

1) LAF imputation should be removed from crop item imputation.

2) The operational expansion for livestock items should be discontinued or the Summary 1 crop item expansion resurrected (they are virtually the same thing).

3) Record-by-record imputation for crop items should be made equivalent to adjusted reweighting for livestock items. This requires a modification in the control values for cropland (recall that x_i = MLSFCROP/m_i, MLSFCRP3/m_i or 1/m_i in the text) and an analogous adjustment of the SPS Summary for livestock items.

4) Crop and livestock item variance estimation should take on the same form as livestock variance estimation presently does in the SPS Summary.

5) Total land item nonresponse should be handled in an analogous manner as total cropland nonresponse.

6) NASS should eventually either impute for all items on the list side with a modified imputation program that incorporates agricultural statistics district information for livestock items as well as crop and stock items, or apply adjusted reweighting to all crop and livestock items at the stratum level. If the second approach is taken, strata should be formed using agricultural statistics district information. Stocks require further study.

7) NASS should investigate further a method of reweighting for area (NOL) expansions suggested here.

8) If NASS chooses to continue using cropland control data, it should consider replacing MLSFCRP3 by MLSFCROP. Moreover, NASS should impute for missing control values on its Edited Data Files. This can be done using full Master File information.

REFERENCES

[1] Ferguson D., Ison, B., Stepanich, J., and Williams, D. Survey Processing System: Data Validation Sub-System. National Agricultural Statistics Service, 1988.

[2] Atkinson, Dale. The Scope and Effect of Imputation in Quarterly Agricultural Surveys. RAD Staff Report No. SSB-8804, February 1988.

[3] Kott, P. S. and Thorson, J. Improving Variance Estimates for Livestock Surveys. RAD Staff Report No. SRB-8901, February 1989.

[4] NASS. Agricultural Surveys - Supervising and Editing Manual: June 1989 - March 1990, Section 3, p. 339, January 1989.

[5] Kott, P.S. "Variance Estimation with Imputed Data," Bulletin of the International Statistical Institute: Contributed Papers, Book 1, p. 532-3, 1989.

[6] Foreman E., and Brewer, K., "The Efficient Use of Supplementary Information In Standard Sampling Procedures," Journal of the Royal Statistical Society, B, p. 391-400, 1971.

[7] Kott, P.S. Mathematical Formulae for the Survey Processing System (SPS) Summary. National Agricultural Statistics Service, forthcoming.

APPENDIX: HOW THE BAIP WAS SIMULATED

There were three differences in how the BAIP was simulated for this study and how it actually was run for the June 1989 Agricultural Survey:

1) Since an Imputed Data File was used in the analysis, all records with an item code 138 equal to 1 or 2 were treated as if the cropland (item code 802) value was missing, while all records were treated as if the cropland value was valid.

2) LAF imputation was streamlined. If a reporting unit had an item code 921 equal to 11, then its LAF was imputed as the mean LAF among all reporting units in the stratum with item codes 921 less than 11 and not equal to 10. If a reporting unit had an item code 921 equal to 12, then its LAF was imputed as the mean LAF among all reporting units in the stratum with item codes 921 less than 11.

3) When a MLSFCRP3 value was not available for a record it was imputed using its MLSFCROP value ($MLSFCRP3^*$ = MLSFCROP x [Σ MLSFCRP3 / Σ MLSFCROP], where both summations ares over all the sampled records in the imputation cell with known MLSFCRP3 <u>and</u> MLSFCROP values). If that value was also unavailable, the MLSFCROP value was imputed by the cell mean of MLSFCROP values.

The same collapsing rules were applied to the simulated BAIP as the BAIP itself. These rules were also applied to imputing MLSFCRP3 and MLSFCROP values.

The cumulative effects of the three modifications of the BAIP are displayed in the following table:

State	% Difference in the expansion for total cropland when the simulated BAIP replaces to the BAIP
Illinois	- 0.16
Indiana	- 0.40
Iowa	0.08
Kansas	0.34
Michigan	0.04
Minnesota	- 0.09
Mississippi	- 0.69
Nebraska	0.94
Ohio	- 0.14
South Dakota	0.07
Wisconsin	- 0.17
11 State Aggregate	- 0.06

They are not very large.