United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Research Division

SRB Research Report
Number SRB-94-01

January 1994

Estimating List Frame Duplication

Orrin Musser
James W. Mergerson

ESTIMATING LIST FRAME DUPLICATION, by Orrin Musser, Survey Quality Research Section, Survey Research Branch and James W. Mergerson, Technology Research Section, Survey Technology Branch, Research Division, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, DC 20250-2000, January 1994, Report No. SRB-94-01.

ABSTRACT

Duplication in a survey organization's list sampling frame is potentially a very serious problem. Differences between the National Agricultural Statistics Service's multiple frame and area frame direct expansions may be due in part to duplication of list population elements. Maintaining a large list without any duplication is not feasible since the necessary resolution and maintenance efforts are prohibitive in terms of time and cost. However, it is essential that duplication be measured, monitored and minimized over time.

This paper presents a methodology for measuring the percent of duplication present within 1992 list sampling frames in two States (North Carolina and Ohio). Based on four possible linkage variables (Social Security number, Employer Identification number, telephone number and linkage cross reference number) sample level records were compared to list sampling frame records to determine linkage groups. Each linkage group was resolved to determine the number of duplicate records associated with each sampled record. These counts were used to determine estimates of the percent of duplication present in 1992 list sampling frames in the two States. The estimated percentages of duplication were very low. The estimates of the percent of duplication among all records classified for the Agricultural Survey were less than 0.5 percent in both North Carolina and Ohio.

KEY WORDS

Record linkage; Data adjustment factor; Duplication check; Survey.

ACKNOWLEDGEMENTS

The authors would like to thank the North Carolina and Ohio State Statistical offices for their support of this project. A special thanks goes to Carol House for her review of the report, and to Jeanne McCarthy-Kersey and Rex Patterson of the List Frame Section for their assistance in reviewing and resolving linkage groups. Thanks to management (Ron Bosecker, George Hanuschak, Jim Davies, Roberta Pense, and Dale Atkinson) for their support of this project.

TABLE OF CONTENTS

SUMMARY

The focus of this project was evaluating duplication on the NASS List Sampling Frame, and the current Agency process to deal with it. More specifically, the goals of this project included:

- Developing a methodology for the estimation of duplication, and applying it in two States, North Carolina and Ohio.
- Assessing the effectiveness of current duplication handling, and its impact on survey indications.
- Assessing the impact of incorrect duplication handling on survey indications.
- Better understanding the causes of duplication, and suggesting strategies to avoid duplication.
- Suggesting improvements in the current duplication checking procedures.

Each spring, after all major survey list samples have been drawn for the year, each State Statistical Office (SSO) receives a listing of potential duplicate record groups, commonly referred to as the duplication check printout. SSO staff attempt to resolve these groups and determine correct survey action to adjust for duplication, when found.

Shortly after the 1992 June Agricultural Survey, a team of two persons went to each of the two State Offices chosen for this study (North Carolina and Ohio). These teams performed a thorough review of the State's duplication check decisions and actions using various available resources. These resources included State office staff, notes from the State office copy of the duplication check printout, June Agricultural Survey data listings, the real-time mail and maintenance system and telephone calls to the farm operator. Questionnaires were also checked for correct coding.

For each potential duplicate group, the following were recorded:

- Number of records linked to each sampled record. This provided data necessary to obtain estimates of list duplication percentages.
- Correct survey codings for each sampled record. These were compared to the actual survey codings in the four quarters of the Agricultural Survey to assess the effectiveness of the State's duplication resolution process over an entire survey cycle.
- Reasons for duplication and other comments to give some insights which could improve the current process of list frame maintenance and duplication resolution.

Estimates of the percentage of duplication among all records classified for the Agricultural Survey are 0.21 percent (with a standard error of 0.06 percent) in North Carolina and 0.19 percent (with a standard error of 0.04 percent) in Ohio. These percentages are based on duplication found only through the duplication check record linkage program which performs record linkage by exact matching on Social Security number, Employer Identification number, telephone number, and link cross reference number. Since some of the records classified in these States have no possibility of matching, some duplication may not be detected. Adjusted estimates of the percentage of duplication based on the proportion of conforming pairs of records among all pairwise comparisons of records are presented. However, the adjustment procedure assumes that records with and without the requisite matching data are equally likely to be duplicates.

A more thorough record linkage using additional variables would be helpful in improving our estimates of the true percentage of duplication. The Survey Technology Branch is currently evaluating a DOS version of a record linkage system which shows much promise in providing efficient and thorough matching. The acquisition and implementation of such a product could potentially provide more efficiency in the duplication check process.

An extremely important aspect of duplication handling is the correct coding of the questionnaire. When the SSO list frame statistician determines that duplication exists, comments which provide detailed coding instructions are usually posted to the list frame. These comments are then printed and attached to the questionnaire at survey time, and must be entered correctly for the duplication adjustment process to work. These codings were checked in each of the quarterly surveys to gauge how well the system works over the entire Agricultural Survey cycle. The results of this quality check indicated that there is room for improvement in this follow through process. System changes are recommended to improve the process and catch errors during survey edits.

To measure the effects of duplication on survey indications, we summarized data for several important variables from the June Agricultural Survey under two scenarios. In the first "worst case scenario," we assumed that none of the duplication found on the frame had been detected. If data were collected for only one of several sampled records in a linkage group, those data were copied to the other records, and the data adjustment factors were set to one. If data were not collected at all for a sampled record (due to duplication with a record in a higher stratum that was not sampled) then we imputed stratum means for that record. In the second "best case scenario," we took the opposite approach and assumed that the duplication resolution and survey action had been performed "correctly," and compared survey indications to those obtained in the actual survey. In both cases, the effect on survey indications was small, due to the low duplication rates.

INTRODUCTION

The Agricultural Survey Program is a yearly cycle of surveys conducted by The National Agricultural Statistics Service (NASS) to provide inventory and production estimates for various agricultural commodities. This program includes the Quarterly Agricultural Surveys which provide estimates for crop production, grain stocks, and hog inventories. The target population for these surveys is the U.S. population of farms, defined as all establishments that sold or would normally have sold more than $1,000 of agricultural products during the year. These surveys use a multiple frame sampling design with a list frame of farms and farm operators, and a supplemental area frame of all land in the U.S. The list frame provides very cost efficient sampling, and good coverage of medium and large farms. The area frame provides complete coverage of all farms and also a measure of the incompleteness of the list.

NASS devotes considerable resources in its SSO's to maintain the list frame. Because the list frame provides more efficient sampling than the area frame for most commodities, there is a strong Agency commitment to maintain current and accurate control data and to increase list coverage. SSO's update the list frame on an ongoing basis, and attempt to find agricultural lists and other sources of farm operations for list building. This activity has the potential to add duplication to the list frame, which can be a serious problem for survey estimation.

Motivation for this project came from observations by some Agency statisticians that multiple frame survey indications for some commodities in some States have historically been higher than those obtained solely from the area frame samples. One possible explanation for the observed level differences is undetected duplication in the list frame. The purpose of this research study was to develop methodology to estimate the percentage of duplication present on list sampling frames and to apply this methodology to estimating duplication in a couple of States.

BACKGROUND

Duplication Control at NASS

Duplication control has been an important part of NASS's list maintenance activities since the late 1970's when the States' current list frames were built. There are three main methods of duplication detection: list frame resolution, on-line searches, and the annual duplication check process.

The first method, list frame resolution, is a process in which every record on the list frame is matched against every other record in an attempt to remove all duplication from the frame. This matching is performed by the Record Linkage Sub-System (RECLSS), developed by NASS in the late 1970's. RECLSS resides on the centrally located mainframe computer system used by NASS on a contract basis. It is capable of performing very thorough record linkage using name and address standardization and matching, but is not "user friendly" and requires experienced users to achieve good results. Because of the computer intensive nature of record linkage, the

cost of running the entire resolution process on the mainframe is high, in the $300-$600 range for a single State depending on the size of the list frame. Typically a full resolution requires a large amount of manual review through on-line searches, and this adds significantly to the total cost of list frame resolution. For these reasons, full resolution is not performed on an annual basis, but is performed periodically, usually every three to five years. During the period 1990-1992, the list frame sampling unit was changed from farm operation to farm operator ("operator-dominant conversion") and full resolution was run for each State as part of this conversion. North Carolina went through this process in 1990 and Ohio in 1991.

The second method of duplication control uses on-line searches of a State's list frame database to search for duplicates or to review potential duplicates found by the other methods. List maintenance is carried out in the State Statistical Offices (SSO's) using the Real Time Mail and Maintenance System (RTMMS), the database of all list records which also resides on the mainframe. Selected staff in each SSO use this system to perform list maintenance activities including resolution of potential list duplication, using on-line searches. When adding a new farm operator to the list, searches can be performed based on name, zip code, phone number, Social Security number, etc. to make sure that this operator is not already on the list. This system is also used to resolve output from list frame resolution and the NASS duplication check process, described in detail in the next section. This system was implemented in the late 1980's, and has been a very useful tool in prevention and reduction of duplication.

The third method of duplication control used by NASS is the "duplication check" process in which sampled records for each of the major surveys are matched against all other list frame records which were classified for that survey. Matching is performed by a much simpler record linkage program (than RECLSS), in which two records match if any of the fields selected as "matching variables" are exact matches. This also resides on the central mainframe, but is much less expensive to run, and is designed to produce a manageable volume of output which can be resolved prior to the survey. The duplication check process is the focus of this research project and is described in detail in the next section.

The NASS Duplication Check Process

The NASS survey cycle begins with list classification, a process in which all records on each State's list frame are classified into ordered design strata for each of the major surveys (Quarterly Agricultural Surveys, Cattle and Sheep Survey, Farm Cost and Returns Survey and Agricultural Labor Survey), based on control data on the frame. While a farm operator may qualify for inclusion in several strata, he will be classified only for the highest numbered stratum, which is of highest priority in the NASS ordered stratification. List classification is followed by sample selection, in which stratified random samples are chosen for the major surveys.

Sample selection is followed by the duplication check process, in which records classified for major surveys are matched against all active records on the list frame. The duplication check program matches records based on four list frame variables: Social Security number (SSN),

Employer Identification number (EIN), telephone number and link cross reference number. A match occurs if any one of these fields matches. Records with none of these matching variables are rare, with 0.9 percent in Ohio, and 2.8 percent in North Carolina.

The output of the duplication check program consists of two listings of "linkage groups" (sets of records that "link" or match). The primary listing consists of all linkage groups which contain at least one sampled record. SSO staff check each of these, prior to the survey, to determine if the records which matched are in fact duplicates. Often records which match only on telephone number actually represent separate operations. In these two States, less than one third of the linkage groups contained actual duplicate records. When duplication is found, questionnaire coding adjusts for the duplication and "Record Status" codes are modified on the list frame to correct the problem for the future. A secondary listing of linkage groups which contain at least one classified record, but no sampled records is also provided for the States to resolve during the year as time permits.

When duplication is found prior to or during the survey, the questionnaire is coded as follows for individual operators:

1. The "reporting unit" (item 921) is set to the total number of times the operator is found on the list. The acceptable range for duplication coding is 2-5.

2. The strata codes of the duplicate records are recorded in the "partner strata" fields (items 925-928), with a maximum of four duplicate record strata codes that may be recorded.

While managed operations and large and/or complex operations are treated somewhat differently, questionnaire coding instructions are designed to achieve consistent results for these operations.

Strategies for Estimation in the Presence of Duplication

When each population element is represented by exactly one list unit, the probability that a given element is in the sample is just the probability that its associated list unit is selected. List duplication occurs when some population elements are represented by, or "linked" to, more than one unit on the list frame. When there is duplication the selection probabilities are not equal, and if we do not make adjustments, our estimators will have an upward bias. Lessler and Kalsbeek (1992, p.75) quantify this bias for estimates of totals under simple random sampling. Suppose we divide the population into domains based on the number of times each population element is represented on the list frame, with domain k being all population elements which appear k times. They show that if the population means of a variable are equal for these domains then the relative bias is equal to the ratio of the number of duplicate frame records to the true population size.

There are several strategies that are commonly used for conducting surveys in the presence of list duplication. One of these is to establish a "unique counting rule" which links each

population element to a single list frame unit. The population element is included in the sample only if the frame unit identified by the rule is selected. An example of this would be a rule that a population element is included in a sample only if the frame unit with the largest identification number among the duplicate records is selected. If duplication is detected after data collection, then some data may be lost.

Another strategy is to use a weight or data adjustment factor to adjust the expansion factor of records which have duplicates on the frame, and thus are more likely to be sampled. If a population element k appears on the sampling frame M_k times, then when sampled, the data are multiplied by $1/M_k$. Even if the same population element appears multiple times in the sample, each sampled unit reports. Cox (1993) describes this procedure as an adjustment of the weight "associated with sampled frame units to reflect the multiple selection opportunities for the desired population unit." This adjustment, obtained by multiplying the sampling weight N/n times $1/M_k$, results in an overall weight of $N/(n*M_k)$. This new weight is not, in general, equal to the reciprocal of the probability of selection. Nonetheless, this procedure does result in unbiased estimation (see Appendix C).

A third strategy is to compute correct inclusion probabilities for each sampled population element, based on the number of duplicate records on the frame. Under the assumption that all duplicate records on the frame may be identified for each unique sampled record, this calculation is straightforward for a simple random sample. Use of the Horwitz-Thompson "pi estimator" provides unbiased estimation (Musser, 1993).

In stratified survey designs, the latter two techniques are difficult to apply because duplicate records may appear in different strata. In that case, applying these techniques across strata would not maintain unbiased estimation within strata. A counting rule which specifies which stratum is the "correct" one is needed.

The NASS Solution: A Dual Approach to Deal with Duplication in Stratified Designs

The NASS strategy for dealing with the problem of duplication across strata is to apply a combination of the first two techniques. A unique counting rule is used when duplication occurs across strata and the adjustment technique is applied for within-stratum duplication.

If members of a linkage group are not all in the same stratum, only those in the highest stratum will account for the data for that operator or operation. Records from lower strata are considered out of scope for data collection. When duplication is discovered after data collection, data fields will be set to zero for sampled records in lower strata. If there are M_{kh} members of linkage group k in the highest stratum h, each sampled record will have reported data "adjusted," i.e. multiplied by $1/M_{kh}$. This quantity is referred to as the "data adjustment factor"(DAF). In practice, if more than one of the members of linkage group k in stratum h are sampled, the data are usually collected once and copied for the other records. Thus, if there are two records in the highest stratum, and both are sampled, then data for each will be multiplied by 1/2, and in effect the data for that operation will receive the full sampling weight. If only

one of the two is sampled, the weight or expansion factor for that operation will be halved. In some States, when there are two duplicate records sampled in the highest stratum, a shortcut is used to avoid entering the same data twice. The reporting unit code is used to set the DAF to 1 for one sampled record and to 0 for the other. The Survey Processing System (SPS) is the series of computer programs which edit survey data files and produce estimates and standard errors for NASS surveys. It uses the reporting unit code and partner strata codes to compute the correct DAF.

METHODS AND RESULTS

Linkage group resolution and review activities were conducted after the June Agricultural Survey in two States (North Carolina and Ohio) by two teams of two persons. Each team included a person from the Research Division and a person from the List Frame Section. The primary focus of these activities was the list duplication check listing. The SSO duplication check actions were reviewed and final duplication check linkage group resolution decisions were made. Various resources were utilized to resolve linkage groups. These resources included State Office staff, notes from the State Office copy of the duplication check printout, June Survey data listings, the real-time mail and maintenance system and telephone calls.

In Ohio, a total of 116 duplication check linkage groups (250 total records) were reviewed and resolved. In North Carolina a total of 100 duplication check linkage groups were reviewed and resolved. Information was recorded to provide the correct data adjustment factors to recompute June Survey indications of major agricultural items when duplication was detected among records included in that survey. Counts of the number of resolved records linked to sampled records listed in the duplication check output were also recorded. These counts were aggregated to the stratum level for the four major surveys for use in computing estimates of list frame duplication.

Estimates of the percent of duplication involved in sampling for the four major survey series in Ohio and North Carolina were computed using sample unit measures of duplication. A sampling approach is desirable, since a complete resolution of all possible pairwise comparisons is not feasible. Such an approach would be too time consuming and expensive. Using list duplication check listings which included only those linkage groups which contained at least one sampled record, a resolution process was performed. For each sampled record, a decision was made as to the number of other records (link count) considered to be duplicates of the sampled record.

Overall estimates of the percent of duplication in each State's list frame relative to major survey series are shown in Table 1. Overall estimates of the percent of duplication in the North Carolina frame for the four survey series ranged from 0.03 to 0.27 percent. Overall estimates of the percent of duplication in the Ohio frame relative to the four survey series ranged from 0.14 to 0.75 percent.

Table 1. Estimates of the Percent of List Frame Duplication - by State and Survey.				
	QAS	Cattle	FCRS	Labor
North Carolina				
Estimate	0.21%	0.14%	0.03%	0.27%
Standard Error	0.06	0.08	0.03	0.11
Ohio				
Estimate	0.19%	0.14%	0.75%	0.64%
Standard Error	0.04	0.08	0.55	0.41

One concern about these estimates of frame duplication is that they are perhaps under-estimates of the true percent of frame duplication due to a limitation of the duplication check process. In order for a linkage group to be formed, a match must occur on at least one pair of Social Security numbers, Employer Identification numbers, telephone numbers or linkage cross reference numbers. This procedure fails to detect duplication of records without any of these identifiers present. To evaluate the potential impact of this limitation, the frequencies of occurrence of these identifiers for records classified for the Quarterly Agricultural Surveys in each State was determined.

Table 2 lists the relative frequencies of various combinations of these identifiers for each State. At least one of the numbers is present on 97 percent of the records in North Carolina and 99 percent of the records in Ohio. In North Carolina 97 percent of the records have a Social Security number or a telephone number present, while in Ohio, almost 99 percent of the records have a Social Security number or a telephone number. As indicated by the percentages in Table 3, most of the duplication check matches were due to matches on Social Security numbers and/or telephone numbers.

Table 2. The Relative Frequencies of Social Security Numbers (SSN), Employer Identification Numbers (EIN) and Telephone Numbers (Phone) Contained on List Frame Name and Address Records - North Carolina and Ohio.

State	Numbers Present	Population %
North Carolina	None	2.8
	SSN only	0.8
	EIN only	0.1
	(SSN and EIN) only	<0.1
	Phone only	19.6
	(SSN and Phone) only	70.0
	(EIN and Phone) only	0.7
	ALL	6.0
Ohio	None	0.9
	SSN only	12.1
	EIN only	0.3
	(SSN and EIN) only	0.2
	Phone only	5.4
	(SSN and Phone) only	73.0
	(EIN and Phone) only	0.9
	ALL	7.2

Table 3. Duplication Check - Relative Frequencies of Link Groups Formed Due to Matches on Social Security Numbers (SSN), Employer Identification Numbers (EIN) and Telephone Numbers (Phone).

State	Matches ON	Percent
North Carolina	SSN only	9.0
	EIN only	1.0
	PHONE only	67.0
	(SSN and EIN) only	1.0
	(SSN and PHONE) only	20.0
	(EIN and PHONE) only	1.0
	ALL	1.0
Ohio	SSN only	5.2
	EIN only	2.6
	PHONE only	79.3
	(SSN and PHONE) only	12.0
	(EIN and PHONE) only	0.9
	ALL	<0.1

An adjustment to the estimates of the percent of list frame duplication from the duplication check procedure, was obtained by dividing the initial estimate by the percent of duplication check match possibilities p(**DCMP**). The p(**DCMP**) was calculated by determining the percent of all pairwise groupings of frame units with conforming duplication check variable(s) (Employer Identification numbers, or Social Security numbers or telephone numbers) present on both records. Table 4 illustrates all possible identifier presence/absence scenarios in matching two records.

Table 4. Pairwise Comparisons Possibilities.

	None	E	S	ES	P	EP	SP	ESP
None	X	X	X	X	X	X	X	X
E	X	**DCMP**	X	**DCMP**	X	**DCMP**	X	**DCMP**
S	X	X	**DCMP**	**DCMP**	X	X	**DCMP**	**DCMP**
ES	X	**DCMP**	**DCMP**	**DCMP**	X	**DCMP**	**DCMP**	**DCMP**
P	X	X	X	X	**DCMP**	**DCMP**	**DCMP**	**DCMP**
EP	X	**DCMP**	X	**DCMP**	**DCMP**	**DCMP**	**DCMP**	**DCMP**
SP	X	X	**DCMP**	**DCMP**	**DCMP**	**DCMP**	**DCMP**	**DCMP**
ESP	X	**DCMP**	**DCMP**	**DCMP**	**DCMP**	**DCMP**	**DCMP**	**DCMP**

The letters **E**, **S**, and **P** and the various combinations represent Employer Identification number (**E**), Social Security number (**S**) and telephone number (**P**). The percent of duplication check match possibilities for the Quarterly Agricultural Surveys are listed in Table 5. Adjusted estimates of the percent of list frame duplication for the Quarterly Agricultural Surveys (based on these percentages) are also shown in Table 5.

Table 5. Percent of Conforming Pairs (**Duplication Check Match Possibility**) Among Records Classified for Possible Inclusion in the Quarterly Agricultural Survey, Revised Estimates of the Percent of Duplicate Records in the Corresponding List Sampling Frame, and the Standard Errors of the Estimates (SE).

State	% (**DCMP**)	% Dup	SE
North Carolina	93.95	0.23	0.06
Ohio	96.09	0.20	0.05

To assess the effects of duplication on NASS survey indications, expansions were computed for three important variables from the 1992 June Agricultural Survey for North Carolina and Ohio under two different scenarios:

1. **"Worst Case Scenario."** Here we assumed that there was no duplication check of any kind performed, and that data would be collected for all sampled records of all linkage groups containing duplicates. The intent here was to simulate what indications would have been if no duplication check had been performed.

2. **"Best Case Scenario."** Here we assumed that all duplication was resolved "correctly" (i.e., as determined by our investigation), to simulate indications under a "perfect" duplication check.

These two sets of expansions were compared to the actual survey expansions for each State. The intent in the first case was to measure the impact of the current duplication handling process as it was performed in these two States. In the second case, the goal was to measure the effect of errors in carrying out the process. A detailed discussion of methods and results for each scenario follows.

To assess the impact of the duplication check process, we attempted to undo any actions taken as a result of the process, and assume that even at survey time, duplication would not be detected. Indications from these data were then compared to actual June Agricultural Survey indications to estimate the effect of the process. This "worst case scenario" assumed that data would be collected for every sampled record in each linkage group. If data were actually collected for only one of several records in a linkage group, those data were moved to the other records, and the DAF was set to "1" for each. If data were not collected at all for a sampled record (due to duplication with a record in a higher stratum that was not sampled) then we imputed stratum means for that record.

In Ohio, there were only 7 sampled duplicate records in the June Agricultural Survey, one of which had 3 tracts giving a total of 9 data file records. One of these was correctly coded as a duplicate, but because this record was duplicated with a record in a lower stratum, data were collected and the DAF was "1." The remaining 8 records were not coded as duplicates, data were collected, and the DAF's were set to "1" for all. Thus in Ohio, in the June Agricultural Survey, data were collected for all of the duplicate records and no data adjustment was made for duplication. Therefore the duplication check process had no effect on survey indications.

In North Carolina, there were 17 sampled duplicate records in the June Agricultural Survey, two of which had 2 tracts, giving a total of 19 data file records. Of these, 16 records were coded correctly, and 8 of these codings caused a change in the DAF. The procedure above was used to approximate the percentage change in survey indications in the absence of any duplication checking process. The results are summarized in Table 6.

Table 6. Impact of the Duplication Check Process on June Agricultural Survey Indications in North Carolina and Ohio. Relative Differences Between the "Worst Case" No Duplication Check Expansions and Actual Survey Expansions by State for Corn and Soybean Planted Acres and Total Hogs Indications.

State	Corn Planted	Soybean Planted	Total Hogs
North Carolina	0.42%	0.70%	0.31%
Ohio	0.00%	0.00%	0.00%

While these percentages suggest that the duplication check process in North Carolina did not have a substantial impact on survey indications, the process was successful in achieving correct data adjustment for 17 of the 19 records. The low relative differences are due to low rates of duplication.

To measure the impact of errors in the duplication check process on June Agricultural Survey indications, the June data were edited to achieve the "correct" DAF's and indications were compared to actual June Agricultural Survey indications.

Table 7. Effect of "Best Case" Duplication Check in North Carolina and Ohio. Relative Differences between Survey Expansions and "Correct" Expansions by State for Corn and Soybean Planted Acres and Total Hogs Inventory.

State	Corn Planted	Soybean Planted	Total Hogs
North Carolina	0.08%	0.20%	0.00%
Ohio	0.10%	0.12%	0.25%

In both States the impact of duplication check errors was very small with relative differences less than 0.25 percent, again due to low duplication rates. From a practical standpoint however, failing to correctly adjust for duplication detected by the current duplication check program will always produce a positive bias in estimation, which must be added to bias present from "undetected" duplication. With an improved duplication detection process, including more thorough record linkage, the importance of correct handling would be increased.

To estimate the total (currently "detectable") duplication bias that would have occurred without a duplication check process, we simply subtract the "correct" expansions from the "no duplication check" expansion. The following table gives the relative differences for both States:

Table 8. Estimates of Total "Detectable" Bias in North Carolina and Ohio. Relative Differences between "No Duplication Check" Survey Expansions and "Correct" Expansions by State for Corn and Soybean Planted Acres and Total Hogs Inventory.

State	Corn Planted	Soybean Planted	Total Hogs
North Carolina	0.50%	0.90%	0.31%
Ohio	0.10%	0.12%	0.25%

Comparing this estimate of the total detectable bias to the bias removed by the duplication check (Table 7) shows that the duplication check process in North Carolina was successful in removing a substantial percentage (corn acres: 85 percent; soybean acres: 77 percent; hogs: 100 percent) of the total bias that otherwise would have occurred.

Correct implementation of the NASS system of adjustment for duplication depends on several key phases of the system. The resolution of the potential duplicate linkage groups identified by the duplication check program must be done correctly. Then comments and a record status of "78" to indicate special handling must be attached to the list frame which clearly define the correct questionnaire codings for the survey cycle. These comments must be attached to the questionnaire, and retained through the survey cycle. After data collection these codes must be entered correctly on the questionnaire and data file. Questionnaire codings and DAF's were checked for each quarter, and results are given in the tables below. For each survey quarter, row 1 gives the total number of sampled records that were members of duplicate linkage groups. Row 2 gives the number of these records with "correct" coding of the 921 and 925-928 items (as defined by the NASS Supervising and Editing Manual). Since States sometimes use short cuts to manipulate these codes and still get the correct DAF, this row might be low without necessarily implying error in the State's duplication handling. Row 3 gives the number of records with the correct DAF. Row 4 gives the overall percentage of correct DAF's.

Table 9. Counts and Percentages for Quarterly Agricultural Survey Questionnaire Codings in **North Carolina** and **Ohio** Combined- by Quarter.

	June	Sept	Dec	March
Number of duplicated sample records	28	20	25	24
Number of records with correct coding of 921 and 925-928 boxes (NASS S&E manual)	17	16	16	14
Number of records with correct Data Adjustment Factor (DAF)	23	18	18	17
Percent of records with correct DAF	82	90	72	71

The table show that there is room for improvement in the current NASS duplication check process in order to correctly adjust survey data for duplication which is found by the duplication check program.

CONCLUSIONS AND RECOMMENDATIONS

The North Carolina and Ohio State Statistical Offices appear to be doing a very good job in controlling list frame duplication. The data from North Carolina and Ohio clearly suggest that both States have low rates of duplication on their list frames and that duplication is not causing serious inflation of survey indications in either State. Since our estimates of duplication are really only estimates of that portion of the total list frame duplication which can be detected by the current (limited) duplication check program, we believe that they are underestimates of the true frame duplication. To more accurately measure the amount of duplication that is "undetected" by the current program, future research should perform an independent, more thorough resolution of sampled records versus classified records with RECLSS, using name and address matching. Nonetheless, counts of list frame records in these States with the necessary data for matching are high and "adjusted" estimates of duplication based on these counts did not produce substantial changes in the estimates of percent duplication. Allowing for limitations in this adjustment procedure, duplication percentages appear to be less than one half of one percent in both States.

Due to these low rates, the effects of errors in the duplication check process on indications for the June Agricultural Survey were relatively small in both States. However, the table of coding performance over the full Agricultural Survey cycle shows that there is room for improvement in the process and we believe that a built in quality check could bring this about.

All sampled records listed in the duplication check print are flagged in the SPS edit. The problem with this is, usually only about 20 to 30 percent of the records identified in the print are actually in true duplicate linkage groups. This will tend to discourage States from carefully checking the coding of these records, in every survey and every quarter of the Agricultural Survey. We suggest an improvement whereby, prior to the June Survey, each SSO would have to record its determination of duplication for each record in the duplication check print. Records for which duplication could only be determined at survey time would be coded as "potential" duplicates. Only definite and potential duplicates would be flagged in the SPS edit. This flagging would be more helpful and less of a burden for the SSO's. It might even be possible for the edit to flag only those duplicates that appear to be coded incorrectly (by comparing the DAF to the correct duplication adjusted DAF).

An alternative plan under serious consideration at NASS is to drop the flagging of duplicate records (by SPS edit) altogether, while implementing improvements to the duplication check process. One suggested improvement is to implement a thorough duplication removal process prior to classification. Another suggestion is to replace the SPS flagging which occurs during the survey with a pre-survey quality check, which could help performance especially in the "follow on" surveys.

Recording decisions for each record in the duplication check process would have other benefits. In addition to providing motivation for the States to make clear determinations and record them in a timely manner, it would provide an annual measure of list duplication for each State. This would be very useful to monitor over time, and could be included in the "Red Book", an annual NASS publication on list coverage and quality. Comparison of these estimates to duplication percentage estimates obtained from a more thorough periodic full record linkage on the entire list frame would provide an evaluation of how well the duplication check process is working in identifying and controlling duplication.

A very important aspect of an efficient duplication control process is a good record linkage program. The NASS duplication check program (which matches only on telephone number, SSN, EIN, and linkage cross reference number) is used to provide a manageable set of potential duplicates at a relatively low cost, which can be manually reviewed under pre-survey time constraints. This may be a reasonable strategy if combined with a periodic (every 3 or 4 years) full resolution with RECLSS or another good record linkage program. However, improvements in "off the shelf" record linkage software and declining cost of very powerful PC's may make it possible to perform the annual duplication check with a much more powerful record linkage software on Agency PC's. The Survey Technology Branch is currently evaluating a PC-LAN-based record linkage package which shows promise. If such software could be "fine tuned" to provide accurate identification of duplicate records while minimizing "false matches", it might actually produce smaller "potential duplicate" sets than the current program. This could result in substantial improvement in the NASS duplication control process, and lowering of rented mainframe computer costs.

One minor issue which came up in this research is how to handle duplication adjustment when there is nonresponse. Currently, the DAF is imputed, and no adjustment for duplication is made. We believe that to maintain unbiased estimation under the current non-response assumptions, the DAF should be adjusted for duplication. If the record has multiple chances of selection, it must be reflected in a smaller weight, even with imputed data. This would require special coding because the same questionnaire item is used to record duplication and nonresponse.

Another observation concerns the coding of multiple operations under the recently adopted NASS "operator dominant" ("op dom") list frame design. Under this design an operator with several separate operations is normally given one active ("parent") record which is coded with an "active" record status of 85. All of the other operations are linked to this active record but are "inactive" records with record status of 45 (not eligible for sampling). If the active record is sampled (i.e. the operator is sampled) then all of his operations are sampled and questionnaires are generated for each of these. Under the old "operation dominant" design, each operation was sampled separately. During our duplication resolution in North Carolina, we found that roughly half of the duplication found was due to multiple operations of some operator being coded as active records. Some of these were just records which had not been handled correctly under the "op dom" guidelines, and others had been given a special (State specific) coding to avoid respondent burden inherent in the "op dom" system. We have two observations on this situation.

1. There may be a need for clearer documentation and explanation at survey training schools on the correct handling of multiple operations. Because control data are summed across the multiple operations, these operators not only have the burden of multiple questionnaires, but are also sampled more often. In the extreme cases, these should be treated as "special handling" records with an active record status of 99 for each operation.

2. When multiple operations are detected during the duplication check process, the list frame statistician normally takes action to change the list frame record status on these records to the 85-45 scheme for the following year. The real intent of this is to correct the duplication for the next year's sample. But, if the 85 record was sampled for the current year, each of the identified multiple operations will receive a questionnaire as a subtract of the 85 record. This is correct, but it is important that each of these questionnaires receives the same coding as the parent questionnaire. Therefore, any comments concerning duplication and correct questionnaire codings which are posted for the 85 record should also be posted to all of the identified multiple operations. There was a case where the parent record had a DAF of 0 but the data for the additional operation record received a DAF of 1.0. If incorrect coding of duplicates was flagged by the SPS edit, as mentioned above, errors like this would be caught.

In summary, we offer two recommendations for improvement in the duplication control process at NASS. The first is a thorough investigation of alternative PC based record linkage programs for more accurate, efficient identification of potential duplicates. The second would require States to capture their decisions on duplication for each potential duplicate record, which would provide yearly estimates of duplication for each State, and make it possible to build a duplication adjustment check into the SPS edit to catch errors over the whole survey cycle.

REFERENCES

Cox, B. (1993), Weighting Class Adjustments for Nonresponse in Integrated Surveys: Framework for Hog Estimation, United States Department of Agriculture, National Agricultural Statistics Service, Research Division, SRB Research Report Number SRB-93-03.

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

Musser, Orrin (1993), Unbiased Estimation in the Presence of Frame Duplication, Statistical United States Department of Agriculture, National Statistics Service, Research Division, SRB Research Report Number SRB-93-10, pp. 60-63.

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

APPENDIX A. DUPLICATION PROPORTION ESTIMATOR

LET $\hat{p}_{st}$ *denote the estimated proportion of list frame duplication.*

$$\hat{p}_{st} = \sum_{h=1}^{L} \frac{N_h \hat{p}_h}{N} \qquad \textit{and} \qquad \hat{p}_h = \sum_{i=1}^{n_h} \frac{p_{hi}}{n_h}$$

p_{hi} *is the measure of duplication from sample unit i in stratum h.*

$$p_{hi} = \left(1 - \frac{1}{m_{hi}}\right) \qquad \textit{and} \qquad m_{hi} = 1 + l_{hi}$$

l_{hi} *is a count of records linked to sample unit i in stratum h.*

$$V(\hat{p}_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} N_h (N_h - n_h) \frac{s_h^2}{n_h}$$

$$s_h^2 = \frac{\sum_{i=1}^{n_h} p_{hi}^2 - \frac{\left(\sum_{i=1}^{n_h} p_{hi}\right)^2}{n_h}}{n_h - 1}$$

APPENDIX B. DERIVATION OF OUR ESTIMATOR OF LIST DUPLICATION

Estimation of the percentage of duplicate records on a list frame is equivalent to estimation of the number of duplicate records on the frame or the number of unique records on the frame.

Let N be the number of records on the list frame, and let M be the number of unique records on the frame. Then N - M is the number of duplicate records and (N-M)/N, the percentage of duplicate records on the frame. All of these may be estimated by obtaining an estimate for the number of duplicate records on the frame.

If we take a simple random sample of size n_h from a list frame stratum h of size N_h, we estimate a total Y for the population by

$$\hat{Y}=\sum_{h=1}^{L}\frac{N_h}{n_h}\sum_{i=1}^{n_h}y_{hi}.$$

If population element i in stratum h appears on the list frame a total of m_{hi} times, ignoring the strata of duplicate records, then the number of duplicate records associated with this element is $(m_{hi}-1)$. To estimate the total duplicates on the frame, define $y_{hi} = (m_{hi}-1)$, so that

$$\hat{Y}=\sum_{h=1}^{L}\frac{N_h}{n_h}\sum_{i=1}^{n_h}(m_{hi}-1).$$

To maintain unbiased estimation in the presence of duplication on the frame, a common survey practice and one used in NASS surveys, is to use a weight or data adjustment factor to account for the effect of duplication. A population element which is represented multiple times on the frame has an increased selection probability and this must be adjusted for. If a population element hi appears on the frame m_{hi} times, then when sampled the data is multiplied by $1/m_{hi}$. This gives the following estimator of total duplicates:

$$\begin{aligned}\hat{Y}&=\sum_{h=1}^{L}\frac{N_h}{n_h}\sum_{i=1}^{n_h}\frac{1}{m_{hi}}(m_{hi}-1)\\&=\sum_{h=1}^{L}\frac{N_h}{n_h}\sum_{i=1}^{n_h}(1-\frac{1}{m_{hi}}).\end{aligned}$$

To estimate the **percentage** of duplication, this is simply divided by N. For stratum level estimates, divide stratum total estimates by N_h:

$$\hat{Y}_h=\frac{1}{n_h}\sum_{i=1}^{n_h}(1-\frac{1}{m_{hi}})$$

APPENDIX C. UNBIASED ESTIMATION USING DATA ADJUSTMENT

The current NASS strategy provides unbiased estimation[1] in the presence of list frame duplication, under the assumption that all duplication involving sampled records is identified correctly and that proper codings are entered on the data files.

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

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

Thus a reasonable estimate for X would be

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

Note again that this sum is over the entire sample of frame units. This is clearly unbiased for X, since this approach is equivalent to a simple random sample with the frame being the population.

In a sense, this data adjustment simply redefines the connection of the frame to the population, and in the process, adjusts for duplication. If a population element is represented k times on the frame, we just let each of the k frame units account for 1/k of the data to adjust for duplication.

---

[1] "unbiased estimation" in the sense that duplication does not contribute to the overall bias.