Appendix C. **Statistical Methodology**

MAIL LIST MODEL

Classification analysis was performed to predict the probability that an addressee on the 1992 mail list operated a farm, and thereby separated the preliminary mail list into probable farm and probable nonfarm classes. The analysis was used to reduce the preliminary census mail list of 3.78 million records to a final mail list size of 3.55 million records. All 3.55 million addresses on the final mail list received a census of agriculture report form.

Records from the 1987 final census mail list were used to build a 1992 prediction model for the 1992 analysis. Classification and Regression Trees (CART) software analyzed characteristics of known 1987 farm and nonfarm operations to determine which were most useful in predicting farm and nonfarm classes. Record characteristics such as the source of the mail list record, number of source lists on which the record appeared, expected value of agricultural sales, and geographic location were used to separate mail list records into model groups. (Sources included the previous agriculture census mail list, the Internal Revenue Service administrative records, U.S. Department of Agriculture, and special commodity lists.) The proportion of 1987 census farm records in each model group was calculated to provide an estimate of the probability that an addressee in the group operated a farm.

After the model groups were defined, each address record on the 1992 preliminary 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 according to the classification tree methodology. The model, followed by analyst reviews, was used to remove 229,700 records from the preliminary mail list (those in model groups with the lowest farm probability), and thereby designated the 3.55 million records with the highest farm probability to receive the census report form. This procedure was used to obtain a more complete census enumeration of farm operations without excessive respondent burden and data collection cost.

CENSUS SAMPLE DESIGN

Each of the 3.55 million name and address records on the census mail list was designated to receive one of three different types of census report forms. The three forms were the nonsample form, the screener form, and the sample form. Sections 1 through 20 and 27 through 32 of the sample form are identical to sections on the nonsample form. The sample form, sections 21 through 26, contains additional questions on usage of fertilizers and chemicals, farm production expenditures, value of machinery and equipment, value of land and buildings, and farm-related income. The screener form is identical to the nonsample form with questions added in section 1 to allow quick identification of nonfarm addresses. These three 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. Addresses were selected into the sample with certainty (1) if they were expected to have large total value of agricultural products sold or large acreage, (2) if they were multiunit operations (i.e., separate farms in more than one location), (3) if they had other special characteristics, or (4) if they were in a county with less than 100 farms in 1987. Other addresses in counties containing 100 to 199 farms in 1987 were systematically sampled at a rate of 1 in 2, and other addresses in counties containing 200 farms or more in 1987 were systematically sampled at a rate of 1 in 6. This differential sampling scheme was used to provide reliable data for the sample sections of the report form for all counties. When a nonsample large farm was identified during processing, a supplemental form that contained the additional sample data inquiries was mailed.

To determine which mail list records would receive the screener form, all mail list records not designated for the sample were sorted by model group farm probability as specified by the mail list model. The 412,000 mail list records in the model groups with the lowest probability of being farms and with an expected total value of agricultural product sales less than $25,000 were designated to receive the screener report form. The remaining mail list records received the nonsample census report form.

CENSUS NONSAMPLING ERROR

The accuracy of the census counts are affected by 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 on specific operations. Nonsampling errors arise from incompleteness of the census mail list, duplication in the mail list, incorrect data reporting, errors in editing of reported data, and errors in imputation for missing data. These specific nonsampling errors are further discussed in this section.

Respondent and Enumerator Error

Incorrect or incomplete responses to the mailed census report form or to the questions posed by a telephone enumerator introduce error into the census data. Such incorrect information can lead, in some cases, to incorrect enumeration of farms. To reduce all types of reporting error, detailed instructions for completing the report form were provided to each addressee. Questions were phrased as clearly as possible based on tests of the census report form and each respondent's answers were checked for completeness and consistency.

Item Nonresponse

As information flows from data collection to tabulation, various types of item nonresponses are identified on the census report forms. Nonresponse to particular questions on the census report form that logically should be present may create a type of nonsampling error in both complete count and sample count data. When information from reporting farms is used to edit or impute for item nonresponse, the data may be biased due to characteristics of the nonreporting respondents differing from those reporting the item. Any attempt to correct the data items may not completely reflect this difference either at the element level (individual farm operation) or on the average.

Processing Error

All phases of processing for each census report form are sources for the introduction of nonsampling error. The processing of the census report forms includes clerical screening for farm activity, computerized check-in of report forms and follow-up of nonrespondents, keying and transmittal of completed report forms, computerized editing of inconsistent and missing data, review and correction of individual records referred from the computer edit, review and correction of tabulated data, and electronic data processing. These operations undergo a number of quality control checks to ensure as accurate an application as possible, yet some errors are not detected and corrected.

EDITING DATA AND IMPUTATION FOR ITEM NONRESPONSE

The Census of Agriculture Complex Edit and Imputation System performs the following functions:

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

- Ensuring necessary consistencies are present. There are more than 70 distinct consistency requirements.

- Ensuring geographic, legal, and physical constraints are met.

The system must perform these and similar functions for 900 data keycodes for sample records and 850 data keycodes for nonsample records.

For the 1992 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 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 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 were 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 classification 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. The computer records were 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 some 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.

After the initial computer edit, keyed reports not meeting the census farm definition were reviewed to ensure that the data were keyed correctly. Edit referrals were generated for about 25 percent of the reports included as farms; they were reviewed for keying accuracy 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 was reedited.