

Overview of the 2007 Census of Agriculture Methodology

Dave Aune¹, Chris Messer²

¹USDA's National Agricultural Statistics Service, 1400 Independence Avenue, Washington, DC 20250

²USDA's National Agricultural Statistics Service, 1400 Independence Avenue, Washington, DC 20250

1. Development of The Census Report Forms

Prior to release of the results from the 2002 Census of Agriculture, NASS was preparing for the 2007 Census of Agriculture. The first team established was the 2007 Census Content Team. This team was tasked with content determination and report form development. They reviewed the 2002 report form content, solicited input from internal and external customers, developed criteria for determining acceptance and/or rejection of content for the 2007 Census of Agriculture report forms, tested the effectiveness of the report forms for various modes of data collection (mail, telephone, personal interview, and electronic data reporting), and made recommendations to NASS senior executives for final determination.

Throughout development NASS sought advice and input from the data user community. Integral partners included the Advisory Committee on Agriculture Statistics, State departments of agriculture and other State government officials, Federal agency officials, land grant universities, agricultural trade associations, media, and various Community Based Organizations.

NASS conducted the 2005 Census of Agriculture Content Test in early 2006. The test consisted of three phases: cognitive pretesting, national mail-out, and follow-up interviews. Results from the testing produced two final report form types -- a 24-page regionalized form with 7 versions (07-A0201 thru 07-A0207 regional forms and an 07-A0200 general version) and a 12-page national form version (07-A0100). The main difference between the form types is the format used to collect crop and livestock information. The regionalized report forms include crop sections designed to facilitate reporting crops most commonly grown within a report form region. Many items in these sections are either prelisted in the tables or listed below the tables. The national report form collected the same information as the regionalized forms, but it was formatted to fit on fewer pages. It includes an open table format to collect crop and livestock information. Respondents had to write in their crops and/or livestock information. See figure 1.

1.1 Data Changes

Descriptions of the report form changes and their effect on the publication tables are documented in Appendix B of Volume 1, U.S. Summary and can be located on the Internet at:

http://www.agcensus.usda.gov/Publications/2007/Full_Report/index.asp.

2. The Census Population

2.1. The Census Mail List

The National Agricultural Statistics Service (NASS) maintains a list of farmers and ranchers from which the Census Mail List (CML) is compiled. The goal is to build as complete a list as possible of agricultural places that meet the NASS farm definition, that is, an operation that produces, or would normally produce and sell, \$1,000 or more of agricultural products per year.

The CML compilation begins with the list used to define sampling populations for NASS surveys conducted for the agricultural estimates program. Each record on the list includes name, address, and telephone number plus additional information that are used to efficiently administer the census of agriculture and agricultural estimates programs.

NASS builds and improves the list on an ongoing basis by obtaining outside source lists. Sources include State and federal government lists, producer association lists, seed grower lists, pesticide applicator lists, veterinarian lists, marketing association lists, and a variety of other agriculture-related lists. NASS also obtains special commodity lists to address specific list deficiencies.

These outside source lists are matched to the NASS list using record linkage programs. Most names on newly acquired lists are already on the NASS list. Records not on the NASS list are treated as potential farms until NASS can confirm their existence as a qualifying farm. Staff in NASS field offices routinely contact these potential farms to determine if they meet the NASS farm definition. For the 2007 Census of Agriculture, NASS made a concerted effort to work with Community-Based Organizations not only to improve list coverage for minorities but also to increase census awareness and participation.

List building activities for developing the 2007 CML started in 2004. Between 2004 and 2007, NASS conducted a series of Agricultural Identification Surveys (AIS) on approximately 1.7 million records, which included nonrespondents from the 2002 census and newly added records from outside list sources. The AIS report form collected information that was used to determine if an operation met the NASS farm definition. If the definition was met, the operation was added to the NASS list and subsequently to the CML. Addressees that were nonrespondents were also added to the CML and identified with a special status code.

Measures were taken to improve name and address quality. Additional record linkage programs were run to detect and remove duplicate records both within each State and across States. List addresses were processed through the National Change of Address Registry and the Locatable Address Conversion System to ensure they were correct and complete. Records on the list with missing or invalid phone numbers were matched against a nationally available telephone database to obtain as many phone numbers as possible.

The official CML was established on September 1, 2007. The list contained 3,194,373 records. There were 2,198,410 records that were thought to meet the NASS farm definition and 995,963 potential farm records, which included AIS nonrespondents, other records added to the CML by the NASS field offices, and late adds to the CML that were not included in any previous AIS or State screening survey. See figure 2.

2.2 Not on the Mail List

To account for farming operations not on the CML, NASS used its area frame. The NASS area frame covers all land in the U.S. and includes all farms. The land in the U.S. is stratified by characteristics of the land. Segments of approximately equal size are delineated within each strata and designated on aerial photographs. A probability sample of segments is drawn within each strata for the NASS annual area frame survey, known as the June Agricultural Survey (JAS). The JAS sample of segments is allocated to strata to provide accurate measures of acres planted to widely grown crops and inventories of hogs and cattle. Sampled segments in the June Survey are personally enumerated. Each operation identified within a segment boundary is known as a tract.

The 2007 JAS sample was allocated to strata so that it would provide additional measures of small and minority owned farms. The 2007 JAS consisted of 10,912 regular sampled segments, supplemented with 3,692 Agricultural Coverage Evaluation Survey (ACES) segments - segments selected to provide measures of small and minority owned farms. These additional ACES segments targeted farming demographics that typically had lower coverage rates on the list.

The information from each tract (operation) within a segment is matched against operations on the NASS list to determine the amount of undercoverage that exists for a wide range of farming sectors and farmer demographics. The names and addresses collected in the 2007 JAS and 2007 ACES were matched to the CML and checked for duplication. Farms from the June 2007 survey that did not match were determined to be Not on the Mail List (NML) and sent a report form of a different color to be easily identified. Data from the NML operations provided a measure of the undercoverage of the CML operations.

Instructions on the census report form guided the respondent to complete the CML form and mail back both CML and NML forms together if duplicate forms were received. Those who returned a CML census form and an NML census form had been erroneously classified as NML and were removed from the NML.

The percentage of farms not represented on the CML varied 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 were missed for various reasons, including the possibility that the operation started after the mail list was developed, the operation was so small that it did not appear in any agriculture-related source lists, or the operation was erroneously classified as a nonfarm prior to mailout.

The NML consisted of 12,821 tracts. The CML was used with the NML in multiple frame estimation to represent all farming operations across all States, with the exception of Alaska. It is financially and logistically unfeasible to maintain an area frame in Alaska due to its vast land mass and relatively sparse agriculture.

3. Data Collection

3.1 Method of Enumeration

Mailout and mailback was the primary data collection method. It was supplemented with Electronic Data Reporting (EDR) on the Internet and non-response follow-ups by telephone and personal enumeration. The enumeration methods used in the 2007 census were similar to those used in the 2002 census.

3.2 Report Forms

A master report form was developed that included all data items to be collected in the census. From the master, two types of report forms were developed to be used in the 2007 census - a regionalized report form with 7 versions and a national report form. Each of the 24-page regionalized report forms (07-A0201, 07-A0202, 07-A0203, 07-A0204, 07-A0205, 07-A0206, 07-A0207) were designed to facilitate reporting crops most commonly grown within the report form region. The 12-page national report form (07-A0100) was designed for operations throughout the country with few commodities. The national report form collected the same information as the regional form, but it was formatted to fit on fewer pages. All of the forms allowed respondents to write in specific commodities that were not identified on their form. The national form was mailed to approximately 528,000 addresses on the CML (about 20 percent) and the regional form was mailed to 2.67 million addresses on the CML (about 80 percent).

3.3 Report Form Mailings and Respondent Follow-up

The initial mailout took place at the end of December 2007. Approximately 3.2 million packets were mailed. Each packet contained a cover letter, instruction sheet, a labeled report form, and a return envelope. Mailout packet preparation, initial mailout, and two follow-up mailings to nonrespondents were handled by the Census Bureau's National Processing Center (NPC) in Jeffersonville, IN. The first follow-up was mailed during the last two weeks of February 2008 to approximately 1.3 million nonrespondents. The second follow-up was mailed the beginning of April 2008 to approximately 1.0 million nonrespondents. Additionally, NPC received, checked-in, scanned, and keyed (from image) returned report forms. NASS statisticians on site at NPC provided technical guidance and monitored NPC processing activities.

Select groups of census records were identified to receive special handling procedures. Report forms were labeled at NPC and shipped to the field offices for enumeration. These respondents were excluded from the initial and both follow-up mailings, and were referred to as "must" operations. Each "must" operation was enumerated by telephone or face-to-face. If a record was determined to be no

longer in operation, their non-farm status was verified and documented. The field offices were responsible for enumerating or resolving all non-response "must" records in their State. Computer Assisted Telephone Interview (CATI) calling for nonrespondent "must" records was conducted between March 2008 and June 2008. Once enumerated, the report forms were either sent to NPC for check-in and data capture or the data were keyed directly from the form at the field office. The 169,000 "must" records fell into one of five groups.

The first "must" group consisted of 46,000 records "tagged" by the NASS field offices for personal enumeration rather than mailout and mailback enumeration. The second "must" group consisted of 4,000 "specialized" records including such operations as grazing associations, governmental units, research farms, college farms, etc. The third "must" group was characterized by location. All 3,000 records in Alaska and Rhode Island were identified as "must" records because census statistics for these two States were based on responses to the CML because nonresponse was not permitted. The last two groups consisted of a total of 116,000 records expected to have either a large number of acres in farm land or a large value of sales. Threshold levels were identified for each State.

Advanced Follow-up was conducted between February 2008 and April 2008. It focused on three groups of nonrespondents that included: respondents least likely to respond because they were nonrespondents to the 1997 and 2002 Censuses of Agriculture, even though they may have responded to other NASS surveys; respondents viewed as easy and quick interviews based on expected sales of zero, including respondents who received Conservation Reserve Program (CRP) payments and respondents to the AIS with expected future sales; and new records whose farm status was uncertain due to unsuccessful earlier screening attempts. The field offices conducted CATI and field enumeration for operations in their State. This phase was followed by Low-Response County Follow-up to attempt to reach a minimum response rate of at least 75 percent in all counties. It was conducted by the field offices using CATI between March 2008 and June 2008. See Figure 3.

4. Report Form Processing

4.1 Data Capture

All report forms returned to NPC were immediately checked in, using bar codes printed on the mailing label, and removed from follow up mailings. All forms with any data were scanned and an image was made of each page of a report form. Optical Mark Recognition (OMR) was used to capture categorical responses and to identify the other answer zones in which some type of mark was present.

Data entry operators keyed data from the scanned images using OMR results that highlighted the areas of the report forms with respondent entries. The keyer evaluated the contents and captured pertinent responses. Ten percent of the captured data were keyed a second time for quality control. If differences existed

between the first keyed value and the second, an adjudicator handled resolution. The decision of the adjudicator was used to grade the performance of the keyers, who were required to maintain a certain accuracy level.

The images and the captured data were transferred to NASS's centralized network and became available to field offices and headquarters on a flow basis. The images were available for use in all stages of review. Images were computer generated for reports obtained from the telephone interviews and the Internet. See Figure 4.

4.2 Editing Data

Captured data were processed through a format program. The program verified that record identifiers were valid and checked the basic integrity of the data fields. Rejected records were referred to analysts for correction. Accepted records were sent to a batch edit process. Each execution of the computer edit in batch mode consisted of records from only one State and flowed as the data were received from NPC.

All 2007 census records were passed through a complex computer edit. The edit determined whether a reporting operation met the minimum criteria to be counted as a qualifying farm (in-scope). Operations failing to meet the minimum criteria (out-of-scope) were referred to analysts for verification. The edit examined each in-scope record for reasonableness and completeness and determined whether to accept the recorded value for each data item or take corrective action. Actions included removing erroneously reported values, replacing an unreasonable value with one consistent with other reported data, or providing a value for an overlooked item. To the extent possible, the edit determined a replacement value. Strategies for determining replacement values are discussed in the next section.

The edit systematically checked reported data section-by-section with the overall objective of achieving an internally consistent and complete report. NASS subject-matter experts defined the criteria for acceptable data. Problems that could not be resolved within the edit were referred to an analyst for intervention. Analysts in the NASS field offices used additional information sources, examined the scanned image, and determined an appropriate action. Field office analysts used an interactive version of the edit program to submit corrected data and immediately re-edit the record to ensure a satisfactory solution.

4.3 Imputing for Missing Data

Missing data occurred whenever a respondent failed to report in a cell that should have a positive value or when the edit determined a value was not reasonable and should be changed. The edit performed a sequence of steps that determined the best value to impute for the missing item. If an item could not be calculated directly from other data reported on the current form, the edit checked for previously reported data. Acreage, production, and inventory items may have been reported on a recent NASS crop or livestock survey. Operator characteristics, such as race and gender, were brought forward from the previous census if the operator had not changed in five years. Administrative data from the

Farm Service Agency was used for a few items, such as Conservation Reserve Program acreage. When these deterministic sources failed to produce a solution, the edit invoked an automated imputation system which searched for a reporting farm of similar type, size, and location to provide a value for the missing data item. If the imputation algorithm failed to provide a solution, the record was referred to an analyst for resolution. The guiding principal for imputation was to find a close match to the farm with the missing item. The census imputation algorithm relied on pre-established donor pools, one for each State. A donor pool included a collection of completed reports that had successfully navigated the edit.

Each pool was further divided into groups of similar type and size, referred to as profiles. When the edit determined the need to impute an item, it went to the appropriate profile and searched for the best fit. Best fit was determined by calculating "distance" between the incomplete report and each candidate donor using a set of match variables. Match variables were specific to each section of the report form and included the latitude and longitude of the principal county of operation. The distance was the sum of the squared differences between the reported values of the match variables. The donor with the smallest distance was considered the "nearest neighbor" and became the source for the imputation action. The value returned may have been a direct copy of the donor's value. In many cases, a relationship between two related variables on the donor record was applied to a reported value on the incomplete record. Using crop production as an example, the donor's production was divided by its harvested acres (yield) and multiplied by the recipient's harvested acres to obtain imputed production.

The imputation process was imbedded in the edit. When the edit determined an item required imputation, the edit program launched the algorithm, waited for a value to be returned, validated that the returned value was satisfactory, and resumed editing. Since imputation was conducted independently for each occurrence, reports requiring multiple imputations drew from multiple donors.

Initial donor pools were established before the first batch edits were run. These donor pools were "seeded" with 2002 census data that were "mapped" to look like 2007 data and passed through the 2007 edit to ensure they were consistent using the 2007 data relationships. In addition, data from the 2005 Census Content Test were similarly mapped and edited. As 2007 data were successfully processed, new records systematically replaced the older records in the donor pool. The older records disappeared entirely from the donor pool after the first few batch edits.

The donor pool for each State was refreshed weekly during the first couple of months of editing. As the flow of new data slowed, the donor pools were refreshed biweekly. During the early stages of editing, records that needed to impute production for field crops or hay were set aside. When the donor pool no longer contained old data, these records were brought back and passed through the edit, ensuring 2007 yields were imputed.

In some cases, nearest-neighbor imputation was not possible. The requirement of a positive imputed value could have ruled out all available donors, resulting in an imputation failure. An imputation failure could have occurred if there were no donors in the same profile as the report being edited. Records with imputation failures were either held until more records were available in the donor pool or referred to an analyst.

4.4 Data Analysis

The complex edit ensured the full internal consistency of the record. Successfully completing the edit did not provide insight as to whether the report was reasonable compared to other reports in the county. Analysts were provided an additional set of tools, in the form of listings and graphs, to review record-level data across farms. These examinations revealed extreme outliers, large and small, or unique data distribution patterns that were possibly a result of reporting, recording, or handling errors. Potential problems were researched and, when necessary, corrections were made and the record interactively edited again.

5. Disclosure Review

After tabulation and review of the aggregates, a comprehensive disclosure review was conducted. NASS is obligated to withhold, under Title 7, U.S. Code, any total that would reveal an individual's information or allow it to be closely estimated by the public. Cell suppression was used to protect the cells that were determined to be sensitive to a disclosure of information. Farm counts are not considered sensitive and are not subject to disclosure.

Based on agency standards, data cells were determined to be sensitive to a disclosure of information if they violated either of two criteria. First, the threshold rule was violated if the data cell contained less than three operations. For example, if only one farmer produced turkeys in a county, NASS could not publish the county total for turkey inventory without disclosing that individual's information. Second, a dominance rule was violated if the distribution of the data within the cell allowed a data user to estimate any respondent's data too closely. For example, if there are many farmers producing turkeys in a county and some of them were large enough to dominate the cell total, NASS could not publish the county total for turkey inventory without risking disclosing an individual respondent's data. In both of these situations, the data were suppressed and a "(D)" was placed in the cell in the census publication table. These data cells were referred to as primary suppressions.

Since most items were summed to marginal totals, primary suppressions within these summation relationships were protected by ensuring that there were additional suppressions within the linear relationship that provided adequate protection for the primary. A detailed computer routine selected additional data cells for suppression to ensure all primary suppressions were properly protected in all linear relationships in all tables. These data cells were referred to as complementary suppressions. These cells were not themselves sensitive to a

disclosure but were suppressed to protect other primary suppressions. A "(D)" was also placed in the cell of the census publication table to indicate a complementary suppression.

Field office analysts reviewed all complementary suppressions to ensure no cells had been withheld that were vital to the data users. In instances where complimentary suppressions were deemed critically important to a State or county, analysts requested an override and a different complement was chosen. See Figure 5.

6. Products

Many reports in various formats are produced from the census of agriculture. The goal is to make all of the reports available on the NASS Internet site: (<http://www.agcensus.usda.gov/>) with limited availability of hard copy reports. The first products to be released on February 4, 2009 were geographic data for the U.S., and 50 states, Puerto Rico, Guam, U.S. Virgin Islands, and Northern Mariana Islands. In addition, three Fact Sheets were made available that highlighted three different topic areas: farms numbers, demographics, and economics. Three on-line only products were made available as well- a query tool, a downloadable desktop application, and an agricultural atlas. Since the release date, five additional fact sheets, two special studies, and five regional and ranking reports have been released.

6.1 Custom Tabulations

Custom-designed tabulations may be developed when data are not published elsewhere. These tabulations are developed to individual user specifications on a cost-reimbursable basis and shared with the public. The census Volume 1 on CD-ROM is an alternative data source that should be investigated before requesting a custom tabulation.

All special studies and custom tabulations are subject to a thorough disclosure review prior to release to prevent the disclosure of any individual respondent data. Requests for custom tabulations can be submitted via the internet from the NASS home page, by mail, or by e-mail to:

DataLab
National Agricultural Statistics Service
Room 6436A, Stop 2054
1400 Independence Avenue, S.W.
Washington, D.C. 20250 - 2054
or
Datalab@nass.usda.gov

7. Tactical Plans

The Census Planning Branch in Headquarters is responsible for planning and executing the plan for each census of agriculture and the associated follow-on surveys. Eighteen full-time employees (FTE=s) are in the Census Planning Branch in Washington, D.C. Three FTE's are located in Jeffersonville, Indiana at the Census Bureau=s National Processing Center (NPC). One additional Census Bureau staff is detailed to NASS.

The employees at NPC are responsible for mail preparation, mailing, mail returns (check-in, sorting, report form scanning, key from image) and record retention. Employees in HQ either work in the Census Planning and Administration Section (CPAS) or the Census Statistics Section (CSS). CPAS employees serve as the points of contact for the Field Offices (FO's) and are known as Census Administrators. Census Administrators are responsible for seeing that the day-to-day activities are conducted according to the timeline and standards established in the Census Administration Manual (CAM). Six Census Administrators were employed from 2007 through most of 2008 - four covered the U.S., one covered Puerto Rico, and one covered Guam, U.S. Virgin Islands, and Northern Mariana Islands.

The FO's designated two points of contact in each location – one was the coordinator for the data collection activities and one was the coordinator for the data analysis. This structure ensured backup contacts and that no one person was overwhelmed by the longer survey process of fourteen months from mail out to publication of reports. In addition, Census Administrators granted rights to the various editing and analytical stages based on availability of staff resources. Improved esprit de corps resulted as we realized the potential of a synergistic relationship.

Timely communication is key to the overall ability to utilize staff across Divisions and Field Office locations to accomplish this project. The Census Administrators act as a liaison to the Field Offices and communicate via telephone, Email, official memorandum, and for-your-information memorandum. Sixty-eight official memoranda were sent from January 2008 to February 2009 and forty-four for-your-information memoranda were communicated during that same time period.

References

Volume 1, U.S. Summary and State Reports.
http://www.agcensus.usda.gov/Publications/2007/Full_Report/index.asp

Figure 1. 2007 Census Regions.

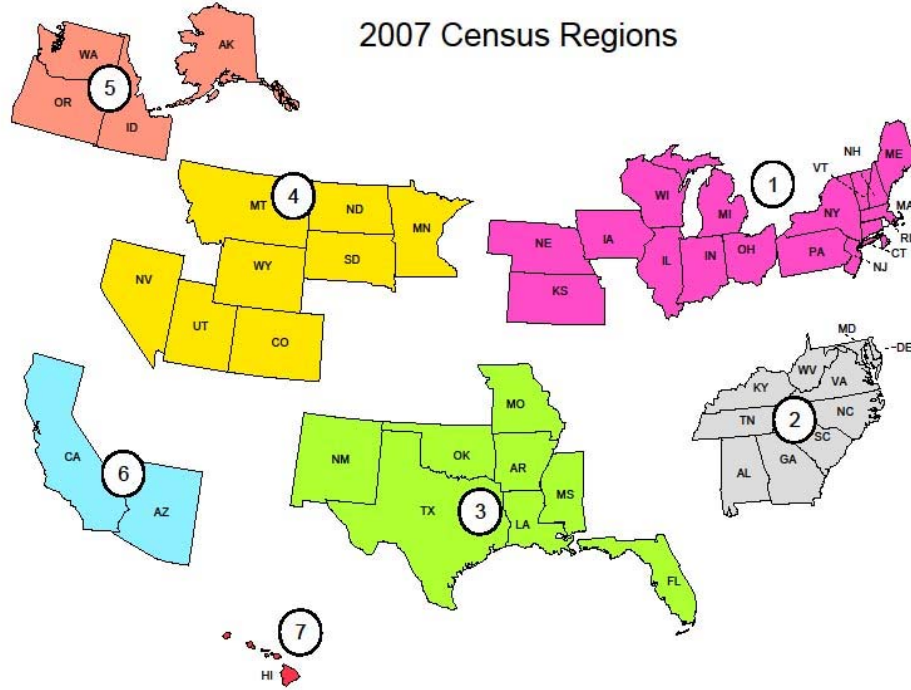


Figure 2. Official NASS Census Mailing List.

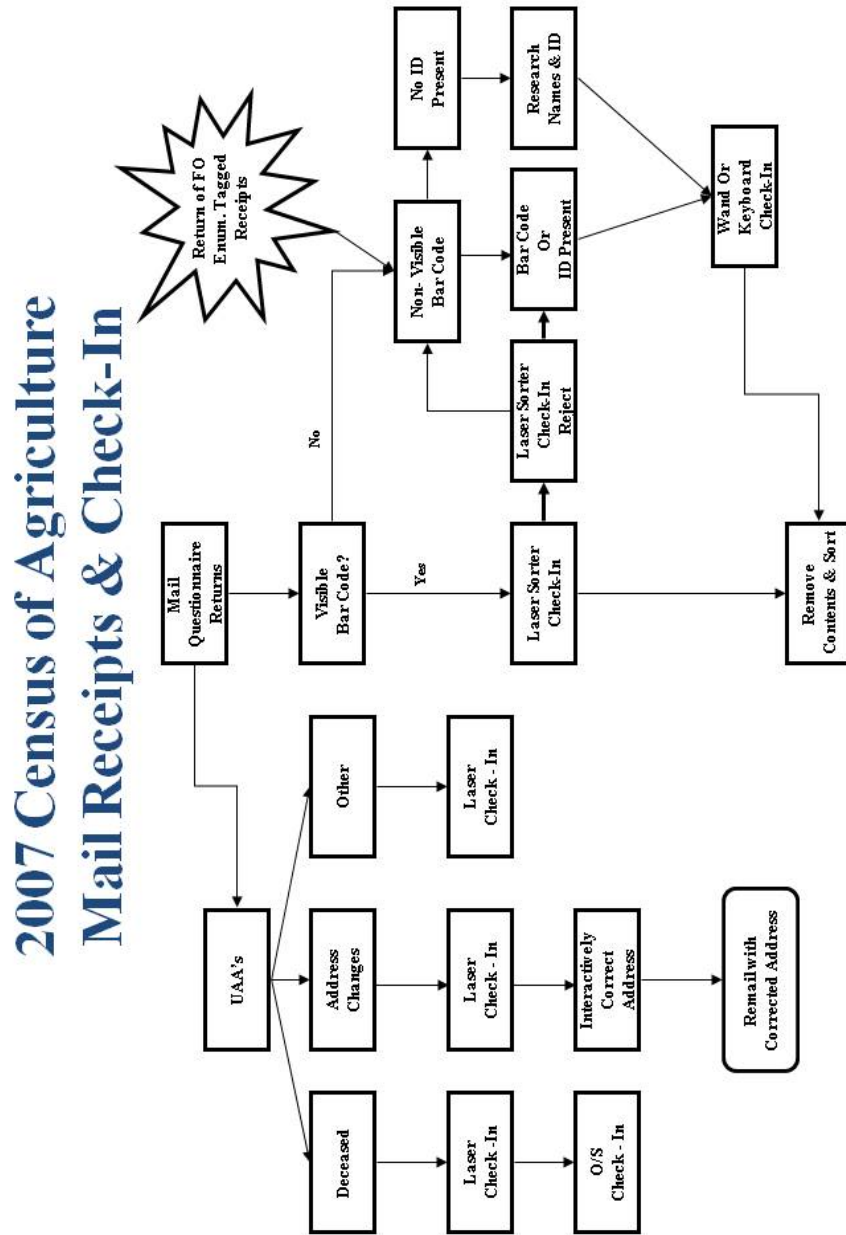
STATE	Potential Farm Records	NASS Farm Definition Records	Total Records
AL	20,021	47,372	67,393
AK	897	896	1,793
AZ	22,019	8,035	30,054
AR	27,809	61,225	89,024
CA	42,949	76,258	119,217
CO	11,065	38,676	49,741
CT	1,800	4,910	6,710
DE	1,441	2,688	4,129
FL	28,067	44,889	72,946
GA	42,411	43,110	85,541
HI	3,061	6,658	9,719
ID	8,817	24,132	32,949
IL	28,306	70,229	98,525
IN	24,555	58,510	83,065

IA	28,653	92,965	121,628
KS	16,686	63,510	80,196
KY	27,918	87,519	115,448
LA	18,350	28,603	46,953
ME	3,394	7,904	11,298
MD	5,119	13,799	18,918
MA	3,924	6,320	10,244
MI	22,899	59,384	82,273
MN	18,234	81,963	100,186
MS	42,369	36,787	79,176
MO	38,301	108,870	147,161
MT	7,345	36,888	44,233
NE	13,472	52,525	65,997
NV	2,338	2,950	5,288
NH	1,115	4,067	5,182
NJ	4,202	12,549	16,751
NM	15,747	16,821	32,568
NY	19,853	34,926	54,779
NC	30,306	48,296	78,612
ND	6,994	29,224	36,218
OH	25,161	78,505	103,636
OK	23,036	78,033	101,059
OR	22,981	39,398	62,379
PA	39,294	58,697	98,001
RI	340	1,065	1,405
SC	15,367	33,392	48,759
SD	15,353	32,471	47,824
TN	45,702	97,824	143,546
TX	136,204	256,020	392,224
UT	7,675	18,373	26,048
VT	1,058	6,448	7,506
VA	16,986	45,163	62,149
WA	24,764	32,565	57,329
WV	3,752	21,185	24,937
WI	24,669	74,907	99,566
WY	3,184	10,906	14,090
TOTAL	995,963	2,198,410	3,194,373

Figure 3. 2007 Census of Agriculture Major Processing Activities.

Activity	Date	Workload
Initial Mailout	Dec. 26, 2007	3,200,000
Thank You Card	Jan. 14, 2008	3,200,000
1 st Follow-Up	Feb. 12-Feb. 26, 2008	1,300,000
2 nd Follow-Up	Mar. 30- Apr.10, 2008	1,000,000
Check-In	Jan. 4- June 10, 2008	2,200,000
Open and Sort	Jan. 7- June 18, 2008	2,200,000
SP. Case Process	Jan. 7- June 18, 2008	330,000
2+ Processing	Jan. 7- June 18, 2008	50,000
Problem Solving	Jan. 7 - June 18, 2008	100,000
Batch/Guillotine	Jan. 14 – June 20, 2008	1,700,000
Scanning/KFI	Jan. 14 – June 20, 2008	1,700,000

Figure 4. 2007 Census of Agriculture Mail Receipts and Check-In



5

Figure 5. 2007 Disclosure Analysis – Summary of Final Disclosure Results.

Level	Total Cell Count	Total Suppressions	Percent of Total Cell Counts That Are Suppressions	Complementary Suppressions	Primary Suppressions
US	6,460	268	4.1%	190	78
US/State	206,275	26,520		10,952	15,568
US	6,460	268		190	78
STATE	199,815	26,252	13.1%	10,762	15,490
STATE and COUNTY	1,652,869	442,358		98,228	344,130
State	199,815	26,252		10,762	15,490
County	1,453,054	416,106	28.6%	87,466	328,640
Cross Tabs	1,000,325	140,836	14.1%	54,670	86,166
Total	2,659,654	583,462	21.9%	153,088	430,374