United States Department of Agriculture

National
Agricultural
Statistics
Service

Research Division

SRB Research Report
Number SRB-93-03

March 1993

WEIGHTING CLASS ADJUSTMENTS FOR NONRESPONSE IN INTEGRATED SURVEYS: FRAMEWORK FOR HOG ESTIMATION

Brenda G. Cox

WEIGHTING CLASS ADJUSTMENTS FOR NONRESPONSE IN INTEGRATED SURVEYS: FRAMEWORK FOR HOG ESTIMATION, by Brenda G. Cox, Research Fellow, American Statistical Association and National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, DC 20250-2000, March 1993, Report No. SRB-93-03.

ABSTRACT

The National Agricultural Statistics Service (NASS) derives its State and national estimates of inventory and production for crops and livestock from the Quarterly Agricultural Survey (QAS). The QAS uses two frames for sample selection: a list frame of agricultural operations derived from various data bases and an area frame constructed by dividing the total land area of the United States into sampling units or "segments." For its dual frame estimates, NASS removes the multiplicity associated with selecting from the full-coverage area frame and the overlapping list frame by using the list frame to represent listed (overlap) operations and the area frame for the unlisted (nonoverlap) operations. This paper presents a weighting plan for hog estimation associated with the list frame portion of the QAS. The presentation is sufficiently general, however, that the results can be adapted to other commodities of the QAS as well as to other integrated surveys such as the Farm Costs and Returns Survey and the January Cattle Survey.

KEY WORDS

Sampling weight; Nonresponse adjustments; Poststratification; Weighting class adjustment; Agricultural operation; Hog operation.

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

ACKNOWLEDGEMENTS

This paper grew out of an assignment to develop weighting specifications for the National Swine Health Survey, which subsampled QAS respondents. The author would like to thank Martha Farrah Roehrenbeck for suggesting this research topic and Ron Bosecker and Bill Iwig for their helpful comments and suggestions. The author also appreciates the many NASS staff who listened to early presentations of this material and suggested how to make the paper more relevant and understandable to NASS staff.

TABLE OF CONTENTS

SUMMARY iii

INTRODUCTION 1

WEIGHTING TO ACCOUNT FOR SAMPLE SELECTION 2

POSTSTRATIFICATION TO QAS FRAME COUNTS 2

THE NASS STRATEGY FOR DEALING WITH FRAME MULTIPLICITY 4

ADJUSTMENT FOR MULTIPLICITY DETECTED FROM NON-INTERVIEW DATA 5

ADJUSTMENT FOR UNCERTAINTY ABOUT AGRICULTURAL STATUS 6

ADJUSTMENT FOR UNCERTAINTY ABOUT HOG STATUS FOR KNOWN OPERATIONS 9

ADJUSTMENT FOR LOSS OF INTERVIEW DATA FROM KNOWN HOG OPERATIONS 12

ADJUSTMENT FOR MULTIPLICITY DETECTED FROM INTERVIEW DATA 13

ASSUMPTIONS UNDERLYING NASS EXPANSION FACTORS 13

ESTIMATION USING ANALYSIS WEIGHTS 18
    Estimation for the Operational Estimator 18
    Estimation for the Adjusted Estimator 21
    Estimation for the Revised Estimator 23

CONCLUDING REMARKS 25

REFERENCES 27

SUMMARY

The Quarterly Agricultural Survey (QAS) is an ongoing survey conducted by the National Agricultural Statistics Service (NASS) to provide inventory and production estimates for crops and livestock at State and national levels. Prior to December 1986, these data were obtained in separate surveys, each targeted to a specific commodity. Frame information as to size of the operation with respect to the commodity of interest was used to stratify the frame and to allocate the sample to strata. With the integrated design, the strata and sample size allocations are compromises between the diverse needs of the various commodities and hence are not optimal for any one commodity. The integrated design is more efficient, however, reducing data costs in ways that also decrease overall respondent burden.

Integration does have its disadvantages; the most important of which is that estimation for individual commodities becomes more involved, particularly the steps to compensate for nonresponse. For the hog portion of the QAS, NASS uses weighting procedures that are simple extensions of the procedures used prior to integration. In this paper, I argue that the current procedures are in fact too simple to reflect the complexities associated with an integrated design.

First, I describe a generalized approach for weighting the list frame portion of the June QAS for hog estimation. This generalized approach incorporates poststratification adjustments using poststrata based upon size of the hog portion of the operation and suggests a different treatment for multiplicity (now combined into the "list adjustment factor"). A stepwise approach to non-response adjustment is described that makes clearer how to use partial data obtained for non-respondents. Comments are also given on variance estimation for this alternate weighting approach.

I then present the assumptions that NASS analysts are implicitly making when they base inferences upon the operational and adjusted hog estimators in current use. These assumptions clarify why the operational estimator can be expected to consistently underestimate population totals. The adjusted estimator is less seriously biased than the operational estimator but also can be expected to underestimate population totals. Both of these estimators fail to make adequate use of the partial data about the agricultural status of nonrespondents; the operational estimator also fails to use the data obtained on the hog status of identified agricultural operations. In addition to being biased, these estimators are less precise than estimators based upon the alternative weighting approach described in this paper.
The final section of the paper provides recommendations for future research in order to apply these weighting concepts.

INTRODUCTION

One of the responsibilities of the National Agricultural Statistics Service (NASS) is to provide inventory and production estimates for crops and livestock at the State and national levels. The Quarterly Agricultural Survey (QAS) furnishes the vehicle to obtain the necessary input data to form these estimates (Bosecker 1987).

The QAS is based on a dual frame design with stratified random sampling from a list frame of operators and stratified cluster sampling from a frame of geographically defined areas. The list frame is a list of names of operations and/or operators constructed by merging lists from various sources and removing duplicate entries. The area frame is constructed by partitioning the entire land area of the United States into sampling units or "segments."

Since agricultural operations can be linked to land, the area frame provides complete coverage of agricultural operations but at a greater cost than list frame interviews. The advantages of the list frame are its capabilities for lower interview costs and for more efficient sampling of specific commodities. Conceptually, the QAS divides the universe of agricultural operations into two components to correspond with presence on the list (overlap operations) versus absence from the list frame (nonoverlap operations). NASS then avoids the complexities of dual frame estimation by combining data for the overlap operators sampled from the list frame with data for nonoverlap operators sampled from the area frame.

For area frame interviews, NASS replaces missing data due to total (and item) nonresponse with logical imputations based upon interviewer observations and expert judgement. Hence, weighting for the area component is straightforward, with simple expansion factors to reflect the sampling of the area segments.

In lieu of imputation for total nonresponse for list frame selections, nonresponse adjustments are made to the weights (with the exception of extreme operators for whom all missing data are logically imputed). In addition, some operations are associated with more than one frame record giving them multiple opportunities of selection. Weighting list frame interviews is more involved as a result of the steps needed to adjust for frame multiplicity and nonresponse as well as the sampling of the list frame records.

In this paper I present a general approach for weighting the list frame portion of the QAS, restricted to the June survey hog data. After presenting the basic approach, I show the conditions under which the current QAS weighting approach can be considered a special case of these procedures and what models are implied by these expansion factors. This paper builds upon the concepts that I summarized for a short course on weighting procedures originally developed for the American Statistical Association and later presented at NASS (Cox, 1991).

WEIGHTING TO ACCOUNT FOR SAMPLE SELECTION

To begin, QAS selects a stratified random sample from the list frame. Hence, the sampling weight for each sample selection (the inverse of the probability of selection) can be expressed as

$$W_{sam}(hi) = \frac{N(h)}{n(h)} \tag{1}$$

where

$W_{sam}(hi)$ is the sampling weight for the i-th sample record from the h-th stratum,

N(h) is the total frame records associated with the h-th stratum, and

n(h) is the stratum h sample size.

This weight reflects the probability of selection of the sampled record from the list frame rather than the agricultural operation and sums to the total number of list frame records.

POSTSTRATIFICATION TO QAS FRAME COUNTS

Commodity data for hogs were originally obtained in a separate survey with strata defined based upon frame information (control data) as to size of the hog portion of the operation and with sample sizes optimally allocated for hog estimates. In 1986, the hog survey was integrated with other surveys being conducted to study crops and storage and cattle and other livestock. For the current integrated crops and livestock survey, the strata and sample size allocations are compromises between the diverse needs of the individual commodities and hence, are not optimal for any one commodity. The integrated design is worthwhile as it reduces costs and respondent burden, but integration makes weight development and estimation more difficult.

Generally, strata definitions for the integrated design are coached in terms of a size measure for a particular crop or type of livestock. A priority ordering is also defined so that operations producing more than one commodity are uniquely assigned to strata. One effect of this ordering scheme is that hog producers of the same size with respect to control data are not always assigned to the same stratum. Hence, the sample weights for hog operations of a particular size in terms of control data will sum to an estimate of the record count rather than the exact known number of frame records.

This furnishes an ideal situation for poststratification adjustment of the sampling weights using poststrata targeted to the commodity of interest, in this case hogs. As Holt and Smith (1979, p. 33) note:

> ... neither the post stratification estimator nor the sample mean is uniformly best in all situations but empirical investigations indicate that post stratification offers protection against unfavorable sample configurations and should be viewed as a robust technique.

The overall effectiveness of the poststratification will of course depend on the quality of the hog control data in terms of its correlation with hog production. (The same constraint holds for the effectiveness of the design strata.)

To illustrate the use of poststratification, let g index poststrata defined using hog control data found on the list frame, where g = 1, 2, ..., G. The poststratification adjustment factor for the g-th poststrata is defined as:

$$A_{ps}(g) = \frac{N(g)}{\sum\limits_{hi \epsilon S(g)} W_{sam}(hi)} \qquad (2)$$

where

N(g) is the total frame records associated with the g-th poststratum, and

S(g) is the set of sample records that are found in the g-th poststratum.

The poststratified adjusted weight for the i-th sample record from the h-th design stratum and the g-th poststratum is then calculated as:

$$W_{ps}(ghi) = A_{ps}(g) \; W_{sam}(hi) \quad . \qquad (3)$$

Note that when summed over members of poststratum g the poststratified weights now total N(g).

Poststratification (of a simple random sample) can produce results almost the same as that of a proportionally allocated stratified random sample when the poststratum sample sizes are sufficiently large (> 20) and errors in the proportions falling into each poststratum [e.g., N(g)/N] can be ignored (Cochran 1977, pp 134-135). In our situation, there is no error in the population proportions as they are simply frame counts.

The best poststrata for adjustment purposes will be those that would have been optimal for use in stratification for an independent survey of hog producers. For populations such as hog producers where significant gains can be made from stratification, NASS follows Cochran's recommendation that the cum $\sqrt{f(y)}$ rule of Dalenius and Hodges (1959) be used for deciding stratum boundaries and his further suggestion that there is little reduction in the variance for more than six poststrata (Cochran 1977, pp. 127-134). Exploiting these concepts and recognizing that NASS does separate estimation for each State, up to six poststrata might be created for each State based upon the frame control data, with additional categories created for records with control data that are missing and records with control data values of zero hogs.

Desirable poststratum sizes can be achieved but the poststrata will have to cross stratum boundaries to do so (i.e., the poststrata cannot be nested within the design strata). A starting point for definition of poststrata could be the size categories used to define the hog design strata within a State adding poststrata for records with zero as their control data for hogs and for records with missing control data for hogs.

THE NASS STRATEGY FOR DEALING WITH FRAME MULTIPLICITY

Agricultural operations can have multiple opportunities of selection when they are associated with more than one list frame record. Surveys routinely use two approaches to resolve the problem posed by frame multiplicity. The first approach is to link the desired population unit (e.g., the agricultural operation) to a particular frame unit and only include the population unit in the survey when that frame unit is selected. This removes the multiple opportunities of selection so that the record's probability of selection is also the probability of selection of the population unit. The second approach is to adjust the weight associated with sampled frame units to reflect the multiple selection opportunities for the desired population unit (agricultural operations) associated with those frame units.

QAS uses a combination of these two approaches to deal with agricultural operations with multiple opportunities of selection. As the first step, QAS applies a priority ordering to the design strata. An agricultural operation's data are used for analysis only when a frame record associated with the highest priority stratum is selected. Records from the lower priority strata are considered out of scope for data collection.

When the sample record is associated with the highest priority stratum, the agricultural operation is eligible for data collection. Agricultural operations have multiple opportunities of selection according to the number of records in the highest priority stratum linked to the agricultural operation. Hence, the next step is to determine if duplicate records exist within the highest priority stratum. The probability of selection of an "on the list" or overlap (OL) agricultural operation is the probability of selection of a record from the highest priority stratum times the number of records from that stratum associated with the agricultural operation.

To construct a weight for agricultural operations (as opposed to list frame records), frame multiplicity is taken into account via a weight adjustment. Two different methods can be used depending upon the desired treatment of agricultural operations that have more than one of their associated frame listings selected. Either a single data record is attached to each agricultural operation regardless of the number of times the operation was selected (this method is described in Cox 1991) or duplicate data are adjusted through weighting for each selection of the operation.

NASS uses the latter "duplicated-data" method for QAS. When an operation has multiple list frame records selected from the highest priority stratum, the interview data are obtained for the "first" selection and then the operation's responses copied for the other selected records. Fractional weights suming to one are used to apportion the data among the duplicated records in the stratum.

Comments made in the 1991 June QAS training school suggest that this rule of duplicating data for repeated selections is not always followed in practice; some State staff manipulate the 921 Box for the "first" record so that it gets a LAF of 1.0 and then code the "second" record as out of business. This approach gives the correct result for current weighting procedures but loses information about multiplicity. It will not produce correct expansion factors for the alternative weighting method described in this paper.

Note that a unique linking could have been developed to eliminate within-highest-priority-stratum duplication. For instance, a rule could be applied that an operation is included in the sample only when the record with the largest ID number from the highest priority stratum is selected.

The present NASS rule deals with all duplication as if it were detected after interviewing was completed. To avoid unnecessary loss of data, the operation is included when any of the highest priority stratum's records is selected. The strata are prioritized to assign operations to the stratum with the largest size measure, thus reducing the occurrence of outliers. Further, this approach avoids the complications of computing selection probabilities when stratum boundaries are crossed among duplicates.

ADJUSTMENT FOR MULTIPLICITY DETECTED FROM NON-INTERVIEW DATA

For all June QAS list sample cases, a duplicate check program is run that prints out sample records for which there are other records in the list frame with the same name, address, telephone number, Social Security Number or Employer Identification Number. Some (but not all) frame duplication can be detected from this detailed comparison of the sample records with other frame records with similar name, location, and identification numbers. Note that this comparison can be made for all records not just those associated with interview respondents. (Some types of frame multiplicity require completed interviews to identify. I describe the use of this type of multiplicity information later.)

When the duplicated-data method is used, the multiplicity-adjusted weight for sampled operations is calculated as:

$$W_{mult}(ghi) = W_{ps}(ghi) \;/\; MAF_{non-intv}(hi) \tag{4}$$

where

$W_{mult}(ghi)$ is the multiplicity-adjusted weight associated with the i-th sample record from the h-th stratum and the g-th post-stratum, and

$MAF_{non-intv}(hi)$ is the multiplicity adjustment factor for the i-th sample record from the h-th stratum, as it is determined from non-interview (non-intv) data only.

$MAF_{non-intv}(hi)$ records the number of records linked to the agricultural operation associated with the hi-th list frame record, as identified from the non-interview data. The W_{mult} weights when summed estimate the number of unique operations associated with the list frame. (The weight sum is a biased estimate to the extent that linkages exist for agricultural operations that can only be identified from interview data.)

Let's look at this multiplicity adjustment factor a little further. Suppose that sample record hi is found in the highest-priority stratum. The multiplicity adjustment factor for sampled record hi then is simply the number of stratum h records linked to the operation associated with the hi-th record.

Now suppose sample record hi is associated with a lower-priority stratum so the operation is not considered to be eligible for data collection when the record is selected. This record gets a multiplicity adjustment factor of $MAF_{non-intv}(hi) = 1.0$ and for later weighting is treated as just another nonagricultural operation.

Finally, suppose that sample record hi is associated with an out of business farm, a nonagricultural operation, or some other entity not in scope for the QAS. Then $MAF_{non-intv}(hi) = 1.0$ as this record is linked to no other agricultural operation.

ADJUSTMENT FOR UNCERTAINTY ABOUT AGRICULTURAL STATUS

The next step in weighting is to adjust for nonresponse. Conceptually, QAS data collection can be regarded as first determining if the record corresponds to a member of the target population of agricultural operations; out of business and nonagricultural operations are not eligible for interview. The first stage and the most serious loss of data, then, is complete loss of all information about the sampled record, including whether the record corresponds to an in-scope agricultural operation.

To be in scope for the study, the operation must during that calendar year presently have, have had, or will have (1) crops grown or hay cut; (2) grains, oilseeds, or hay stored; (3) hogs; or (4) cattle, sheep, goats, livestock, or poultry. This rather long list of conditions arises from the fact that the QAS is not one survey but a set of surveys -- each with its own definition of eligibility -- that have been integrated into one data collection effort.

For convenience, I will use the term "agricultural operation" or "ag-op" from this point on in the paper to mean an operation that satisfies the above conditions for questionnaire administration. Ag-op status is recorded for each case in the 921 Box. Operations known to be agricultural receive codes of 1-8, 10, 11 or 99 while out-of-business or nonagricultural operations are coded as 9's. ("99" is a special 921 Box code used in operator dominant States for specially designated records where the operation rather than the operator is the sampling unit.) A 921 Box code of 12 is assigned to nonrespondents whose ag-op status is unknown.

To adjust for loss of information on ag-op status, an assumption has to be made about the nature of nonresponse for ag-op status determination. I am going to assume that some records with ag-op status unknown are ag-ops while others are nonag-ops. (This is a reasonable assumption to make since the telephone interviews used by QAS may not facilitate identification of the ag-op status of nonrespondents.) In this situation, it is common to assume that classes can be defined such that within classes ag-op status nonrespondents are similar in characteristics to ag-op status respondents. That is, the proportion of code 12 unknown ag-ops equals the proportion of known ag-ops (codes 1-8, 10, 11 or 99) among records whose ag-op status is known (codes 1-8, 9-11, or 99). Note that being a "respondent" at this stage merely means that ag-op status is known for the record.

The best classes within which to adjust for loss of ag-op status information would appear to be the poststrata denoted by the letter "g" earlier; for the sake of generality, however, I am going to assign the letter "c" to the ag-op status weighting classes.

Under this model described above, the nonresponse adjustment factor for ag-op status (AG_ST) determination is calculated as:

$$A_{AG\_ST}(c) = \frac{\sum\limits_{i \in S(c)} W_{mult}(ghi)}{\sum\limits_{i \in S_{AG+nonAG}(c)} W_{mult}(ghi)} , \tag{5}$$

where

$A_{AG\_ST}(c)$ is the weighting class c adjustment factor for ag-op status nonresponse,

S(c) is the set of sample records from weighting class c, and

$S_{AG+nonAG}(c)$ is the set of weighting class c sample records with ag-op status determined (known ag-ops as well as known nonag-ops).

The nonresponse-adjusted weight for weighting class c ag-op status respondents is calculated as the product of this ag-op status nonresponse-adjustment factor and the multiplicity-adjusted sampling weight or

$$W_{AG\_ST}(cghi) = W_{mult}(ghi) * A_{AG\_ST}(c) \quad . \tag{6}$$

This weight is calculated for all ag-op status respondents
-- to reiterate both known ag-ops (codes 1-8, 10-11, or 99) as well as known nonag-ops (code 9).

Ag-op status nonrespondents (code 12's) receive an ag-op status nonresponse-adjusted weight of zero. These cases can be stripped from the data base. However, it is usually best to retain them so that a complete history file exists for all selections. In this case, for all subsequent weighting stages these ag-op status nonrespondents should be given adjustment factors of zero and hence zero as their adjusted weight.

Note that the sum of the W_{AG_ST}(cghi) over all ag-op status respondents in class c sums to the same weight total as the W_{mult}. Since nonrespondents have weights of zero, these weights also sum to the same weight total when added across all records belonging to weighting class c, that is

$$W_{AG\_ST}(c{+}{+}{+}) = W_{mult}(c{+}{+}{+}) \tag{7}$$

where the "+" notation indicates summation over the missing parameters.

In effect what the adjustment does is to take the ag-op status nonrespondents' weights from them and distribute it over the ag-op status respondents. (Stratum totals will not necessarily be preserved when the weighting classes cross stratum boundaries.)

Before going to the next step, I will digress for a moment and mention two changes needed in how ag-op data are obtained to implement this weighting strategy.

First, eligibility for questionnaire administration is derived from Question 1 of the list frame questionnaire. For the 1991 QAS, Part c deals with hogs and asks, "Have or will there be any hogs on this operation at any time during 1991?" Since past, present and future 1991 hog status is of concern, this question should read, "Have there been or will there be any hogs ..." This modification should make the intent for the time frame of this question clearer from the question alone. The Interviewer's Manual should also provide guidance as to the exact definition of those operations in scope for questionnaire administration, either within the manual itself (NASS 1991a) or in a special supplement tailored to the June QAS.

The second change that is needed has to do with the way the Box 921 code of 11 is defined for the list frame questionnaire. For weighting purposes, we need to know when a nonrespondent can be identified as an ag-op, which implies past, present and future 1991 agricultural operations. In short, for a nonresponding case if any part of Question 1 appears to be true now, was true earlier in 1991, or will be true later in 1991, then this information needs to be recorded.

The Supervising and Editing Manual should be revised to provide a more precise description of when to assign the 921 Box code of 11. The present description for the 921 Box for list frame selections suggests that 11 is assigned only for ag-ops presently in operation (NASS 1991b, p. 6069). No explicit directions are given to the editor to attempt to determine if the selection meets the definition of an agricultural operation because of pre-June agricultural activities or planned post-June activities. (Granted this information is more difficult to obtain for operations not presently operating but the editor should still be given the exact definition to apply.)

ADJUSTMENT FOR UNCERTAINTY ABOUT HOG STATUS FOR KNOWN OPERATIONS

The QAS instrument has a separate section for each commodity. In this paper, I focus on estimation for the hog portion of the questionnaire. At one time, hog data were obtained in a separate survey that has since been integrated with other commodity surveys. To be eligible for the hog component of the integrated June questionnaire, the operation must have had hogs or pigs on June 1 or at some time during the period March 1 through May 31 on the land operated by the farm, ranch or individual associated with the frame listing.

Up to now, only one stage in the eligibility process has been examined and that is the determination as to whether the sample unit is an ag-op or not. For hog estimation, this is insufficient. Here only estimation for hogs is desired so that an ag-op that does not raise hogs is not eligible for this portion of the survey. (By the definition given earlier, only ag-ops raise hogs.)

Differential loss of information can occur with some nonrespondents whose status as an ag-op is in question and other nonrespondents that are known to be ag-ops but whose status as a hog operation is unknown. For this next stage of weighting, we restrict attention to known agricultural operations and define an operation to have responded if its hog status is known.

The next step in weighting is adjusting for loss of information as to whether an ag-op had hogs on June 1 or during the previous quarter. Generally, I would assume that the best weighting classes for hog-status determination are the previous weighting classes (denoted by c) cross-classified by ag-op status, although collapsing may be needed to create classes of adequate size. In the following discussion, "d" will denote the weighting classes being used to explain hog-status nonresponse.

For ag-op classes, the nonresponse adjustment factor for hog-status determination is calculated as

$$A_{HOG\_ST}(d) = \frac{\sum_{i \in S_{AG}(d)} W_{AG\_ST}(cghi)}{\sum_{i \in S_{HOG\_AG+nonHOG\_AG}(d)} W_{AG\_ST}(cghi)} , \tag{8}$$

and the nonresponse-adjusted weight for hog-status ag-op respondents from weighting class d as

$$W_{HOG\_ST}(cdghi) = W_{AG\_ST}(cghi) * A_{HOG\_ST}(d) , \tag{9}$$

where

$S_{HOG\_AG+nonHOG\_AG}(d)$ is the set of ag-op records in weighting class d known to be hog operations (HOG_AG) as well as ag-ops known not to raise hogs (nonHOG_AG), and

$S_{AG}(d)$ is the set of class d sample records known to be agricultural operations.

The W_{HOG_ST} weight, as defined above, is applied only to records that belong to the set $S_{HOG\_AG+nonHOG\_AG}$, that is those ag-ops whose hog status is known.

Once a record has been identified as nonagricultural or not in operation during the calendar year, we know everything we need to know about the record. For the hog status determination (and all subsequent stages of data collection), these cases form a class of complete and unique respondents in the sense that we know their status as a hog operation (they cannot be a hog operation since all hog operations are ag-ops by definition) and we know how many hogs they raised (none).

Because nonag-ops have quite different survey responses from ag-ops, ag-op status must be used in defining weighting classes for hog status determination (and for all subsequent stages of nonresponse adjustment). Since nonag-ops are complete respondents by definition, their hog-status nonresponse adjustment factor A_{HOG_ST} is 1.0 and hence their hog-status nonresponse adjusted weight W_{HOG_ST} equals W_{AG_ST}.

As before, ag-ops whose hog status is unknown could be deleted from the data base at this point. However, it is usually preferable for history file purposes to retain them in the data base and to assign them adjustment factors and weights of zero for this stage and subsequent stages of weighting (as well as records whose ag-op status is unknown).

For the above adjustment to be worthwhile, we need to obtain hog status data for some or all of the nonresponding ag-ops (i.e., those with Box 921 codes of 11). Note that an ag-op is defined to be a hog operation for weighting purposes if hogs were present on the total acres operated on June 1 or during the previous quarter. Before going to the next step, I will again digress to mention changes needed in how hog status data are obtained in order to implement this weighting strategy.

Hog status is recorded for known ag-ops using the hog completion box (variable 499) which has codes

0 = Complete, had hogs June 1 or during previous quarter,

1 = Incomplete, has hogs,

2 = Incomplete, hog presence unknown, and

3 = Valid zero.

Code 0 is automatically assigned when the completion box is blank and positive responses are given for hogs. The remaining three codes are entered by the enumerator only when "all data are inaccessible or refused, or when valid zeros are reported for all items in a section." (NASS 1991a, p. 1303)

Interview respondents with hogs on June 1 or the previous quarter have the completion box automatically coded to 0; interview respondents without hogs on June 1 or the previous quarter are assigned to code 3. The phrasing of the completion box suggests that interview nonrespondents are assigned to code 1 when they have hogs on the total acres operated on the interview date. Presumably, nonrespondents are assigned code 3 when they do not have hogs on the interview date.

The implication is left that for nonrespondents the temporal definition of the presence-absence codes is different from that for respondents. This apparent discrepancy needs to be resolved, both in the questionnaire and the interviewer's manual.

The Agricultural Survey Interviewer's Manual does not address this problem specifically in its discussion of the hog completion box (NASS 1991a, p. 1708) or in its general discussion of the completion box (p. 1303). For code 3, the manual reads,

> Enter this code whenever it is known, either through interviews or other sources, that the operator has no positive data for the item of interest on the total acres operated.

To interpret this instruction as only referring to noninterview cases without hogs on June 1 and also without hogs the previous quarter may be too subtle for most enumerators to grasp. Also, code 1's instruction is unlikely to be interpreted to include noninterview cases without hogs now but with hogs on June 1 or the previous quarter. For code 1, the manual states,

> Through observation or other information, you know the operation has the item of interest on the total acres operated.

In addition to these issues, the March survey has a temporal problem not shared by the other quarters. Specifically, the hog portion of the questionnaire seeks data from all operations with hogs on March 1 or with hogs at some time during the previous quarter, or December 1 to February 28. However, question 1 excludes cases that had hogs during the previous December but have not and will not have hogs during the current year. Thus, question 1 could screen out some operations for which hog data are needed.

ADJUSTMENT FOR LOSS OF INTERVIEW DATA FROM KNOWN HOG OPERATIONS

The next step is to adjust for interview nonresponse for agricultural operations known to raise hogs. Again, I assume that the best weighting classes for this adjustment would be the poststrata g this time cross classified by hog status, with perhaps some collapsing to create classes of acceptable size. Let the letter "e" denote the classes used to calculate the adjustment factor and weight created below to adjust for interview non-response.

For known hog operations, the adjustment factor follows that of equation (8), or

$$A_{intvHOG\_AG}(e) = \frac{\sum\limits_{i \in S_{HOG\_AG}(e)} W_{HOG\_ST}(cdghi)}{\sum\limits_{i \in S_{intvHOG\_AG}(e)} W_{HOG\_ST}(cdghi)} \quad , \tag{10}$$

where $S_{intvHOG\_AG}(e)$ refers to the set of hog ag-ops from weighting class e who complete the hog portion of the questionnaire. The interview-nonresponse adjusted weight for hog-interview respondents is calculated as

$$W_{intvHOG\_AG}(cdeghi) = W_{HOG\_ST}(cdghi) * A_{intvHOG\_AG}(e) \quad , \tag{11}$$

which is attached to all interview respondents with hogs.

For nonhog ag-ops and nonag-ops, the responses to the hog questionnaire are radically different (they have no hogs by definition) so these records must be weighted separately. Further, both forms of nonhog operations have provided complete information for hogs and hence receive an adjustment factor, $A_{intvHOG\_AG}$, of 1.0 and a weight, $W_{intvHOG\_AG}$, equal to W_{HOG_ST}.

As before, hog ag-ops who fail to complete the hog portion of the questionnaire could be deleted from the data base at this point; for history file purposes it is usually better to instead assign them zero as their adjustment factor and weight (as should also be done for ag-op status nonrespondents and hog status nonrespondents).

ADJUSTMENT FOR MULTIPLICITY DETECTED FROM INTERVIEW DATA

The last stage of weighting is to adjust for frame multiplicity that can be detected for completed interviews only. When an interview is completed, data are obtained on operation name and the name(s) of owners and/or managers. These interview data items are compared with the list frame to identify whether other list frame records are linked to the sample record (besides those already identified in the pre-data collection duplicate check).

Let $\mathrm{MAF}_{\mathrm{intv}}(\mathrm{hi})$ be the total multiplicity as it is determined from all data sources, both the non-interview duplicate check and the operation, operator, and partner checks done for completed interviews only. Then the final analysis weight for hogs is the multiplicity-adjusted version of the interview-nonresponse adjusted weight or

$$W_{HOG\_ANAL}(cdeghi) = W_{intvHOG\_AG}(cdeghi) * \frac{MAF_{non\text{-}intv}(hi)}{MAF_{intv}(hi)} \ . \tag{12}$$

What this adjustment does is to remove the effect of the previous non-interview adjustment and instead substitute the full multiplicity as it is derived from all data sources.

The question comes up, "Couldn't we have waited to the end to do this adjustment for all cases?" "No" is the answer because the previous adjustment factors would not be calculated correctly. To prove this just suppose that all cases with multiple selection probabilities were nonrespondents. Waiting to the end to adjust here would mean that their multiplicity information would never be used since these cases have interview nonresponse-adjusted weights of zero.

The approach that I have outlined above -- to split the multiplicity adjustment into two steps depending upon the source of the multiplicity information -- is not implementable at the present time since the required data are not being collected. The relevant data items do not distinguish between multiplicity detected via non-interview data versus interview data. Further, sample cases belonging to lower priority strata and hence not eligible for data collection are assigned the same code as out of business cases for the 921 Box.

ASSUMPTIONS UNDERLYING NASS EXPANSION FACTORS

The above weighting procedures are sufficiently general that most expansion factors can be shown to be a special case. In this section, the assumptions are derived that form the basis for the current expansions used in NASS hog estimation.

To begin, both of the current direct expansions calculate a list adjustment factor (LAF) which is used for respondent data only and includes the multiplicity adjustment factor as well as a data adjustment factor (DAF) that is computationally used to zero out the data for out of scope operations. Let DAF(hi) be an indicator variable that equals 0 when the hi-th sample record is out of scope for data collection and 1 otherwise. Then, the LAF as currently defined can be expressed as:

$$LAF(hi) = [1 / MAF_{intv}(hi)] * DAF(hi) \quad . \tag{13}$$

Cochran (1977, pp.35-38) describes the use of such data adjustment factors to zero out of scope units (e. g., nonagricultural operations and nonhog operations) when estimating domain totals (hog producers are a domain of the population of agricultural and nonagricultural operations linked to the list frame).

The LAF is applied only after all other weighting steps are completed. Hence, the operational and adjusted hog direct expansions assume that frame multiplicity is being determined from interview data only and that no frame multiplicity is identified via an examination of non-interview data prior to data collection. That is, that

$$MAF_{non-intv} = 1.0 \tag{14}$$

for all cases and hence

$$W_{mult}(ghi) = W_{sam}(hi) \tag{15}$$

for all hi. Note that this assumption will be true only for States who prior to sample selection clean their complete frame of all duplication that can be detected from non-interview data.

The current expansions occur within the design strata and poststratification is not used. The assumption being made here is that the design strata used for sampling make adequate weighting classes. However, the QAS is now an integrated survey that includes many diverse commodities. Hence, its design strata are compromises between the competing needs of the various commodities and are not the optimal design strata for any one commodity. It is also doubtful that compromise design strata form the best classes for weighting purposes.

Before proceeding, I need to define these sample size components for each stratum h = 1, 2,..., H:

N(h) is the frame count of records for stratum h,

n(h) is the sample size selected from stratum h,

$n_{AG}(h)$ is the number of stratum h sampled records that survey operations identifies as ag-ops,

$n_{nonAG}(h)$ is the number of stratum h sampled records that survey operations identifies as nonag-ops,

$n_{HOG\_AG}(h)$ is the number of stratum h survey-identified ag-ops that are determined to be hog operations,

$n_{nonHOG\_AG}(h)$ is the number of stratum h survey-identified ag-ops that are determined to be non-hog operations,

$n_{intvHOG\_AG}(h)$ is the number of stratum h survey-identified hog operations completing the QAS interview, and

$n_{intv\text{-}nonHOG\_AG}(h)$ is the number of stratum h survey-identified nonhog ag-ops completing the QAS interview.

Under the two assumptions given earlier in this section -- that all multiplicity data are derived from interview results only and that the design strata form acceptable weighting classes -- and ignoring the LAF which is applied at the end, the ag-op status nonresponse-adjusted weight can then be written as:

$$W_{AG\_ST}(hi) = \frac{N(h)}{n(h)} * \frac{n(h)}{N_{AG}(h) + n_{nonAG}(h)} = \frac{N(h)}{n_{AG}(h) + n_{nonAG}(h)} , \qquad (16)$$

which also equals the final weight for nonag-ops. For ag-ops, the hog-status nonresponse-adjusted weight is

$$W_{HOG\_ST}(hi) = \frac{N(h)}{n_{AG}(h) + n_{nonAG}(h)} * \frac{n_{AG}(h)}{n_{HOG\_AG}(h) + n_{nonHOG\_AG}(h)} , \qquad (17)$$

which also equals the final weight for nonhog ag-ops (excluding the LAF).

For responding hog ag-ops, the interview nonresponse-adjusted weight is

$$W_{intvHOG\_AG}(hi) = \frac{N(h)}{n_{AG}(h) + n_{nonAG}(h)} * \frac{n_{AG}(h)}{n_{HOG\_AG}(h) + n_{nonHOG\_AG}(h)} * \frac{n_{HOG\_AG}(h)}{n_{intvHOG\_AG}(h)} . \qquad (18)$$

Neither of the two expansion factors being used for hog estimation uses the above equations.

The oldest hog estimator still computed by NASS is the "operational" estimator, which uses as its weight the "reweighted direct expansion factor" described by Kott (1990, p. 11). This expansion factor is defined for each stratum as the record count N(h) divided by the number of "usables" where usables are defined to be interviewed hog producers (interviewed "positives") and interviewed and noninterviewed cases that have zero hogs ("zeros").

The operational estimator can be shown to use the following weight for "usable" records:

$$W_{OP}(hi) = \frac{N(h)}{n_{usables}(h)} \tag{19}$$

where

$n_{usables}(h)$ is the stratum h sample count of identified nonag-ops, identified nonhog ag-ops and hog ag-ops with interview data, that is,

$$n_{usables}(h) = n_{intvHOG\_AG}(h) + n_{nonHOG\_AG}(h) + n_{nonAG}(h) \ . \tag{20}$$

There really is no justifiable model for this expansion factor, which may be why an alternative expansion factor (the "adjusted" estimator described later) was developed for hog estimation. The problem is that the operational estimator uses the hog-status completion box data for non-hog operators but not for hog operators. By including only one type of interview nonrespondent -- the non-hog "zeros," the operational estimator is biased downward in the sense that it overrepresents non-hog producers.

The operational estimator appears to be a corruption of the following alternative estimator:

$$W_{OP\_ALT}(hi) = \frac{N(h)}{n_{intvHOG\_AG}(h) + n_{intv\text{-}nonHOG\_AG}(h) + n_{nonAG}(h)} \ . \tag{21}$$

This expansion factor can be derived from the above equations (16), (17), and (18) by assuming that valid partial data are not attainable for nonrespondents [that is, that

$$n_{intvHOG\_AG}(h) = n_{HOG\_AG}(h) \tag{22}$$

and

$$n_{intv\text{-}nonHOG\_AG}(h) = n_{nonHOG\_AG}(h) \ ] \tag{23}$$

or if attainable (which it is) is not sufficiently accurate to be used. That is, this estimator proceeds as if valid ag-op status results cannot be obtained for any nonrespondents (e.g., no 11 codes for the 921 box) and further that all cases respond who are known to raise hogs (no 1 codes for the hog completion box). Rather than use any partial data that are obtained, this estimator assumes that each stratum's responding "positives" (regardless of the commodity they produce) plus known non-ag ops can be regarded as a simple random subsample from the full stratum sample.

Since 1986, hog commodity reports have also used the "adjusted" hog estimator to establish the statistics it reports. Kott (1990, p. 12-14) refers to this estimator as a "presence estimator". The expansion factor for each stratum is defined as the record count N(h) times the proportion of records known to be associated with hog producers (all "positives" for hogs regardless of response status) among all records who have hog status known ("zeros" for hogs regardless of agricultural status and all "positives" for hogs) and divided by the number of interviewed hog producers (responding "positives").

Using the notation given above, the adjusted estimator can be shown to use the following weight for all "usable" records:

$$W_{ADJ}(hi) = \frac{N(h)}{n_{HOG\_AG}(h) + n_{nonHOG\_AG}(h) + n_{nonAG}(h)} * \frac{n_{HOG\_AG}(h)}{n_{intvHOG\_AG}(h)} \qquad (24)$$

(including nonag-ops and nonhog ag-ops). There does not appear to be a model that supports this estimator for all usable records. If hog status is known for all cases known to be ag-ops (e.g., there are no records with a 921 code of 11 and a hog completion box code of 2) or the ag-op status data are not sufficiently accurate when hog status is unknown, then the equation (18) weight for hog ag-op respondents simplifies to equation (24), since

$$n_{HOG\_AG}(h) + n_{nonHOG\_AG}(h) = n_{AG}(h) \ . \qquad (25)$$

However, equation 24 will not yield the appropriate equation (16) and (17) weights for nonag-ops and nonhog ag-ops, respectively, as they do not contain the last term in equation (18). That is, for these "zeros," the adjusted weight is defined as:

$$W_{ADJ}(h1i) = \frac{N(h)}{n_{HOG\_AG}(h) + n_{nonHOG\_AG}(h) + n_{nonAG}(h)} \ . \qquad (26)$$

Let us assume for a moment that equation (24) reflects the desired model for respondents and that the weights for nonag-ops and nonhog ag-ops are being calculated in error. Then, the underlying model is one that assumes:

(1) that a stratum's "positives" (both responding and nonresponding identified hog producers) plus its "zeros" (nonhog producers and nonag-ops) can be regarded as a simple random subsample of the stratum sample, and

(2) that hog respondents are a simple random subsample of each stratum's identified hog ag-op sample.

This is essentially the model proposed by Crank (1979).

Assumption (1) is problematic as it fails to recognize that ag-op status may be obtained without obtaining hog status. Suppose for instance that all nonrespondents are known to be ag-ops (i.e., no 12 codes are assigned) and that all ag-ops are hog producers. Forcing Assumption (1) result in underestimates of hogs as some nonrespondents are imputed to be non-ag "zeros" when none are. Admittedly, this is a contrived example but it illustrates that the underlying model is deficient. The adjusted estimator does not distinguish between nonrespondents whose ag-op status is unknown (and hence may be nonag-op zeros, nonhog ag-op zeros or hog ag-op positives) and nonrespondents who are known to be agricultural but whose hog status is unknown (who can only be nonhog ag-op zeros or hog ag-op positives).

ESTIMATION USING ANALYSIS WEIGHTS

The alternative weighting procedure that is outlined in this paper should produce more accurate estimates for hogs. From the NASS prospective, though, this revised procedure has the disadvantage that the resultant reweighted data cannot presently be analyzed with the special purpose Survey Processing System software NASS currently uses for estimation (Kott 1990). In this section, I describe the NASS analysis approach underlying the operational and adjusted estimators and then note the changes needed for the revised weighting strategy proposed in this paper.

When nonresponse occurs, assumptions must be made about the nonresponse mechanism in order to analyze the data. Modeling nonresponse as if it were due to a sampling operation (i.e., as if one had randomly selected the operations to respond) means that a response model-based variance estimate can be derived. The quality of such a variance estimate is directly dependent upon the adequacy of the model for the nonresponse mechanism. This section summarizes the sampling mechanism that is being used to model nonresponse.

Estimation for the Operational Estimator

The operational estimator can be regarded as assuming that the "usable" hog interviews (i.e., nonag-op and nonhog ag-op "zeros" and interviewed hog ag-op "positives") are a simple random subsample of each stratum's sample. Since a simple random subsample of a simple random sample is a simple random sample, this assumption implies that the "usable" cases can be modeled as being derived a stratified simple random sample. (As noted earlier, this is not a reasonable assumption to make for QAS hog estimation.)

Under the stratified simple random sampling assumption for usable interviews, the operational estimate of the total is computed as:

$$\hat{Y} = \sum_{h=1}^{H} \hat{Y}(h) = \sum_{h=1}^{H} \sum_{i=1}^{n(h)} z_{OP}(hi) \tag{27}$$

where

$$z_{OP}(hi) = [W_{OP}(hi) \; / \; MAF_{intv}(hi)] * DAF(hi) * Y(hi) \tag{28}$$

and

$\hat{Y}$ is the estimated population total,

$\hat{Y}(h)$ is the estimated stratum h total, and

Y(hi) is the count of hogs possessed by the hi-th case.

This can be seen to be equivalent to equation (6) of Kott (1990, p.11) when one recognizes that nonrespondents have zero as their operational weight. This expression is the same as the Taylor series linearized value used by SUDAAN in calculating totals (RTI, no date, p. A-20).

For variance estimation for the operational estimator of the total, one applies domain estimation approaches since hog producers constitute a subpopulation of each stratum. (See for instance Cochran 1977, pp. 35-38.) These results can be shown to produce the following estimate of the variance:

$$Var(\hat{Y}) = \sum_{h=1}^{H} \frac{N(h) - n_{usables}(h)}{N(h)} * n_{usables}(h) * s^2(h) \; , \tag{29}$$

where

$$s^2(h) = \frac{\sum_{i=1}^{n_{usables}(h)} [z_{OP}(hi) - \bar{z}_{OP}(h+)]^2}{n_{usables}(h) - 1} \tag{30}$$

and

$$\bar{z}_{OP}(h+) = \frac{\sum_{i=1}^{n_{usables}(h)} z_{OP}(hi)}{n_{usables}(h)} . \tag{31}$$

This estimator of the variance is equivalent to Kott's equation (8) for the variance of the reweighted direct expansion estimator when $n_{usables}(h)$ is greater than one (Kott 1990, p. 13).

Rather than collapse strata that contain only one usable interview, NASS's Survey Processing System software retains the original strata and uses the squared value $[Y^2(hi)]$ to estimate the stratum variance contribution when only one usable stratum interview is obtained. (This is a conservative procedure as zero is being implicitly used as the stratum mean in calculating the stratum variance.) When no usable interview is obtained, the stratum is effectively excluded from estimation of the total and its variance. These two situations point out another deficiency of the operational estimator. As with strata, weighting classes should contain a minimum of 20 respondents and average at least 30 to 50 respondents (Cox 1991). NASS estimation procedures contain no fallback procedures to collapse strata (which form the weighting classes for the adjustment) when the sample size for some strata (in terms of usable interviews) becomes too small for use as weighting classes.

As an aside at this point, note that NASS replaces missing data for extreme operators (operators who raise a large percentage of the hogs in their State) with logically imputed values, which may be derived from interviewer observations, expert judgement, frame control data or some other source (USDA 1991b). NASS includes these imputed data for extreme operators in equations (27) and (28) for estimating totals and in equations (29) and (30) for estimating the variances associated with these estimated totals. Since all missing data are replaced via imputation and hence no nonresponse adjustments are needed, the weight $W_{OP}(hi)$ for the extreme operator is simply the sampling weight W_{sam}. That is, the estimator for the total treats each extreme operator as having "usable" data, regardless of their response status. This is a reasonable approach as the data are logically rather than statistically imputed; imputation error is best regarded as contributing to bias in the survey estimate of the total.

With the growing complexity of sample design and nonresponse adjustment procedures, it is now commonplace to develop analysis weights in a separate step that then allows the use of general purpose analysis software. Equations (29) and (30) are equivalent to the SUDAAN variance expression for estimation of totals under without replacement simple random sampling (RTI, no date, pp. A-7 to A-9). SUDAAN proceeds by calculating a z-value for each record equivalent to equation (28) based upon the weight variable and a domain indicator variable found on each data record. Thus if one inputs $[W_{OP}(hi) / MAF(hi)]$ as the weight, DAF(hi) as the domain indicator variable, and Y(hi) as the variate of interest, then SUDAAN yields the same operational estimator for the total as the special purpose USDA software except when some design strata contain only one respondent. (When a design stratum has only one respondent, SUDAAN uses the overall sample mean in calculating the stratum variance instead of the zero NASS's summary system uses.)

Estimation for the Adjusted Estimator

As noted earlier, the adjusted estimator treats ag-op status and hog status as one entity in data collection terms; either both pieces of information are assumed known or neither is known. In effect, the adjusted estimator proceeds as if the associated nonresponse mechanism is such that (1) the cases providing both ag-op and hog status data are a simple random subsample of the original stratum sample, and (2) those hog operations completing the hog interview are a simple random sample of the survey-identified hog operations. For the purposes of deriving the variance estimate, another assumption is made which is (3) the sample size for each second-phase stratum is determined as a fixed proportion (the response rate) of the first-phase sample.

Returning to the first assumption, we note again that a simple random subsample of a simple random sample is a simple random sample from the population. The first phase of sampling under this model then is the "selection" of the records that respond to both agricultural status and hog status information. This "sample" consists of the identified nonag-ops, the identified nonhog ag-ops, and the identified hog ag-ops. The stratum h sample size for the first phase of sampling under this model is regarded as n'(h) where

$$n'(h) = n_{nonAG}(h) + n_{nonHOG\_AG}(h) + n_{HOG\_AG}(h) \quad . \tag{32}$$

The modeling then regards the first-phase stratum sample as being partitioned into two second-phase strata. One of these strata is composed of the identified nonag-ops and nonhog ag-ops, which are regarded as all responding to the second phase as their hog data are known (they have none). The other stratum is composed of the identified hog ag-ops, who are regarded under the nonresponse model as being subsampled to determine which will provide hog data. Let l=1 designate those with hogs and l=2 those without hogs. The model proceeds as if the stratum h sample of n'(h) is substratified into a stratum of hog operations n'(h1) of size

$$n'(h1) = n_{HOG\_AG}(h)$$

from which a subsample $n'_{intv}(h1)$ is designated to respond to the interview. Using the notation given earlier,

$$n'_{intv}(h1) = n_{intvHOG\_AG}(h) \quad .$$

For nonhog operations, the model proceeds as if this substratum were sampled with certainty, that is

$$n'(h2) = n'_{intv}(h2) = n_{nonAG}(h) + n_{nonHOG\_AG}(h) \quad .$$

Under these assumptions listed above about the nonresponse mechanism, the resultant sample of interviewed hog operations can be modeled as having resulted from a two-phase sample design where the second-phase strata are nested within the first-phase strata, simple random sampling is used at both stages for sampling within strata, and the second-phase sample sizes are fixed proportions of the first-phase samples.

Within each stratum then, we have the classic double sampling for stratification procedures described by Cochran (1977, pp. 327-358). Let j=1 reference the second-phase stratum of identified hog operations and j=2 the second-phase stratum of identified "zeros" within each first-phase stratum (i.e., the nonag-ops and nonhog ag-ops). Using this notation, the sample estimate of the population total can be expressed as

$$\hat{Y} = \sum_{h=1}^{H} \sum_{j=1}^{2} \hat{Y}(hj)$$

$$= \sum_{h=1}^{H} \sum_{j=1}^{2} \sum_{i=1}^{n'_{intv}(hj)} \frac{W_{ADJ}(hji)}{MAF_{intv}(hji)} * DAF(hji) * Y(hji) \qquad (33)$$

where

$\hat{Y}$ is the estimated population total,

$\hat{Y}(hj)$ is the estimated second-phase stratum hj total, and

Y(hji) is the count of hogs possessed by the hji-th case.

For variance estimation purposes, equation (33) can be rewritten as

$$\hat{Y} = \sum_{h=1}^{H} N(h) \sum_{j=1}^{2} w(hj) \left[\sum_{i=1}^{n'_{intv}(hj)} \frac{z'(hji)}{n(hj)} \right] \qquad (34)$$

where

$$z'(hji) = Y(hji) * DAF(hji) \;/\; MAF_{intv}(hji) \qquad (35)$$

$$w(hj) = n'(hj) \;/\; n'(h)$$

This formulation emphasizes that each stratum h sample can be regarded as having been derived from double sampling for stratification. The associated variance can then be derived following formula (12.32) of Cochran (1977, p. 334) as:

$$Var(\hat{Y}) = \sum_{h=1}^{H} \left[N(h)^2 \, \frac{N_h - 1}{N_h} * \sum_{j=1}^{2} \left[\frac{n'(hj)-1}{n'(h)-1} - \frac{n'_{intv}(hj)-1}{N_h - 1} \right] \frac{w(hj) \;\; s^2(hj)}{n'_{intv}(hj)} \right. \\ + \sum_{h=1}^{H} \left[N(h)^2 \, \frac{[N(h)-n'(h)]}{N(h)\,[n'(h)-1]} \sum_{j=1}^{2} w(hj) \;\; [\overline{z}'(hj)-\overline{z}'(h)]^2 \right] \tag{36}$$

where

$$s^2(hj) = \frac{\sum_{i=1}^{n'_{intv}(hj)} [z'(hji)-\overline{z}'(hj+)]^2}{n'_{intv}(hj)-1} \tag{37}$$

Recognizing that $s^2(h2) = 0$ leads to some simplification of the variance estimate. This expression is not exactly equivalent to the formula that NASS uses to derive the variance estimate (Kott, 1990).

Estimation for the Revised Estimator

Describing the underlying model for nonresponse for the revised estimator is more difficult due to its use of weighting classes that are not necessarily nested within the design strata. Basically, the nonresponse mechanism may be conceptualized as resulting in a four-phase sample design. The first phase under this model is the actual sampling that occurs within design strata.

The second phase begins with a "restratification" based upon hog characteristics only and then post-stratification adjustment of the sampling weights. Within these post-strata, the model assumes that the units providing agricultural status are a random subsample of the full poststratum sample. This random subsampling is in proportion to the unit's sampling weight.

The third phase begins by restatifying the units, generally by the second-phase poststrata crossed with the ag-status indicator. The non-ags are regarded as being sampled with certainty (as their hog data are universally known to be zero) while the ag-ops are regarded as being subsampled to determine which will provide hog status information, with the subsampling this time in proportion to their ag-op status nonresponse adjusted weight.

The fourth and final phase would be regarded as again restratifying the units, this time by second-phase poststrata crossed with ag-op status and hog-status. The non-ags and nonhog operations are again regarded as sampled with certainty while the hog operations are treated as if they were subsampled to decide who will provide hog data, with the subsampling proportional to their hog-status adjusted weight.

Further research is needed to decide the most effective approach for calculating a variance for the revised estimator. The complex nature of the assumed nonresponse mechanism makes derivation of a variance estimate quite complex. The resultant formula when derived might be too complex to program for production use. Most of the complexity noted above would disappear if design strata were used as the basis for classing. The need to use weighting classes that differ from the design strata, however, is the result of the integration of diverse commodity samples within one design resulting in design strata that as not as informative as they could be. Although calculation of a model-based variance estimate is easier when design strata are used to form weighting classes, the estimator may not be as effective in reducing nonresponse bias.

Till such research is completed, there is an approach that should yield a reasonable approximation for variance of the the revised estimator under this nonresponse model. This is to use the formulas for the operational estimator (27 and 29) but substitute the revised weight. Simulations by Jones and Chromy found that this approach produced variance approximations that were about five percent less than the appropriate estimate. In their simulations Jones and Chromy used weighting classes that were extensively interwoven with the design strata. In the case of the QAS, the most effective strategy for classing would probably begin with the definitions used to define hog strata. Except for cases that were classified elsewhere due to the priority classification scheme, most operations would be expected to have the same weighting class and design strata.

CONCLUDING REMARKS

NASS should reconsider the weighting procedures being used in estimation. I recommend that NASS adopt poststratification and the stepwise approach outlined above for deriving nonresponse adjustments with a shift in emphasis to one of determining the most appropriate assumptions to use in modeling nonresponse. Once these assumptions are decided, the expansion factor can be derived by following the steps described in this paper.

Consider the following questions as examples of the issues to addressed:

> How extensive are the follow-up and tracing procedures for "inaccessibles?" Can inaccessibles or a definable subgroup of them be safely assumed to be all out of business?
>
> Is it valid to assume that all refusals are agricultural operations? Why not?
>
> Under what circumstances (if any) should control data be used to decide ag-op status and/or hog status?
>
> Are there modifications to NASS data collection procedures that would increase response for one or more of these components of response (e.g., ag-op status response, hog status response, or interview response)?
>
> Are different models needed for nonresponse depending upon the methods used by the State for follow-up and conversion of initial nonrespondents?

Answers to questions such as these are needed to evaluate the model implied by the current estimation approaches and how appropriate the assumed model is for explaining the nonresponse encountered by the survey.

In doing a survey, I usually have a good idea about the answers for questions such as these. In this case, I do not have sufficient experience to know. Other Headquarters staff may have a similar problem since not one but 45 different State offices implement the study, each perhaps using slightly different procedures. A questionnaire to State statisticians might solicit the information needed to determine the appropriate model for nonresponse.

Another issue is whether there are eligibility data not presently being used in weighting such as control data and/or data that the enumerators could but are not presently collecting. For instance, NASS could consider setting targets for the response rates for partial data items such as ag-op status and hog presence-absence. For the January 1991 Cattle Survey, for instance, the percentage of nonresponding operations where it could not be determined whether cattle were present or absent averaged 66.4 percent across States and ranged from a low of 5 percent unknown for Nebraska to over 90 percent not determined for 5 of the States (Vogel 1991).

Clearly, some States may have room for improvement with respect to presence-absence determination.

As a first step in this regard, I suggest that stepwise response rates be calculated for past QAS surveys using the historical data that are available. These calculations would mimic the steps of weighting described above, that is, (1) the response rate to ag-op status determination, (2) the response rate to hog status determination among known ag-ops, (3) the response rate to the interview from known hog operations, and (4) the overall response rate (e.g., the product of these three factors). These calculations should be done for the nation as a whole and then by State, type of agriculture (e.g., hog EOs, wheat EOs, etc.), and State by type of agriculture (e.g., stratum).

The quality of the information being derived on ag-op status and hog status also needs to be assessed as well to decide how (and whether) these data should be used. Reinterview results should prove useful in this regard. Stability over time might be studied by comparing over years the proportion of ag-ops among list frame records and the proportion of hog operations among known ag-ops.

NASS expansion factors assume that the design strata form the most appropriate classes for nonresponse adjustment. The choice of variables for defining weighting classes should be reexamined to verify that the design strata are indeed optimal for modeling the response mechanism. NASS's increasing use of integrated surveys is leading to the use of compromise strata that tend to be less than optimal for any particular survey. Such compromise strata are unlikely to form the best weighting classes.

Even when the design strata are optimal to define weighting classes, they may need to be collapsed to obtain classes of sufficient size or to prevent extreme adjustments. I do not see any evidence that weighting class sizes and adjustment factors are being monitored. Since strata now form the weighting classes, it might be of benefit to use historical data to look at the strata arrayed by number of respondents per stratum and size of the adjustment factor. Note that stratum size is immaterial for strata with 100% response; "complete" response occurs for extreme operator strata where all missing data are imputed.

Finally, it should be noted that while departing from the use of design strata as weighting classes is probably the path to the most accurate commodity estimate, it does make variance estimation more complicated. This is another issue that needs further investigation. The difficulties in variance estimation should not deter or delay NASS, however, in implementing a revised weighting strategy designed to reduce nonresponse bias. After all, a more accurate point estimate for hogs is more desirable -- even with a slightly biased variance estimate -- than a biased point estimate for hogs with an unbiased estimate of the variance.

REFERENCES

Bosecker, R. R. (1987). *The Quarterly Agricultural Surveys*, Washington, DC: National Agricultural Statistics Service, U. S. Department of Agriculture.

Cochran, William G. (1977). *Sampling Techniques*, New York: John Wiley & Sons.

Cox, Brenda G. (1991). *Weighting Survey Data for Analysis*, Short course documentation. Research Triangle Park, NC: Research Triangle Institute.

Crank, Keith N. (1979). *The Use of Current Partial Information to Adjust for Nonrespondents*, Washington, DC: Economics, Statistics, and Cooperatives Service, U. S. Department of Agriculture.

Dalenius, T. and Hodges, J. L., Jr. (1959). "Minimum Variance Stratification," *Journal of the American Statistical Association*, 54, pp. 88-101.

Holt, D. and T. M. F. Smith (1979). "Post Stratification," *Journal of the Royal Statistical Society*, Soc. A, 142, Part 1, pp. 33-46.

Kott, Phillip S. (1990). *Mathematical Formulae for the 1989 Survey Processing System*, National Agricultural Statistics Service Staff Report No. SRB-90-08, Washington, DC: U. S. Department of Agriculture.

National Agricultural Statistics Service (1991a). *Agricultural Surveys: Interviewer's Manual*, Washington, DC: U. S. Department of Agriculture.

National Agricultural Statistics Service (1991b). *Agricultural Surveys: Supervising and Editing Manual*, Washington, DC: U. S. Department of Agriculture.

Vogel, Frederic A. (1991). "Quarterly Agricultural Surveys," Internal Memorandum to Cynthia Clark dated March 20, Washington, DC: National Agricultural Statistics Service, U. S. Department of Agriculture.