Appendix C.
Statistical Methodology

THE SCREENING PHASE AND THE MAIL LIST MODEL

The 1997 Census of Agriculture featured a pre-census screening phase that surveyed selected records, by mail or telephone, for presence or absence of agricultural activity. Records selected for screening had a low probability of qualifying as farms. All records responding to the screener and reporting no agricultural activity were removed from the census mail list. Eliminating nonfarm records from the mail list reduced respondent burden and data collection costs.

The screening phase included nearly 500,000 records. Records were selected for screening using one of the following criteria:

1) Records on selected agriculture specialty lists that had no other list source,

2) Records identified by a mail list model as having a low probability of being a farm.

A mail list model predicted the probability that an addressee on the 1997 preliminary census mail list operated a farm. The model defined groups based on combinations of characteristics such as source(s) of the mail list record, expected value of agricultural production, and geographic location. Farm proportions were estimated for these groups by calculating the proportion of 1992 census respondent records that were farms which exhibited the characteristics defined by the group. This proportion, also called the in-scope rate, provided an estimate of the probability that an addressee in the group operated a farm.

Each address record on the 1997 preliminary census mail list was assigned to a model group by matching record characteristics to model group characteristics. Records belonging to the groups with the highest farm probability were those more likely to be farms. Records with a farm probability of approximately 30 percent or less were selected for screening, along with records included on selected agriculture specialty lists as noted above.

Before screening, the preliminary census mail list consisted of 3,314,790 records. There were 478,298 records selected for screening. Of these, 125,570 records were determined to be nonfarms as a result of the screening phase and were removed. These records were removed from the final census mail list. The remaining 3,189,220 records received census report forms.

CENSUS SAMPLE DESIGN

All name and address records on the final census mail list were designated to receive a 1997 Census of Agriculture report form. Two different types of census report forms, sample and nonsample, were used to collect data. Sections 1 through 20 and 28 through 32 of the sample form were identical to sections on the nonsample census form. Sample form sections 21 through 27 contained additional questions on usage of fertilizers and chemicals, farm production expenditures, value of machinery and equipment, value of land and buildings, farm-related income, and hired workers. There were 11 regional versions of the nonsample form and 13 regional versions of the sample form with listings of crops varying by region. These different forms were used to reduce the response burden of the census, while providing reliable information on a large number of data items.

The sample form was mailed to all mail list records in Alaska, Hawaii, and Rhode Island and to a sample of records in other States selected from the final mail list. Mail list records were selected into the sample with certainty if they (1) were expected to have large total value of agricultural products sold or large acreage, (2) were multi-unit operations (i.e., separate farms producing under one company organization), (3) were in a county with less than 100 farms in 1992, or (4) had other special characteristics. Farms with special characteristics were abnormal farms, such as institutional farms, experimental and research farms, and Indian reservations. Mail list records in counties containing 100 to 199 farms in 1992 were systematically sampled at a rate of 1 in 2; records in counties containing 200 to 299 farms in 1992 were systematically sampled at a rate of 1 in 4; and records in counties containing 300 or more farms in 1992 were systematically sampled at a rate of 1 in 6. The remaining mail list records not chosen to receive the sample form received the nonsample census form. This differential sampling scheme was used to provide reliable data for the sample sections of the report form for all counties.

EDITING DATA AND IMPUTATION FOR ITEM NONRESPONSE

The census of agriculture complex edit and imputation system is an automated computerized system that performed the following functions:

- Ensured reasonable relationships between/among data items, values for various sizes of farms, combinations of commodities, and economic interactions.

- Ensured necessary consistencies were present (there were more than 70 distinct consistency requirements).

- Ensured climatic, geographic, legal, and physical constraints were met.

The system performed these and similar functions for more than 900 data key codes for sample records and approximately 850 data key codes for nonsample records.

For the 1997 Census of Agriculture, as in previous censuses, all reported data were keyed and then edited by computer. The edits were used to determine whether the reports met the minimum criteria to be counted as farms in the census. The complex edit and imputation system provided the basis for deciding to accept, impute (supply), delete, or alter the reported value for each data record item.

Whenever possible, edit imputations, deletions, and changes were based on component or related data on the respondent's report form. For some items, such as operator characteristics, data for that record from the previous census were used when available. Values for other missing or unacceptable reported data items were calculated based on reported quantities and known fixed price parameters.

When these and similar methods were not available and values had to be supplied, the imputation process used information reported for another farm operation in a geographically adjacent area with characteristics similar to those of the farm operation with incomplete data. For example, a farm operation that reported acres of corn harvested, but did not report quantity of corn harvested, was assigned the same bushels of corn per acre harvested as that of the last nearby farm with similar characteristics that reported acceptable yields during that particular execution of the computer edit. The imputation for missing items in each section of the report form was conducted separately; thus, assigned values for one operation could come from more than one respondent.

Prior to the imputation operation, a set of default values and relationships was assigned to the possible imputation variables. The relationships and values varied depending on the item being imputed. For example, different default values were assigned for several Standard Industrial Classifications and total value of sales categories when imputing hired farm labor expenses. These values and item relationships for the possible imputation variables were stored in the computer in a series of matrices.

Each execution of the computer edit consisted of records from only one State sorted by reported State and county. For a given execution of the edit, the stored entries in the various matrices were retained in memory only until a succeeding record having acceptable characteristics for the same sections of the report form was processed by the computer. Then the acceptable responses of the succeeding operation replaced those previously stored. When a record processed through the edit had unreported or unacceptable data, the record was assigned the last acceptable ratio or response from an operation with a similar set of characteristics. Once each execution of the computer edit for a State was completed, the possible imputation variables were reset to the default values and relationships for subsequent executions. An edit run usually consisted of 10,000 or more records.

After the initial computer edit, all keyed reports not meeting the census farm definition were reviewed to ensure that the data had been keyed correctly. Edit referrals were generated for 17 percent of the reports included as farms; they were reviewed for keying accuracy and to ensure that the computer edit actions were correct. If the results of the computer edit were not acceptable, corrections were made and the record re-edited.

CENSUS NONSAMPLING ERROR

The accuracy of the census counts is affected jointly by sampling errors and nonsampling errors. Extensive efforts were made to compile a complete and accurate mail list for the census, to design an understandable report form with instructions, and to minimize processing errors through the use of quality control measures. Nonsampling errors arise from many sources, including respondent or enumerator error or incorrect data keying, editing, or imputing for missing data. These nonsampling errors are further discussed in this section. Nonsampling error due to mail list incompleteness and duplication as well as misclassification of records on the mail list is called coverage error. The section titled ''Coverage Evaluation'' discusses the evaluation studies conducted to measure the extent of this error in the census.

Respondent and Enumerator Error

Incorrect or incomplete responses to the census report form or to the questions posed by an enumerator can introduce error into the census data. To reduce reporting error, detailed instructions for completing the report form were provided to each respondent. Questions were phrased as clearly as possible based on previous tests of the report form. In addition, each respondent's answers were checked for completeness and consistency by the complex edit and imputation system.

Item Nonresponse

As information flowed from data collection to tabulation, various types of item nonresponses were identified on the census report forms. Nonresponse to particular questions on the census report form that logically should have been present created a type of nonsampling error in both complete count and sample count data. In this case, information from a similar farm was used to impute for these missing data items. The resulting data may have been biased if the characteristics of the nonreporting respondents were different from those of reporting respondents for those items.

Processing Error

All phases of processing for each census report form were potential sources for the introduction of nonsampling error. An automated check-in recorded that the report had been returned and excluded from further followup mailings. Approximately one-third of the mail returns were reviewed to resolve questions dealing with multiple reports, respondent remarks, or no reported data. The remaining mail returns (about two-thirds) were batched and sent directly to data keying, along with some of the reviewed cases containing farm data. Keyed records were transmitted, formatted, and run through the complex edit and imputation system. About one-fifth of all forms edited were clerically reviewed for inconsistencies, omissions, or questionable values. While reviewing these forms, the edit review staff determined if the action taken by the computer edit and imputation system was correct. Edited records were tabulated to the county level. Each county was reviewed and, when necessary, individual records were corrected prior to publication.

Developing accurate processing methods is complicated by the complex structure of agriculture. Among the complexities are the many places to be included, the variety of arrangements under which farms are operated, the continuing changes in the relationship of operators to the farm operated, the expiration of leases and the initiation or renewal of leases, the problem of obtaining a complete list of agriculture operations, the difficulty of contacting and identifying some types of contractor/contractee relationships, the operator's absence from the farm during the data collection period, and the operator's opinion that part or all of the operation does not qualify and should not be included in the census. During data collection and processing of the census, all operations underwent a number of quality control checks to ensure as accurate an application as possible.

COVERAGE EVALUATION

Coverage Overview

The primary objectives of the census of agriculture are to accurately count U.S. farms, measure commodity production and sales, and measure demographic characteristics of farm operators. Since 1945, an evaluation of census coverage has been conducted for each census of agriculture to provide estimates of the completeness of census farm counts. These results help to identify problems and focus improvements for future censuses.

According to coverage evaluation results, the past five censuses of agriculture included an average of 92 percent of U.S. farms and 98 percent of agriculture production. Complete enumeration of agricultural operations satisfying the farm definition of $1,000 or more in agricultural sales is complicated by the variety of arrangements under which farms are operated, the multiplicity of names used for an operation, the number of operations in which an operator participates, and the difficulty in classifying those operations just around the $1,000 sales range. In 1997, extensive efforts were made to compile as complete and accurate a mail list as possible, while reducing the duplication and number of nonfarm operations on the list.

The 1997 coverage evaluation program was designed to measure four components of error in the census farm counts. These components include:

1. Undercount due to farms Not on the Mail List (NML)
2. Overcount due to farms Duplicated or enumerated more than once (DUP)
3. Undercount due to farms Incorrectly Classified as nonfarms (ICU)
4. Overcount due to nonfarms Incorrectly Classified as farms (ICO).

The first component, mail list undercount, is by far the largest component of coverage error. Duplication, though occurring far less frequently, can involve larger farms and have a larger impact on acreage and sales estimates. The last two components involve the misclassification of either farms or nonfarms. Misclassification can arise from errors in either reporting or processing the data.

Table A - Coverage Estimates - illustrates the effect of coverage adjustments on census farm counts by demographic characteristics, land in farms, and total value of sales. The coverage total is defined as the net difference between undercounted and overcounted farms. The adjusted census total is the sum of the census total and the net coverage total. The relative standard error is shown for the final census coverage adjusted number. This number will be similar to the relative standard error for the census number, except when the coverage total is negative or close to zero. The coverage adjustment percentage shows the coverage total as a percentage of total census adjusted farms for that characteristic.

The 1997 Census of Agriculture is the first census to include all four components of coverage error in the coverage table. Previous publications only included the coverage error component due to farms not on the mail list (NML). Because of this, caution should be taken when comparing coverage estimates with previous years. In addition, the coverage total is a negative number for some characteristics. This means that the number of farms overcounted for this characteristic was greater than the number of farms undercounted.

Area Frame Surveys to Measure Mail List Undercoverage

Names and addresses collected in the 1997 June Agricultural Survey and 1997 Fall Area Survey were used to estimate the undercount due to farms not on the census mail list (NML). These names were matched to the census mail list, and those that did not match were contacted by telephone or person. The enumerator verified whether the operation had reported in the census, and if not, a census of agriculture report form was completed.

The percentage of farms missed in the census varies considerably by State. In general, farms not on the mail list tended to be small in acreage, production, and sales of agricultural products. Farm operations could be missed for various reasons, including the possibility that the operation started after the mail list was developed, the operation may be so small as not to appear in any agriculture-related source lists, or the operation may have been falsely classified as a nonfarm prior to mailout.

Classification Error Survey to Measure Three Types of Coverage Error

The remaining three types of coverage error were measured by the Classification Error Survey. This survey was used to estimate the number of farms counted more than once (DUP), the number of farms misclassified as nonfarms (ICU), and the number of nonfarms misclassified as farms (ICO). A sample of census of agriculture respondents was selected for reinterview to determine their farm/nonfarm status and collect information to identify potential duplication. The farm classification from this interview was compared with the classification on the census of agriculture report form. Any differences between these two classifications were reconciled to determine the true farm status. Each operation was reviewed for duplication by matching the additional information received from the reinterview (landlords, tenants, other names, etc.) to the list of census respondents. Potential duplication was reviewed and discrepancies reconciled.

In general, the classification error rate is higher for small farms close to the $1,000 agricultural sales requirement. This rate is also higher for farms with small acreage (less than 49 acres), higher for tenant farms than for full- or part-owner farms, and higher for farms where farming is not the operator's principal occupation.

Coverage Estimation

The adjusted census total, T, is estimated as the census farm count, C, plus undercount and minus overcount adjustments. Undercount includes 1) farms not on the mail list (NML) and 2) farms incorrectly classified as nonfarms (ICU). Overcount includes 3) nonfarms incorrectly classified as farms (ICO) and 4) farms duplicated in the census (DUP). Altogether, the adjusted census total is:

$$T = C + (NML + ICU) - (ICO + DUP).$$

In some States, estimates of misclassification of farms owned by operators having rare demographic characteristics were based on particularly small sample sizes. Where such small sample sizes occurred, a form of small area estimation was used in which data from similar States contributed to that State's estimates. In these cases, the coverage totals are weighted totals of the direct State estimate and the direct estimate from the region. Direct estimates were used to the largest extent possible, based on the amount of survey cases available for the particular item being estimated.

Table A. **New England Coverage Estimates: 1997**

[For meaning of abbreviations and symbols, see introductory text]

Item	Census total	Coverage total[1]	Adjusted census		Coverage adjustment (percent)
			Total	Relative standard error (percent)	
Farms .... number..	24 571	7 008	31 579	3.8	22.2
Land in farms .... acres..	3 821 702	410 022	4 231 724	2.9	9.7
Average size of farm .... acres..	156	59	134	(X)	(X)
Farms by size of farm:					
Less than 10 acres	3 491	1 022	4 513	14.3	22.6
10 to 49 acres	6 466	3 508	9 974	9.5	35.2
50 to 179 acres	8 080	1 933	10 013	7.1	19.3
180 acres or more	6 534	545	7 079	4.7	7.7
Farms by value of sales:					
Less than $2,500	7 539	4 642	12 181	8.2	38.1
$2,500 to $9,999	6 309	1 271	7 580	8.0	16.8
$10,000 or more	10 723	1 095	11 818	4.0	9.3
Market value of agricultural products sold .... $1,000..	1 988 736	16 872	2 005 608	.9	.8
Farms by type of organization:					
Individual or family	20 591	6 833	27 424	4.2	24.9
Partnership, corporation, or other	3 980	175	4 155	4.8	4.2
Farms by tenure of operator:					
Full owners	15 759	4 971	20 730	4.8	24.0
Part owners	6 961	1 693	8 654	7.0	19.6
Tenants	1 851	344	2 195	17.6	15.7
Operators by place of residence:					
On farm operated	19 638	6 375	26 013	4.2	24.5
Not on farm operated	3 488	704	4 192	7.8	16.8
Not reported	1 445	–71	1 374	20.6	–5.2
Operators by principal occupation:					
Farming	12 553	1 229	13 782	4.1	8.9
Other	12 018	5 779	17 797	6.8	32.5
Operators by sex:					
Male	20 859	5 776	26 635	4.1	21.7
Female	3 712	1 232	4 944	9.6	24.9
Operators by race:					
White	24 464	6 987	31 451	3.8	22.2
Black and other races	107	21	128	58.6	16.4
Operators by years on present farm:					
4 years or less	2 279	1 065	3 344	16.8	31.8
5 years or more	18 854	5 403	24 257	4.1	22.3
Not reported	3 438	540	3 978	11.3	13.6

[1] See text in Appendix C regarding coverage estimates.