1992 CENSUS OF AGRICULTURE COVERAGE EVALUATION ESTIMATION

Paul J. Lewis, Trilogy Consulting Corp., Glenn S. Wolfgang, E. Ann Vacca, Bureau of the Census
E. Ann Vacca, Bureau of the Census, Washington, D.C., 20233[1]

KEY WORDS: Census completeness, classification error, mail list error, area frame sample

1. Introduction

A coverage evaluation program for a census is an important means of assessing the completeness and accuracy of the data collected during the census. An evaluation of coverage has been conducted for each census of agriculture since 1945. This paper presents the objectives, sample design, and estimators for the 1992 Census of Agriculture Coverage Evaluation Program.

Although the goal of each census of agriculture is to enumerate all farms in the nation, continuing change in operational units, inadequacies of source lists, difficulty in communicating census definitions and concepts, and other factors contribute to errors and incompleteness in the published census farm count. There are two general types of error encountered: list error and classification error. List error includes a measurement of farms not on the census mail list and a measurement of farms that are duplicated on the mail list. Classification error includes a measurement of farms incorrectly classified as nonfarms and nonfarms incorrectly classified as farms. Farms not on the census mail list and farms incorrectly classified as nonfarms contribute to an undercount of the true number of farms, while duplicated farms and nonfarms incorrectly classified as farms contribute to an overcount of the true number of farms. For all sizes of farms, the list error of farms not on the mail list dominates other errors although the components vary considerably by state.

A total of 2,243,648 farms were enumerated during the 1987 Census of Agriculture. There was a net undercount of farms of approximately 7.2% (±0.3%). This includes an estimated 242,853 (±7,613) farms not on the mail list, 54,080 (±5,354) farms incorrectly classified as nonfarms, 72,310 (±6,920) nonfarms incorrectly classified as farms, and 63,290 (±6,579) duplicated farms. (Note: The numbers in parentheses represent 1.645 standard error above and below the estimates). Estimates of the components of error in the 1992 Census of Agriculture will be provided in a separate report in the census of agriculture publication series. In addition, a net coverage error will be provided.

2. 1992 Census of Agriculture Coverage Evaluation Program

Data from the 1992 Census of Agriculture Coverage Evaluation Program will provide an independent measure of the number of farms not on the census mail list, the number of incorrectly classified farms, the number of duplicated farms, and the characteristics of these farms. The evaluation provides information about problem areas to allow for future improvements in developing the census mail list and in collecting and processing the data.

The 1992 Coverage Evaluation is conducted using the 1992 June Agriculture Survey (JAS), conducted by the National Agricultural Statistics Service (NASS) of the United States Department of Agriculture, and the Census Bureau's 1992 Classification Error Survey (CES).

The JAS is used in estimating the number and characteristics of farms not on the mail list (NML) for the 1992 census. State level estimates of farms NML their characteristics are being published in Table G of Volume 1, Geographic Area Series, Appendix C. The CES will be used to estimate the number and characteristics of farms incorrectly classified and duplicated. Results from both the NML study and the CES will be published in Volume 2, Subject Series, Coverage Evaluation. This publication will be available in early 1995.

2.1 Census of Agriculture Data Collection

The census of agriculture is a major source of data for the nation's agriculture production. It is the only source of uniform, comprehensive data on agricultural production and operator characteristics for the nation and for each county and state in the United States. Censuses are conducted on a five year cycle for years ending in 2 and 7. Report forms for the 1992 Census of Agriculture were mailed to farm and ranch operators in late December 1992 to collect data for the 1992 calendar year. All those who received a census form were asked to return their report forms by February 1, 1993. Those not responding by that date were contacted by mail followups or telephone calls. Upon receipt by the Bureau of the Census, the forms were checked for completeness and accuracy. The data are now being tabulated and published for the 1992 Census.

2.2 Not on the Mail List Study

The 1992 Census of Agriculture Coverage Evaluation Program is being used to provide state-level estimates of farms and characteristics of farms not on the census mail list (estimates will not be made for Alaska or Hawaii). Not on the mail list (NML) estimates are made using an independent estimate of total farm count in a dual-system estimation model (Wolter 1986). Rather than constructing an area frame, selecting a sample, and conducting a field enumeration survey, the Census Bureau's Agriculture Division (AGR) has a cooperative agreement with NASS

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau or Trilogy Corp.

to use data collected in its JAS. In this agreement, the AGR provided requirements for the 1992 JAS data collection to ensure that the resulting data were appropriate for the census coverage evaluation program, including the collection of additional data items by NASS.

The JAS is an annual area sample conducted by NASS to measure planted acreage of crops and numbers of livestock. It provides the basis for several subsequent NASS surveys including the September, December and March Agricultural Surveys. Enumeration for the JAS is done by personal interview during the first two weeks of June. The reference date used for reporting is June 1.

NASS's area frame is created by dividing the land in a state into six to eight land-use strata such as intensively cultivated land, urban areas, agricultural urban areas, and rangeland. The land-use strata are identified on county highway maps using permanent and easily recognizable land features. Cluster analysis is used to delineate into substrata with similar agricultural makeups. Substrata are then divided into segments using aerial photographs. A typical segment contains portions of 2 to 4 farm operations. Since the land area within each segment is completely enumerated, the segment and not the farm is the basic unit of analysis for the JAS. Refer to Cotter and Nealon (1987) for more details on the JAS design.

Once the segments are chosen, an enumerator visits them and establishes who operates the land within the segment, defining the ultimate reporting unit, the tract. Only one farm operation is associated with a tract; however, a farm may be represented by tracts in more than one segment.

The 1987 Census of Agriculture estimate of farm count from the JAS was based on an open segment estimator -- data was collected only from those farms whose operator's residence was located within the sampled segment. The 1992 Census of Agriculture estimate of farm count from the JAS is based on a weighted segment estimator, which uses a proportion of data from each farm operation in the segment regardless of where the operator's residence is located.

2.3 Census Processing of the JAS

The Census Bureau received two files of JAS data from NASS. One file contained the names, addresses, and other identifier information for all sample area segment tracts that had any indication of agricultural activity. An initial computer match of these records to the census mail list identified all JAS area sample records as either matched or nonmatched to the census mail list. All JAS sample records not matched to the mail list were assigned a census file number (CFN) and were added to the mail list to be included in the mailout. The other JAS data file contained supplemental data for each JAS tract, including identifying characteristics and whole-farm commodity data.

All census report forms mailed in the census were returned to the Census Bureau's Jeffersonville, Indiana processing center. JAS-census cases were sorted out from the main census cases and microfilmed. The forms were returned to the main census processing after they were microfilmed. All cases went through normal census processing including data entry and the edit and imputation system. The data from nonmatched JAS cases were removed from census processing before tabulation and analytical review of aggregate census estimates.

JAS-census cases that did not respond to the census received the same type of mail follow-up as regular census cases. However, after prespecified cut-off dates, all JAS-census nonrespondents were telephoned in an attempt to obtain information. Those cases which still had not responded after telephone follow-up had data imputed for them based on the JAS data.

The 1992 JAS reference period was from January 1, 1992 to June 1, 1992, while the census reference period covered January 1, 1992 to December 31, 1992. Consequently, some JAS records which were considered to be nonfarms in June would have been farms if the JAS had been conducted at the same time as the census. If census but not JAS processing determined that a place was a farm, the NASS farm definition was applied to the census data for that place to classify the record. Farms that became nonfarms between June and December were not a concern since an operation that qualifies as a farm anytime during the year is counted as a farm by the census.

During processing, the evaluation unit compared the census data to the JAS data for each matched case to ensure that it was a valid match. The census database was searched to find potential matches to the nonmatched cases using the names, addresses and data from the JAS as well as the census collected data. Discrepancies between the JAS farm status and the census farm status were resolved to ensure that there were no incorrectly classified farms or nonfarms among the nonmatched cases.

2.4 Classification Error Survey

The Classification Error Survey (CES) is used to measure classification error and list duplication error. It is designed to measure those farms incorrectly classified as nonfarms, nonfarms incorrectly classified as farms, and farms duplicated on the mail list. This independent sample is selected from the census mail list and is designed to provide census region level estimates of the number of incorrectly classified farms and duplicated farms with a designed coefficient of variation of 10 percent.

The classification sample was selected from the final census mail list independently within census regions using a systematic sample design. The sampling rates within census regions were based on the estimated proportion of farms incorrectly classified and duplicated in 1987 and a specified coefficient of variation. Records ineligible for selection include operations in Alaska and Hawaii, operations with expected sales of $500,000 or more, and multi-unit or abnormal operations. Abnormal operations include Indian reservations, research farms, experiment farms, institutional farms, etc.

2.5 Data Processing of the CES

A CES questionnaire was mailed independently of the regular census mailing to those addresses which had responded to the census. Two mailings were made; one on April 1, 1993, and the second on July 1, 1993. CES nonrespondents were sent a postcard follow-up two weeks after initial mailout and a form follow-up four weeks after initial mailout. Any remaining nonrespondents were telephoned six weeks after initial mailout.

A clerical review of the CES forms was conducted to classify each record as either a farm or nonfarm. The CES farm status will be compared to census farm statues to identify cases which have been incorrectly classified or duplicated in the census. All error cases will be referred to an analyst for further review. The analyst will attempt to reconcile differences using telephone follow-ups. Estimates of the number of farms incorrectly classified as nonfarms and characteristics of those farms will be derived using the data from the CES questionnaires. The estimate of the number of farms overcounted due to classification error and characteristics of those farms will be derived using census data since these data are what were overcounted. Duplicated farms will be evaluated case by case since past CES's have shown that not all data from duplicated farms are actually reported more than once.

3. Coverage Estimators

Estimates are computed for farms not on the census mail list, farms incorrectly classified as nonfarms, nonfarms incorrectly classified as farms, and duplicated farms.

The total farm population (T) is defined as the census published number (C) of farms plus the undercounted (U) farms minus the overcounted (OV) farms:

$$T - C + U - OV \quad (3.1)$$

The undercount can be grouped into those farms not on the census mail list (NML) and those farms on the census mail list that were incorrectly classified as nonfarms (ICU). Likewise, the overcount can be divided into two groups: those nonfarms on the census mail list that were incorrectly classified as farms (ICO) and those farms duplicated in the census (DUP). The universe total can then be restated as:

$$T - C + NML + ICU - ICO - DUP \quad (3.2)$$

The estimate of the universe total for some farm characteristic x is defined similarly:

$$T_x - C_x + NML_x + ICU_x - ICO_x - DUP_x \quad (3.3)$$

The estimation of the various components of T and T_x are discussed in the following sections.

3.1 Estimation of Farms Not on the Mail List

The estimate of the number of farms not on the census mail list is derived using a coverage error model based on dual-system estimation theory. Its properties and derivation are discussed by Wolter (1986). The model can be extended to provide estimates of the characteristics of farms not on the census mail list as well. There are several assumptions implicit in the coverage error model:

1) Census and JAS are independent of one another; that is, the probability of a farm being on the census mail list is independent of the probability of a farm being enumerated by the JAS and vice versa. Since the NASS list frame is used to build the census list, the AGR depends on independence in the NASS list and area frame. (Nealon 1984)
2) The probability of being missed by either the census or the JAS is the same for all farms within a given size category.
3) It is possible to match the JAS sample results to the census results without error.
4) Spurious events have been eliminated, e.g. duplicates on the census mail list, nonexistent cases in either the JAS or the census, and out-of-scope census cases.
5) Enough information is collected about the nonresponse cases in both the census and the JAS to allow accurate classification.
6) The reference periods for both the census and the JAS are well defined.

By matching the JAS respondents to the census mail list, each record can be classified into one of the cells of the contingency table shown in Figure 1.

JAS Area List

		Captured	Not Captured	
Census List	Captured	N_{11}	N_{12}	N_{1+}
	Not Captured	N_{21}	N_{22}	NML
		N_{+1}		T

The notation used in the table is;

N_{11} = number of farms on the census mail list and in the JAS universe resulting from the match of the JAS sample farms to farms in the census,

N_{12} = number of farms on the census mail list but not in the JAS farm universe,

N_{21} = number of farms in the JAS but not on the census mail list resulting from the match of the JAS sample farms to farm records in the census,

N_{22} = number of farms not on the census mail list and not in the JAS farm universe,

N_{1+} = number of farms on the census mail list,

N_{+1} = number of farms in the JAS,

NML = total number of farms not on the mail list, and

T = total number of farms in the population.

Letting;

$\hat{N}_{11}$ = JAS expanded number of farms on the census mail list and in the JAS universe resulting from the match of the JAS area sample farms to farm records in the census (matched farms)

$\hat{N}_{21}$ = expanded number of farms in the JAS but not on the census mail list resulting from the match of the JAS area frame sample farms to farm records in the census (nonmatched farms)

$\hat{N}_{1\cdot}$ = number of farms on the census mail list adjusted for classification error and list duplication.

then we can estimate $\hat{N}_{12}$, $\hat{N}_{22}$, and $N\hat{M}L$ as follows:

$$\hat{N}_{12} = \hat{N}_{1\cdot} - \hat{N}_{11} \quad (3.4)$$

$$\hat{N}_{22} = \frac{\hat{N}_{21}\hat{N}_{12}}{\hat{N}_{11}} \quad (\because \textit{ of the independence assumption}) \quad (3.5)$$

The estimate of the total number of farms not on the mail list ($N\hat{M}L$) can then be shown to equal:

$$N\hat{M}L = \hat{N}_{21} + \hat{N}_{22} = \hat{N}_{21}\frac{\hat{N}_{1\cdot}}{\hat{N}_{11}} \quad (3.6)$$

Some characteristic x of the farms not on the census mail list ($N\hat{M}L_x$) is estimated by:

$$N\hat{M}L_x = \hat{Q}_x\frac{\hat{N}_{1\cdot}}{\hat{N}_{11}} \quad (3.7)$$

where $\hat{Q}$ is the unbiased weighted segment estimate of the total of the characteristic x for farms not on the mail list but in the JAS sample. The number of matched farms, $\hat{N}_{11}$ is estimated by:

$$\hat{N}_{11} = \sum_{i=1}^{L}\sum_{j=1}^{p_i}\sum_{k=1}^{n_{ij}} e_{ijk}y_{ijk} \quad (3.8)$$

where:

L = number of land-use strata

p_i = number of substrata in the ith land-use stratum

n_{ij} = number of sample segments in the ijth substratum

e_{ijk} = segment expansion factor - the inverse of the probability of selection for the ijkth segment

y_{ijk} = number of matched farms in the ijkth segment

The number of nonmatched farms, $\hat{N}_{21}$, is calculated the same as $\hat{N}_{11}$ except that y_{ijk} is the number of nonmatched farms in the kth segment, jth substratum, and ith land-use stratum. The estimator for $\hat{Q}_x$ is analogous:

$$\hat{Q}_x = \sum_{i=1}^{L}\sum_{j=1}^{p_i}\sum_{k=1}^{n_{ij}} e_{ijk}y_{ijk} \quad (3.9)$$

where y_{ijk} is now the value for the characteristic of the farms in the segment instead of the number of farms. The estimates of $N\hat{M}L$ **and** $N\hat{M}L$ are found by substituting the $\hat{N}_{11}$ **and** $\hat{N}_{21}$ **or** $\hat{Q}_x$ into equation 3.6 or 3.7. Note: $\hat{N}_{21}$ is a special case of $\hat{Q}_x$ and is equivalent to $\hat{Q}_x$ when the farm characteristic is farm count. $\hat{Q}_x$ is used in the notation to represent both variables.

Segment data values are generated by summing farm data, which is prorated by the amount of the farm's acreage in the sampled segment. The weighted value of the farms in segment y_{ijk} is:

$$y_{ijk} = \sum_{m=1}^{f_{ijk}} a_{ijkm}y_{ijkm} \quad (3.10)$$

where:

a_{ijkm} = farm weight (total tract acres / census entire farm acres) for the mth tract in the ijkth segment

y_{ijkm} = total value of the farm represented by the mth tract in the ijkth segment

3.2 Poststratification of the JAS Records

Coverage evaluations of previous censuses of agriculture have shown that small farms have a greater chance of not being on the mail list than large farms. Thus, poststratification is used in the NML estimation to account for the heterogenous capture probabilities. The JAS records are poststratified based on the total value of products sold (TVP) using $2,500 as the cut-off. The poststrata are collapsed if less than ten records (matched or nonmatched) are in either strata. The poststratified estimator is the sum of $N\hat{M}L_x$ applied within the hth poststrata:

$$N\hat{M}L_x = \sum_{h=1}^{2}\hat{Q}_{(h)}\frac{\hat{N}_{1\cdot(h)}}{\hat{N}_{11(h)}} \quad (3.11)$$

3.3 Variance of the Not on the Mail List Estimates

Since two versions of $N\hat{M}L_x$ are possible (poststrata or no poststrata), two variance estimators are needed. The Delta or Taylor Series methods outlined in Chiang (1980), Wolter (1985), and Bishop *et al.* (1975) were used to derive the asymptotic variance of the estimates for each case. The details of the derivations may be found in Lewis (1993a). The variance for the nonpostratified case is given by:

$$\mathrm{Var}(N\hat{M}L_x) = (N\hat{M}L_x)^2\left[\frac{\mathrm{Var}(\hat{Q}_x)}{\hat{Q}_x^2} + \frac{\mathrm{Var}(N_{1\cdot})}{N_{1\cdot}^2} + \frac{\mathrm{Var}(\hat{N}_{11})}{\hat{N}_{11}^2} - \frac{2\mathrm{Cov}(\hat{Q}_x,\hat{N}_{11})}{\hat{Q}_x\hat{N}_{11}}\right] \quad (3.12)$$

The variance for the poststratified case is given by:

$$Var(N\hat{M}L_x) = \sum_{k=1}^{2} (N\hat{M}L_{x_{(h)}})^2 \left[\frac{Var(\hat{Q}_{x_{(h)}})}{\hat{Q}^2_{x_{(h)}}} + \frac{Var(\hat{N}_{1_{(h)}})}{\hat{N}^2_{1_{(h)}}} + \frac{Var(\hat{N}_{11_{(h)}})}{\hat{N}^2_{11_{(h)}}} - \frac{Cov(\hat{Q}_{x_{(h)}},\hat{N}_{11_{(h)}})}{\hat{Q}_{x_{(h)}}\hat{N}_{11_{(h)}}}\right]$$
$$+ 2\, N\hat{M}L_{x_{(h)}} N\hat{M}L_{x_{(h)}} \left[\frac{Cov(\hat{Q}_{x_{(h)}},\hat{Q}_{x_{(h)}})}{\hat{Q}_{x_{(h)}}\hat{Q}_{x_{(h)}}} - \frac{Cov(\hat{Q}_{x_{(h)}},\hat{N}_{11_{(h)}})}{\hat{Q}_{x_{(h)}}\hat{N}_{11_{(h)}}} + \frac{Cov(\hat{N}_{1_{(h)}},\hat{N}_{1_{(h)}})}{\hat{N}_{1_{(h)}}\hat{N}_{1_{(h)}}}\right.$$
$$\left. - \frac{Cov(\hat{N}_{11_{(h)}},\hat{Q}_{x_{(h)}})}{\hat{N}_{11_{(h)}}\hat{Q}_{x_{(h)}}} + \frac{Cov(\hat{N}_{11_{(h)}},\hat{N}_{11_{(h)}})}{\hat{N}_{11_{(h)}}\hat{N}_{11_{(h)}}}\right] \qquad (3.13)$$

For both the poststratified and non-poststratified case, $V(\hat{Q}_x)$, **and** $V(\hat{N}_{11})$ are the stratified variance estimators (S^2_{N11} and S^2_{Qx})due to the JAS sample design (Kott 1988):

$$S^2_{N_{11}} = \sum_{i=1}^{L} \sum_{j=1}^{P_i} \frac{n_{ij}}{n_{ij} - 1} \sum_{k=1}^{n_{ij}} \left(e_{ijk} y_{hijk} - \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} e_{ijk} y_{hijk}\right)^2 \qquad (3.14)$$

where,

y_{hijk} = weighted sample estimate of farm value in the ijk^th segment and h^th poststrata, and

An analogous statement can be shown for S^2_{Qx} where "estimate of farms" is replaced by "value of the characteristic of the farms".

3.4 Percent of Farms Not on the Mail List

The percent of farms not on the mail list is estimated by:

$$\hat{P} = \frac{N\hat{M}L}{N\hat{M}L + \hat{N}_{1+}} (3.15)$$

The percent of a characteristic of farms not on the mail list is: $\hat{P}_x = \dfrac{N\hat{M}L_x}{N\hat{M}L_x + \hat{C}_x}$ (3.16)

where $\hat{C}_x$ is the census level estimate for the particular characteristic.

The variances of the percent of farms and characteristics of farms not on the mail list are derived using the same techniques as above (see Lewis 1993b for the derivations.) The variance of the non-poststratified case is:

$$Var(\hat{P}_x) = \left[\frac{N\hat{M}L_x C_x}{(N\hat{M}L_x + C_x)^2}\right]^2 \left[\frac{Var(\hat{Q}_x)}{\hat{Q}^2_x} + \frac{Var(N_1)}{N^2_1} + \frac{Var(\hat{N}_{11})}{\hat{N}^2_{11}} + \frac{Var(C_x)}{C^2_x}\right.$$
$$\left. - 2\frac{Cov(\hat{Q}_x,\hat{N}_{11})}{\hat{Q}_x\hat{N}_{11}} - 2\frac{Cov(N_1,C_x)}{N_1 C_x}\right] \qquad (3.17)$$

The variance of the poststratified case is:

$$Var(\hat{P}_x) = \sum_{k=1}^{2}\left(\left[\frac{C_x N\hat{M}L_{x_{(h)}}}{(N\hat{M}L_x + C_x)^2}\right]^2 \left[\frac{V(\hat{Q}_{x_{(h)}})}{\hat{Q}^2_{x_{(h)}}} + \frac{Var(N_{1_{(h)}})}{N^2_{1_{(h)}}} + \frac{Var(\hat{N}_{11_{(h)}})}{\hat{N}^2_{11_{(h)}}}\right] + \left[\frac{C_{x_{(h)}} N\hat{M}L_x}{(N\hat{M}L_x + C_x)^2}\right]^2 \left[\frac{V(C_{x_{(h)}})}{C^2_{x_{(h)}}}\right]\right.$$
$$\left. - \frac{2C_x^2 N\hat{M}L^2_{x_{(h)}}}{(N\hat{M}L_x + C_x)^4}\left[\frac{Cov(\hat{Q}_{x_{(h)}},\hat{N}_{11_{(h)}})}{\hat{Q}_{x_{(h)}}\hat{N}_{11_{(h)}}}\right] - \frac{2C_{x_{(h)}} C_x N\hat{M}L_{x_{(h)}} N\hat{M}L_x}{(N\hat{M}L_x + C_x)^4}\left[\frac{Cov(N_{1_{(h)}},C_{x_{(h)}})}{N_{1_{(h)}}C_{x_{(h)}}}\right]\right)$$
$$+ \frac{2C_x^2 N\hat{M}L_{x_{(h)}} N\hat{M}L_{x_{(h)}}}{(N\hat{M}L_x + C_x)^4}\left[\frac{Cov(\hat{Q}_{x_{(h)}},\hat{Q}_{x_{(h)}})}{\hat{Q}_{x_{(h)}}\hat{Q}_{x_{(h)}}} - \frac{Cov(\hat{Q}_{x_{(h)}},\hat{N}_{11_{(h)}})}{\hat{Q}_{x_{(h)}}\hat{N}_{11_{(h)}}} + \frac{Cov(N_{1_{(h)}},N_{1_{(h)}})}{N_{1_{(h)}}N_{1_{(h)}}}\right.$$
$$\left. - \frac{Cov(\hat{N}_{11_{(h)}},\hat{Q}_{x_{(h)}})}{\hat{N}_{11_{(h)}}\hat{Q}_{x_{(h)}}} + \frac{Cov(N_{11_{(h)}},\hat{N}_{11_{(h)}})}{\hat{N}_{11_{(h)}}\hat{N}_{11_{(h)}}}\right] - \frac{2C_{x_{(h)}} C_x N\hat{M}L_{x_{(h)}} N\hat{M}L_x}{(N\hat{M}L_x + C_x)^4}\left[\frac{Cov(N_{1_{(h)}},C_{x_{(h)}})}{N_{1_{(h)}}C_{x_{(h)}}}\right]$$
$$- \frac{2C_{x_{(h)}} C_x N\hat{M}L_{x_{(h)}} N\hat{M}L_x}{(N\hat{M}L_x + C_x)^4}\left[\frac{Cov(C_{x_{(h)}},N_{1_{(h)}})}{C_{x_{(h)}}N_{1_{(h)}}}\right] + \frac{2C_{x_{(h)}} C_{x_{(h)}} N\hat{M}L^2_x}{(N\hat{M}L_x + C_x)^4}\left[\frac{Cov(C_{x_{(h)}},C_{x_{(h)}})}{(C_{x_{(h)}}C_{x_{(h)}})}\right] \qquad (3.18)$$

3.5 Estimation of Incorrectly Classified and Duplicated Farms

The estimates of the number of incorrectly classified farms and the estimate of the number of duplicated farms will come from the CES. They will be calculated at the region level for the total number of farms, categories of farms (e.g. the number of incorrectly classified farms smaller than 50 acres in size), and characteristics of farms.

To calculate the total number of farms which are incorrectly classified as nonfarms (ICU) in the i^th region, let a_{ij} be an indicator variable equal to 1 if the j^th farm in the i^th region is incorrectly classified as a nonfarm, and 0 otherwise. Then,

$$ICU_i = WGT_i \sum_{j}^{n_i} a_{ij} \quad (3.19)$$

where WGT_i is the weight of an individual record in the i^th region:

$$WGT_i = \frac{N_i}{n_i} \quad (3.20)$$

with N_i equal to the total number of records on the mail list for the region, and n_i equal to the total number of farms in sample from the i^th region. Because the weight is constant within a region, an individual record's contribution to the total error is ($WGT_i a_{ij}$). The total number of farms incorrectly classified as nonfarms in the U.S. ,ICU, is then:

$$ICU = \sum_{i}^{4} WGT_i \sum_{j}^{n_i} a_{ij} \quad (3.21)$$

The amount of a characteristic x on farms incorrectly classified as nonfarms in the U.S., ICU_x , is estimated by:

$$ICU_x = \sum_{i}^{4} WGT_i \sum_{j}^{n_i} a_{ij} y_{ij} \quad (3.22)$$

where y_{ij} = the value of the x^th characteristic on the j^th incorrectly classified farm in the i^th region.

The number of farms incorrectly classified as nonfarms within a certain category, ICU_c is estimated by:

$$ICU_c = \sum_{i}^{4} WGT_i \sum_{j}^{n_i} a_{ij} c_{ij} \quad (3.23)$$

where c_{ij} is an indicator variable equal to 1 if the j^th farm in the i^th region is contained in the particular category, and 0 otherwise.

The amount of a characteristic x on farms incorrectly classified as nonfarms within a certain category, ICU_{cx} , is estimated by:

$$ICU_{c_x} = \sum_{i}^{4} WGT_i \sum_{j}^{n_i} a_{ij} c_{ij} y_{ij} \quad (3.24)$$

The number of nonfarms incorrectly classified as farms (ICO) and the number of duplicated farms (DUP) are calculated in a similiar manner. The total number of nonfarms incorrectly classified as farms and the total number of duplicated farms and their characteristics are estimated by making the appropriate substitutions into equations 3.23 and 3.24.

3.6 The CES Variance Estimators

Even though the CES sample was drawn using systematic sampling, the variance of the estimates of incorrectly classified and duplicated cases is computed as if a simple random sample was selected from each region. This assumption is thought to be valid since the records on the census mail list are ordered sequentially by census file number (CFN) and there is no evidence that CFN's contain a periodic or linear trend.

The variance for the number of farms incorrectly classified as nonfarms, Var(ICU), in region i is based on Cochran's (1980) variance for a simple random sample without replacement (ignoring the finite population correction factor):

$$\mathrm{Var}(ICU_i) = WGT_i^2 \frac{n_i}{n_i - 1} \sum_{j=1}^{n_i} (a_{ij} - \bar{a}_i)^2 \qquad (3.25)$$

The variance for the number of farms incorrectly classified as nonfarms in the U.S. is the sum across the regions:

$$\mathrm{Var}(ICU) = \sum_{i=1}^{4} WGT_i^2 \frac{n_i}{n_i - 1} \sum_{j=1}^{n_i} (a_{ij} - \bar{a}_i)^2 \qquad (3.26)$$

The variance for some characteristic of farms incorrectly classified as nonfarms in the U.S., Var(ICU_x), is:

$$\mathrm{Var}(ICU_x) = \sum_{i=1}^{4} WGT_i^2 \frac{n_i}{n_i - 1} \sum_{j=1}^{n_i} \left(a_{ij} y_{ij} - \bar{a}_{ij} \bar{y}_{ij}\right)^2 \qquad (3.27)$$

The variance of the number of farms incorrectly classified as nonfarms in a particular category in the U.S., Var(ICU_c), is estimated by:

$$\mathrm{Var}(ICU_c) = \sum_{i=1}^{4} WGT_i^2 \frac{n_i}{n_i - 1} \sum_{j=1}^{n_i} \left(a_{ij} c_{ij} - \bar{a}_{ij} \bar{c}_{ij}\right)^2 \qquad (3.28)$$

The variance for the characteristics of farms incorrectly classified as nonfarms in a particular category, V(ICU_{cx}), is estimated by:

$$\mathrm{Var}(ICU_{c_x}) = \sum_{i=1}^{4} WGT_i^2 \frac{n_i}{n_i - 1} \sum_{j=1}^{n_i} \left(a_{ij} c_{ij} y_{ij} - \bar{a}_{ij} \bar{c}_{ij} \bar{y}_{ij}\right)^2 \qquad (3.29)$$

The variances for the number of nonfarms incorrectly classified as farms and the number of duplicated farms are estimated in a similiar manner as shown above.

3.7 Total Undercount

The estimate of the total number of undercounted farms (U) is the sum of $\widehat{NML}$ from the JAS and ICU from the CES. The variance of U is the sum of the variances of $\widehat{NML}$ and ICU. The total of some characteristic of undercounted farms and its variance are similarly found.

3.8 Total Overcount

The estimate of the total number of overcounted farms (OV) is the sum of ICO and DUP from the CES. The variance of OV is the sum of the variances of ICO and DUP. The total amount of some characteristic of overcounted farms and its variance are similarly found.

3.9 Total Farms

The estimated total number of farms is calculated as shown in equation 3.2. The variance of the estimated total number of farms is computed by summing the variances of $\widehat{NML}$, ICO, ICU, and DUP. The total and variance of some characteristic of farms are similarly computed..

4. Acknowledgments

The success of the 1992 Coverage Evaluation Program is credited to the team effort of Steve Ash, Melody Atkinson, Beverly Battle, Demos Birbillis, Esther Darnell, William Davie, Tommy Gaulden, Bruce Hughes, Jerry Imel, Vince Matthews, Bessie Marz, Marsha Milburn, Dora Monroe, Jane Sandusky, Anthony Williams, Antoinette Wooten under the direction of Ruth Detlefsen, Ruth Ann Killion and Nash Monsour. The authors wish to thank Richard Sigman and Anthony Williams, Census Bureau reviewers, for their helpful comments.

5. References

Cochran, W.G. 1977. Sampling Techniques, 3rd Edition. John Wiley & Sons, New York. 428 pp.

Cotter, and J.P. Nealon. 1987. Area Frame Design for Agricultural Surveys. USDA./NASS, Washington, D.C. 67 pp.

Kott, P. S. 1988. Estimating Variance for the June Enumerative Survey. STB-88-06, USDA/NASS, Washington, D.C. 12 pp.

Lewis, P.J. 1993a. Estimating Farms not on the Census Mail List. Interoffice memo #92EAG-A.

Lewis, P.J. 1993b. Estimating Percent of Farms not on the Census Mail List. Interoffice memo #92EAG-A.

Nealon, J.P. 1984. Review of the Multiple and Area Frame Estimators. SF & SRB Staff Report No. 80. USDA/SRS, Washington, D.C.

Wolter, K.M. 1986. Some Coverage Error Models for Census Data. JASA. 81:338-346.

Wright, K.E., W.C. Davie, J.D. Sandusky, E.A. Vacca. 1989. Evaluation of Census Coverage. 1989 Proceedings of the section on Survey Research Methods, ASA, pp. 599-604

Yates, F. 1981. Sampling Methods for Censuses and Surveys, 4th Edition. Charles Griffin & Company, London. 458 pp.