Four-Digits or No-Digit Social Security Numbers -- Impact on the National Agricultural Statistics Service Record Linkage Maintenance Processes

Denise A. Abreu, Kara Daniel, Bill Iwig, Stan Hoge
National Agricultural Statistics Service
1400 Independence Ave, SW Washington DC 20250, denise_abreu@nass.usda.gov

Key Words: Social Security Numbers, personally identifiable information[1] (PII), list frame, record linkage

Abstract

The National Agricultural Statistics Service (NASS) relies on Social Security Numbers (SSNs) and Employer Identification Numbers (EINs) as important matching variables for use in record linkage processing and other list maintenance activities that are conducted in order to maintain a high quality list of U.S. farmers and ranchers. Record linkage is used to match the NASS list frame to new lists and administrative data sources in order to identify new farmers and ranchers not previously identified by NASS. Record linkage is also used to match the list frame to administrative sources for maintenance purposes, for example updating telephone numbers on the list frame. However, maintaining nine digit SSNs/EINs on the list frame is a PII security concern for NASS. This research evaluated the potential impact of eliminating SSNs/EINs from the list frame or of only maintaining four digit SSNs/EINs on the various list building and maintenance record linkage processes. Overall, the results showed that using four digit SSNs/EINs for record linkage would result in missing 1 to 2 percent of the actual SSN matches and 4 to 8 percent of the actual EIN matches. However, having no SSNs/EINs for record linkage would result in missing 4 to 6 percent of the actual SSN matches and 9 to 13 percent of the EIN matches. In general, the percentage of missed matches will increase as the size of the data sets being matched increases. Since NASS processes several record linkage projects per state per year, the cumulative effect of either approach on the quality of the list frame is a concern.

I. Introduction

The National Agricultural Statistics Service (NASS) spends considerable effort safeguarding both respondent and employee personally identifiable information (PII). In a 2007 memorandum sent to all employees regarding PII, NASS‟ former administrator, Ronald R. Bosecker, wrote “It is critically important for all NASS employees and contractors to understand the definition of PII, recognize PII when they encounter it, and understand their responsibility for safeguarding it.” Furthermore, he emphasized the agency‟s incredible track record in handling PII. He wrote “We have been proud throughout our history that we protect the confidentiality of all information relating to those who entrust us with their data. It is the responsibility of each employee and contractor, regardless of grade, job series, or location throughout the country.” It is very clear in his message that safeguarding PII is not only important to NASS‟ employees but also a major focal point of its upper management team. The agency complies with all Federal Information Security Management Act (FISMA) regulations in securing all its Information Technology (IT) systems. As required by FISMA, NASS conducts Privacy Impact Assessments (PIA) to evaluate how the agency processes, handles, and stores privacy information. As required by the Privacy Act of 1974, NASS also published System of Record Notices (SORN) on the Federal Register. These publicly available notices document how NASS uses, stores, retrieves, etc. public records that are under its control. Both PIA and SORN help NASS senior management assess its security posture in safeguarding public information that are entrusted to NASS.

The ability to request PII from respondents has been greatly jeopardized due to numerous reports of loss of such information by other government agencies in the past few years. These highly scrutinized events raised a great deal of controversy and the public‟s trust in the government‟s ability to safeguard their information was weakened. In an attempt to restore the public‟s

[1] The term “personally identifiable information” refers to the information which can be used to distinguish or trace an individual‟s identity, such as their name, social security number, biometric records, etc. either alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother‟s maiden name, etc.

trust, the Office of Management and Budget (OMB) began a government wide awareness initiative on the collection and safekeeping of PII. Additionally, OMB tightened the rules around the collection of PII. In response, government agencies began implementing measures to eliminate or diminish the collection of PII, specifically Social Security Numbers (SSNs) and Employer Identification Numbers (EINs). In 2006, NASS voluntarily stopped requesting SSNs and EINs on all its major surveys. In addition, NASS has considered completely eliminating SSNs and EINs from their processing systems or only maintaining 4-digit SSNs and EINs to help reduce the risk of any inappropriate release of sensitive PII.

NASS relies on SSNs and EINs as important matching variables for use in record linkage processing and other list maintenance activities that are conducted in order to maintain a high quality list of U.S. farmers and ranchers. Record linkage is used to match the NASS list frame to new lists and administrative data sources in order to identify new farmers and ranchers not previously identified by NASS. Some of these new sources contain SSNs and EINs, which are valuable matching variables. Record linkage is also used to match the list frame to administrative sources for maintenance purposes, for example updating telephone numbers on the list frame. Other list maintenance activities also utilize SSNs and EINs, such as interactive look-ups of records on databases.

This research report evaluates the potential impact of eliminating SSN/EIN from the list frame or of only maintaining four digit SSN/EIN on the various list-building and maintenance record linkage processes and other maintenance activities.

II. Record Linkage

NASS began utilizing its current record linkage system over 9 years ago. The NASS system was built using AutoStan and AutoMatch (formerly sold by MatchWare Technologies) as its base for the standardization and matching of records. These software programs were developed using the probabilistic record linkage techniques proposed by Ivan Fellegi and Alan Sunter in their 1969 JASA paper. NASS developed front and back end companion products to assist in setting up record linkage match parameters and reviewing results.

NASS begins a typical record linkage project by obtaining an outside list source. This list is then transformed into a standard fixed field ASCII text file. Data in this file are formatted such that they meet the standards of the List Frame. Individual names are transformed to signature format, and variable length restrictions are imposed so that the length of fields going into record linkage match the length of the corresponding fields on the List Frame. For example, name and address fields are limited to 30 characters on the List Frame, so name and address fields for each record from the outside source list are reformatted to 30 characters. A list identification number is also generated for each record in the outside source list, allowing the processor to easily identify the record before and after the match is run. Each outside source list is further assigned a record source code, uniquely identifying that list source. Finally, each outside source record is assigned a status code. Examples of the status code are: known farm, potential farm, or non-farm record. The status code is continually updated based on information received about the record.

Before processing a record linkage project, a fixed field ASCII text file is pulled from the NASS List Frame database. The layout of this file is identical to the layout of the outside source ASCII file. The List Frame extract typically contains all records on the frame, including known farms, potential farms, and non-farms.

After the outside source and List Frame fixed field ASCII text files are created, a SAS program is run to determine if the cities and ZIP Codes from the outside source correspond to the United States Post Office standard cities and ZIP Codes. If a postal standard city and ZIP Code cannot be found, a report is generated noting that the city or ZIP Code information is incorrect and should be updated before the match is processed. The SAS program also verifies that telephone numbers, SSNs, and EINs meet certain validity standards. If they are not valid, a report is generated. Once the verify program is finished and the file is free of all errors, a standardization process is run on both the outside source and the List Frame files.

The standardization process parses the names and addresses into their component parts. For example, a person name could be parsed into a prefix, first name, middle name, last name, last name suffix and title. During the standardization process, input

name and address components are replaced with standard values. This standardization removes the effect of common nicknames and spelling variations. It also ensures that like information is compared during the match process.

NASS utilizes AutoMatch software to match the outside source and List Frame files. For each project, a set of blocking variables is used to divide the data into mutually exclusive subgroups. Records with common values for the blocking variables are compared during the match process, and records where the blocking variables differ are considered non-matches. For example, one pass may block on the ZIP Code. Records with common ZIP Codes are compared during the match process, while records with different ZIP Codes are considered non-matches. Multiple passes with different blocking and matching variables are run to compensate for inaccuracies in the data. Once the blocks have been determined, values for a series of match variables are compared. If the values agree, a positive weight is generally assigned. If the values disagree, a negative weight is generally assigned. The weights for each of the variables are then summed up to come up with a composite weight, which is a measure of the likelihood that a record pair is a match.

After a pass is run, a report is generated showing possible links. Linked pairs are sorted according to their composite weights, and links are reviewed in SAS. During the review process, two cutoff values, referred to as the upper and lower cutoffs, are set for each pass. Any record pairs with a composite weight higher than the upper cutoff are considered matches; record pairs with composite weights lower than the lower cutoff value are considered non-matches; and record pairs with weights between the two values are considered possible matches. The possible match records must be manually reviewed before a final review status will be set. Ideally this review would be done between each pass; however, this would become logistically infeasible because review of possible matches is done by personnel in the Field Offices. Rather than review possible matches between each pass, NASS combines the results from all passes into one final review using a SAS program. This SAS program brings all related records together into groups which NASS refers to as link groups. A link group contains all match pairs involving the same outside source or List Frame records. Additionally, all records marked on the List Frame as associated with an operation are brought into the group (for example, partner or manager records). Each link group is then classified as either a match, possible match, or non-match based on how the pairs that make up the link group were classified during the matching process.

NASS developed its own resolution system for reviewing record linkage projects. After an outside list source has been matched to the List Frame, a record linkage database is populated with the results. This database is independent of the List Frame. However, the resolution system has many features which allow those performing resolution to view and update List Frame information as needed. Field Office (FO) staff does the majority of the resolution work. The employees in the FO have experience with the agriculture in that State and work closely with important State Agriculture contacts. To resolve the link group, FO staff reviews the records and determines the match status. At times, phone calls are made to verify operating arrangements.
When the FO reviews the link groups, the reviewer goes through the link group to determine which List Frame record best matches the outside source record. If an outside source record does not best match the first List Frame record in the link group, the reviewer can change the link group number so that the outside source record and the best List Frame record are in the same link group. The reviewer can create up to 10 new link groups (sub link groups) for a particular existing link group. The NASS record linkage system creates a composite record which represents each operation contained in a link group. Once all the records are reviewed and each one is grouped with its best List Frame record, a composite record is regenerated for each sub link group. These composite records are the ones used in generating any needed transaction files for possible name and address, control data, and any other updates. Additionally, the composite records are used to create new add records to be posted to the List Frame. Users have the ability to alter information in the composite record so that it represents the operation as completely and accurately as possible. See Appendix A for a flowchart of the entire record linkage processing guideline.

III. Methodology

Analysis was conducted to evaluate the impact of not using SSNs and EINs and of just using four digit SSNs and EINs on record linkage and NASS current list maintenance activities and procedures. The 2007 Farm Service Agency (FSA) data files for Arizona (AZ), Montana (MT), South Carolina (SC), South Dakota (SD), and Texas (TX) were used for matching to the list frame.

FSA data are especially valuable for multiple record linkage projects as well as various maintenance activities since most records contain full nine digit SSNs or EINs and provide efficient matching results. This affords an easier matching process and shorter review time for field office (FO) staff. However, FSA data are also prone to error. One concern with these data is that SSNs are not always correct. Family members or other individuals associated with the same operation often use the same SSNs when reporting their operations. This can present problems when setting cutoffs and during review.

The files being matched were processed through multiple passes (blocks/subsets) which did not include any special identifiers such as SSNs and/or EINs. These will be identified as the „general passes.‟ The general passes consist of blocks of variables using several of the parsed items from the information available for respondents. For example, one pass would consist of a data subset based on surname NYSIIS[2], first name and middle name. Thus, pairs of records with the same surname NYSIIS, first name and middle name will come together as matches. All other pairs will be considered non-matches. Another example would be a pass on operation name and city. Twenty-two general passes were created using these criteria. See Appendix B, Table B1.

Since the full nine digit SSN was available for both the FSA and list frame files, FSA records were flagged whenever any FSA SSN matched an SSN on the list frame (nine digit to nine digit). This allowed tracking FSA SSNs already on the list frame throughout the entire research project. Flagged SSNs in the non-matched group were considered missed SSNs since they should have been linked at some point in the match process. In other words, these records should have been possible or definite matches instead of non-matches. The same steps undertaken for SSNs were also applied to EINs.

Four record linkage models were created to assess the impact of not maintaining SSNs and/or EINs on the list frame: a standard model (nine digit), a four digit option 1 model, a four digit option 2 model, and a no SSN/EIN model. The standard model consisted of the 22 general passes plus 2 additional passes -- one blocking on nine digit SSN and other blocking on nine digit EIN. The four digit option 1 model consisted of the 22 general passes in addition to 6 other passes involving the use of the four digit SSNs and four digit EINs in combination with several other fields such as last name, first name, operation name, etc. The six passes selected for this model yielded similar number of matches when compared to their nine digit counterparts. The four digit option 2 model consisted of the 22 general passes plus 2 additional passes -- one blocking on four digit SSN & city and the other blocking on four digit EIN & city. The underlying basis for the former model came from an evaluation of the SSNs on the list frame. Counts of unique SSNs and four digit SSNs along with other fields (such as last name, county, city, etc.) were obtained. The results of the evaluation showed that for 96.2 percent of the list frame records, the combination of the four digit SSN, city and state was unique. This strong “uniqueness characteristic” made the combination of four digit SSN/EIN and city a viable replacement for SSN/EIN in the record linkage processing. Additionally, this model had a small number of missed SSNs and EINs relative to the other passes evaluated. The no SSN/EIN model consisted of the 22 general passes, all which excluded both SSNs and EINs. Table 1 provides a brief description of each model evaluated.

Table 1: Four Models Evaluation and their Descriptions

Model Name	**Description**
Standard Model (9-digit)	General Passes[3] w/ nine digit SSN & nine digit EIN
4-digit Option 1 Model	General Passes w/ four digit SSN w/ City, First/Last, Last only and four digit EIN w/ city, operation name, keyword on operation name
4-digit Option 2 Model	General Passes w/ four digit SSN w/ City and four digit EIN w/ city
No SSN/EIN Model	General Passes w/out SSN/EIN

In general, we accept the nine digit SSN as a unique identifier, and substantial weight is placed on its ability to identify like

[2] NYSIIS (New York State Identification and Intelligence System) is a phonetic algorithm for indexing names by their sound when pronounced in English. Its basic aim is to encode names with the same pronunciation to the same string so that matching can occur despite minor differences in spelling.

[3] General passes refer to the 22 passes which did not involve the use of SSNs and EINs.

records. One of the strengths of SSN is that we are fairly confident that we have a definite match if SSN and other pieces of information agree. Having SSN is a great tool to reduce both HQ staff's processing time and FO's review time. We are able to classify more records as definite matches having SSN than we would be able to otherwise.

IV. Results

We evaluated the effect the four digit SSN and four digit EIN had in processing time, review time, "miss rate" of SSN/EIN matches, and staff time. Table 2 presents the results for each model, for all five states combined. The table provides the number of definite and possible matched link groups, number of FSA records that were definite matches to records on the list frame, FSA records that did not match any records on the list frame (non-matches); and FSA records that were possible matches to list frame records

Table 2: Number of Definite, Possible and Non-Matched Link Groups[4] & Records by Model-Overall

Model	Definite Matched Link Groups	FSA Records Definite Matches to List Frame Records	Possible Matched Link Groups	FSA Records Possible Matches to List Frame Records	FSA Records Not Matching to List Frame Records/Link Groups[5]	Totals FSA Records
Standard (9-digit)	69,088	87,810	276,431	990,643	978,457	2,056,910
4-digit Option 1	66,790	85,238	272,661	990,130	981,542	2,056,910
4-digit Option 2	69,047	87,289	270,582	964,871	1,004,750	2,056,910
No SSN/EIN	66,326	84,879	274,149	963,373	1,008,658	2,056,910

The results show that the no SSN/EIN and the four digit option 2 models had a higher number of non-matching FSA records, indicating that duplication would be added to the list frame. The no SSN/EIN model would add over 30,000 records as duplicates to the list across all five states (1,008,658 - 978,457). The four digit option 2 did not do much better than the no SSN/EIN model since it would add over 26,000 duplicates to the list frame. Even though the standard model contained slightly larger review workloads, it had a significantly smaller number of non-matches. In other words, the model identified more FSA records requiring review, while identifying noticeably fewer records that should have been added to the list frame. The compounded effect of adding duplicate records would increase as the number of states processed increases.

The four digit option 1 model performed about the same as the standard model. As compared to the four digit option 2 and no SSN/EIN models, it identified more FSA records that required review and fewer that should be added to the list frame. However, this model and the four digit option 2 model consisted of a more complex record linkage structure. The four digit option 1 model required additional staff processing time to prepare and to set cutoffs since it had four additional passes. For this model, staff had to cycle through more data to achieve reasonable pairs from which to set cutoffs. Cutoffs vary depending on the composition of the records on the data files, and a known error rate is assumed when setting them. Staff would allow some unreasonable linked pairs to filter through to be able to identify a linked pair which obtained a lower weight than expected. It is assumed that these unreasonable records will be identified by FO staff during their review of clerical matches. This practice of allowing unreasonable linked pairs to filter through was more prominent in the case of the models involving four digit SSNs and EINs (options 1 and 2) than it was for the standard nine digit model. There were often a higher number of non-matched FSA records in the cases utilizing four digit or no SSN/EIN matching, which led to missing valid matches. The alternative matching procedures also often required more data review to get to a good pair of records. Record linkage cutoffs were typically higher for these models (compared to cutoffs that staff are accustomed to), since a large number of unreasonable pairs would filter in between good pairs, thus increasing the models" miss rate. Overall, the predicting power of the four digit option 1 model was very close to that of the standard model when minimizing or limiting the amount of duplicate records

[4] A link group may contain multiple FSA and list frame records. Hence, the number of records is greater than the number of link groups.

[5] Non-match link groups only contain one record. Link groups with more than one record are classified as either matches or possible matches.

added to the list frame.

An error rate measure based on nine digit SSN/EIN was obtained whenever any FSA SSN or EIN matched an SSN or EIN on the list frame (nine digit to nine digit). This allowed tracking FSA SSNs and EINs already on the list frame throughout the entire research project. Within the match, clerical, and non-match FSA records, all matching SSNs and EINs were identified. Flagged SSNs and EINs in the non-matched group were considered missed since they should have been linked at some point in the matching process. Table 3 presents the number and percentage of FSA SSNs and EINs not linked during the matching process for each of the models for all five states combined. The table also illustrates the overall total amount of duplication that would be added to the list frame as a result of the SSNs and EINs not linked or missed.

Overall, the results show that using four digit SSN/EIN option 1 for record linkage would miss 1.8 percent of the actual SSN matches and 7.2 percent of the actual EIN matches. Furthermore, using this model would add about 1.3 percent duplication to the list frame. However, having no SSN/EIN for record linkage would result in missing 5.5 percent of the actual SSN matches and 11.9 percent of the EIN matches. Additionally, totally excluding SSN/EIN from record linkage procedures would add over 3 percent duplication to the list. It is expected that significantly more records would be missed in larger states.

Table 3: SSNs and EINs not Linked During Matching Process – All States

Model	Total FSA SSNs not Linked	FSA SSNs not Linked (%)	Total FSA EINs not Linked	FSA EINS not Linked (%)	Duplication[6]	Duplication (%)
Standard (9-digit)[7]	0	0.0%	2,819	4.5%	2,819	0.3%
4-digit Option 1	8,449	1.8%	4,490	7.2%	12,939	1.3%
4-digit Option 2	23,267	5.0%	5,484	8.8%	28,751	2.9%
No SSN/EIN	25,576	5.5%	7,405	11.9%	32,981	3.3%

V. Impact on List Frame Maintenance Activities -- Eliminating SSNs/EINs or Using Four Digit SSNs and EINs

List frame maintenance activities are highly expensive both in human and equipment resources. The NASS record linkage system was designed such that lists could be accurately matched to one another while minimizing human resources. Furthermore, the system was designed such that additions and updates to the list frame could be made as efficiently and effectively as possible. A number of maintenance updates to the list frame are based heavily on the use of SSNs and EINs. Headquarter staff have made various changes in processing to incorporate the use of SSNs and EINs during record linkage processing such that specific records are excluded from initial review, thus minimizing staff time and maximizing productivity. Matching on SSNs and EINs has proven to be very beneficial in achieving these goals. Various list frame maintenance activities were evaluated to determine the overall impact of moving to the four digit SSN or not maintaining SSNs at all would have on the list frame over time.

FSA SSN/EIN Updates

Every year the SSN/EIN Update record linkage projects are conducted. The goal of these projects is to use FSA data to post SSN/EIN updates to the list frame. Some of the original FSA data files used to process the FSA SSN/EIN updates for MT and SC in 2006 were available for this research. For the purpose of this research, the files were matched to the list frame records, and SSNs were flagged and „removed." This provided an idea of how not performing this maintenance activity would impact the list frame. The initial run for MT had 53,816 matching SSNs. The number of matching SSNs was reduced to 37,059 after removing SSNs that had been updated through the FSA SSN updates. This represents a reduction of 31 percent fewer matching SSNs.

[6] Duplication = SSNs not linked + EINs not linked

[7] Not all matches involving nine digit EINs are valid matches since EINs are recycled.

In the initial run for SC, there were 38,049 matching SSNs, but only 28,568 for the reduced run -- a reduction of 25 percent. It is important to note that all the files used for these yearly SSN/EIN update projects were no longer available. This indicates that the percentage of SSNs „removed" would have been higher. Tables C1, C2, C3 and C4 in Appendix C show the results for MT and SC. The four models were processed again and counts were obtained with fewer matching SSNs. The analysis showed the impact on record linkage of having fewer records with SSN/EIN on the list frame. For example, in the standard model for MT the number of definite matches goes down from 5,595 in Table C1 to 4,434 in Table C2, Appendix C. Also in Appendix C, the number of non-matches goes up from 137,223 (Table C1) to 141,425 (Table C2). This demonstrates that having more SSNs on the list frame provides more definite matches and fewer non-matches. This is a clear indication of the long-term impact of fewer records on the list frame with SSN/EIN values.

The SSN/EIN update record linkage projects currently give the FOs the option to review records prior to Headquarters performing SSN updates. Moving to using four digit SSN, the record review portion for FOs would not be optional, in hopes of reducing the error rate added due to the shortening of SSN. Thus, shortening of the SSNs implies more review for states. Additionally, FO staff are instructed to review one project, before staff can process another one. Thus, this will create a larger backlog of projects for FO personnel.

FSA Missing Phone Match

For the FSA missing phone match project, record linkage uses SSNs and EINs as its strongest passes. Similar to the SSN/EIN update projects, definite matches are automatically updated without FO staff reviewing any of the records. Using four digit SSNs/EINs would require incorporation of a review for FO staff. The same issues arise regarding error rates and FO review backlogs.

FSA Minority Code Updates

FSA data are also used to update minority codes on the list records. List frame records with missing minority codes are updated by matching the list frame to FSA records with minority codes. The stronger passes involve using SSNs and EINs. Eliminating SSNs or using only the four digit SSNs will greatly limit the ability to properly classify records as matches. Even though we could still perform this project using names and addresses, the predicting power will not be as strong and thus will miss valid matches between FSA and the list frame. The ability to identify minority records on the list frame is important to NASS since it helps improve the accuracy of coverage measurements of minority farmers in our surveys and the Census of Agriculture.

NASS Across and Within State Unduplication

The agency also performs across-state and within state duplication removal efforts using primarily SSNs and EINs. This project would have to be performed on a very limited basis. SSN and EIN are the strongest identifiers for uniquely matching records for across state duplication. Without nine digit SSN/EIN, the power of any across state duplication record linkage would be very limited, primarily because the aim is to find the same person with different addresses in different states. Only using four digit SSN/EIN with name would greatly weaken the linkage process. This would especially hold true in this situation where the state code is eliminated from the matching process, thus the only field left to match is four digit SSN/EIN.

Social Security Administration (SSA) Death Match

NASS receives death master files from the Social Security Administration (SSA) on a yearly basis. The only available information on SSA files is name, date of birth, date of death, zip code and SSN. These data are used to identify the records of deceased operators. This project relies heavily on the use of SSNs as a matching field and thus will have to be eliminated from our current maintenance activities. Even though we could potentially match on name and zip code, this would represent a massive review effort for FO staff. Matching on these fields is not nearly as effective as matching on SSN. For the SSA death match, we are dealing with 69M records, a number significantly higher than the total number of records on the list frame. Thus, matching only on name and zip code would increase the margin of error significantly. With this number of records, there are more chances of identifying records with the same four digit SSNs than there are on the list frame. It becomes much more

difficult to discriminate between matched record pairs. Especially since the data fields provided by SSA are much more limited.

The Census of Agriculture mail list (CML) is comprised of all records eligible to receive a census questionnaire. The SSA death match project identifies records of deceased operators on the list frame. Based on this project, over 32,000 records received a special code indicating that their corresponding operators could be deceased and should be given special attention during census data collection.

Currently, FO staff interactively searches information on Social Security Death Index (SSDI) sites using SSN whenever they receive indication that an operator is deceased. Without SSNs available, these searches would not be possible. Additionally, the search engines are not designed to search on four digit SSNs and a larger number of returns would be expected since four digit SSNs alone are not a strong, unique matching variable.

Federal Tax Information (FTI) Processing

During the Federal Tax Information (FTI) processing some records are deemed matches and removed from further processing through the use of the SSNs and EINs. This processing step reduces the number of records required for further processing and at the same time reduces the review workloads of FO staff. Appendix D provides results for TX, showing the additional number of records which would need to be processed as a result of eliminating the SSN passes from the FTI project. The appendix provides the number of records that would be processed using nine digit SSN/EIN and four digit SSN/EIN with placename. The results show that eliminating the SSN and EIN passes from processing will quadruple the number of records requiring review by NASS staff. This also signifies that it will take longer for the project to be processed and FO staff will require more time to conduct their review. However, using either nine digit or four digit SSNs/EINs would reduced the workload by ¼. For FTI, IRS will not provide only four digits to avoid incorrect identification of respondents. Thus, NASS will be receiving nine digits from IRS regardless of what is internally used to match to the list frame. It is important to note that safeguard procedures are in place which prohibit IRS information from being used to update the list frame. IRS data are used for processing purposes only.

General Database Search

We should also expect an increase in staff time since currently staff are proficient with database searches based on nine digit SSNs, and it will be much harder to discriminate between different operations with similar data without SSNs. For example, if there were two records with differing suffixes (JR/SR) and similar addresses, having SSNs allow reviewers to make a more definitive decision than they could without SSNs. Without SSNs, reviewer might call these operators the same; however, with the aid of SSNs it will be clear that the two are different.

VI. Conclusions

Overall, the results show that using four digit SSNs/EINs for record linkage would result in missing 1 to 2 percent of the actual SSN matches and 4 to 8 percent of the actual EIN matches, with a higher miss rate in large states. However, having no SSNs/EINs for record linkage would result in missing 4 to 6 percent of the actual SSN matches and 9 to 13 percent of the EIN matches, with a higher miss rate in large states. NASS processes several record linkage projects per state per year and if we add this same level of duplication per project and missed these percentages of SSNs and EINs the compounded effect of using four digit would be significantly higher as the years go on. The impact of not using any SSNs/EINs at all would decrease the quality of the list frame even more.

The results show that completely eliminating SSNs and EINs greatly impacts NASS‟ ability to maintain and run smooth processes. If SSNs and EINs are completely eliminated from the list frame and all record linkage processing, HQ staff will have to process significantly more data to set record linkage parameters. Additionally, FO staff will have a significantly increased record linkage review workload.

VII. Recommendation

NASS needs to maintain nine digit SSNs/EINs on the list frame for use in record linkage and maintenance activities in order to avoid significant erosion of the quality of the list frame over time, including:

- Adding duplication from new list sources;
- Missing drop records from the SSA Death Match;
- Missing minority code updates; and
- Missing removal of duplicate records from within-State and across-State un-duplication projects.

VIII. References

Bates, N. (2005). Development and Testing of Informed Consent Questions to Link Survey Data With Administrative Records. ASA Proceedings of the 2005 Joint Statistical Meetings.

Bosecker, R. (2007). "Personally Identifiable Information." Internal memorandum, National Agricultural Statistics Service. November 8, 2007.

Johnson III, C. (2007). Safeguarding Against and Responding to the Breach of Personally Identifiable Information. Memorandum for the Heads of Executive Departments and Agencies. Executive Office of the President- Office of Management and Budget. May 22, 2007. M-07-16

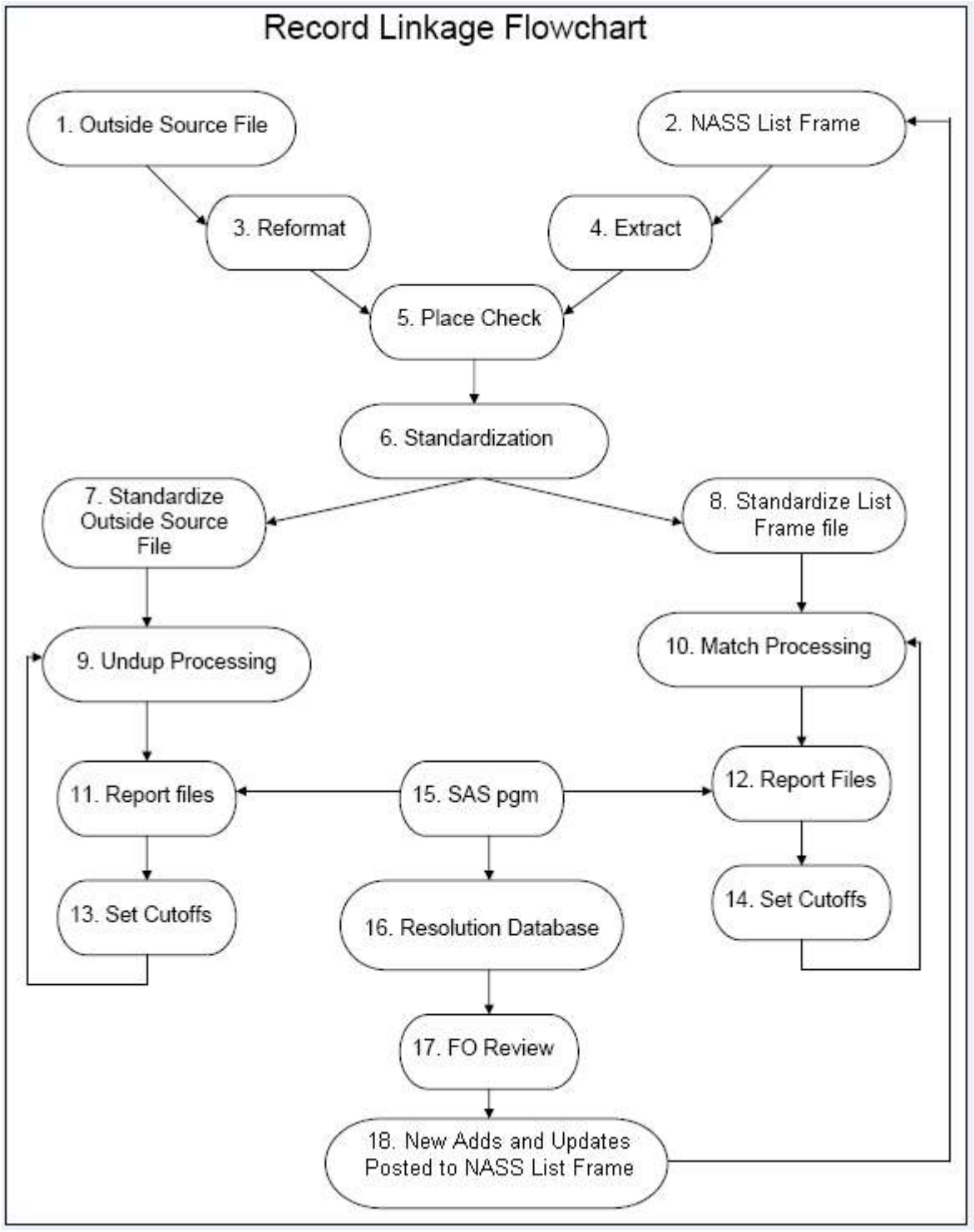
Meyer, P.S., Dahlhamer, J.M., and Pleis, J.R. (2006). Developing New Methods and Questions for Improving Response and Measurement on Sensitive Questions on the National Health Interview Survey. ASA Proceedings of the 2006 Joint Statistical Meetings.

Nealon, J. (2007). "NASS Plant to Eliminate Use of SSN." Internal memorandum, National Agricultural Statistics Service.

Smith, H., Daniel, K., Abreu, D., Hoge, S., and Iwig, W. (2006). Record Linkage and Automatic Maintenance Activities. ASA Proceedings of the 2006 Joint Statistical Meetings.

USDA Privacy Impact Assessment (PIA). UNIX MA/ USC Systems. Prepared by DSD Laboratories. March 15, 2007.

Appendix A


Record Linkage Flowchart
1. Outside Source File
2. NASS List Frame
3. Reformat
4. Extract
5. Place Check
6. Standardization
7. Standardize Outside Source File
8. Standardize List Frame file
9. Undup Processing
10. Match Processing
11. Report files
15. SAS pgm
12. Report Files
13. Set Cutoffs
16. Resolution Database
14. Set Cutoffs
17. FO Review
18. New Adds and Updates Posted to NASS List Frame

Appendix B

Table B1: Description of Blocking Variables for 22 General Passes

Pass #	Description of Blocking Variables
1	Surname NYSIIS, first name & middle name
2	City, street NYSIIS, house number & surname NYSIIS
3	10-digit telephone number
4	Surname NYSIIS, first name, middle name & city
5	Surname NYSIIS, first name & city
6	Surname NYSIIS & city
7	City, house number & street NYSIIS
8	City & PO Box
9	NYSIIS of operation name surname & city
10	Specific keyword on operation name (oopr_name)
11	Operation name
12	First word on the operation name & city
13	Operation name & city
14	Zip Code
15	NYSIIS of surname, first name, NYSIIS of operation surname, operation first name
16	Surname NYSIIS
17	Specific key word on the operation name & city
18	Operation name
19	7-digit portion of telephone number & 7-digit portion of operation‟s telephone number
20	Surname NYSIIS, house number & street NYSIIS
21	Surname NYSIIS & PO box
22	Surname NYSIIS & house number

Appendix C

Table C1: Totals by Model for MT *Before* Removing SSNs from List Frame File

Model	Definite Matched Link Groups	FSA Records Definite Matches to List Frame Records	Possible Matched Link Groups	FSA Records Possible Matches to List Frame Records	FSA Records Not Matching to List Frame Records/Link Groups[5]	Totals FSA Records
Standard (9-digit)	5,595	7,606	25,864	114,529	137,223	259,358
4-digit Option 1	5,596	7,606	25,992	113,684	138,068	259,358
4-digit Option 2	5,272	7,230	26,167	112,429	139,699	259,358
No SSN/EIN	5,259	7,198	26,344	111,801	140,359	259,358

Table C2: Totals by Model for MT *After* Removing SSNs from List Frame File

Model	Definite Matched Link Groups	FSA Records Definite Matches to List Frame Records	Possible Matched Link Groups	FSA Records Possible Matches to List Frame Records	FSA Records Not Matching to List Frame Records/Link Groups[5]	Totals FSA Records
Standard (9-digit)	4,434	5,236	26,965	112,697	141,425	259,358
4-digit Option 1	4,581	5,398	26,398	112,365	141,595	259,358
4-digit Option 2	4,088	4,797	26,740	111,221	143,340	259,358
No SSN/EIN	4,039	4,730	27,406	109,339	145,259	259,358

Table C3: Totals by Model for SC *Before* Removing SSNs from List Frame File

Model	Definite Matched Link Groups	FSA Records Definite Matches to List Frame Records	Possible Matched Link Groups	FSA Records Possible Matches to List Frame Records	FSA Records Not Matching to List Frame Records/Link Groups[5]	Totals FSA Records
Standard (9-digit)	12,040	13,851	25,481	71,856	121,387	207,094
4-digit Option 1	11,090	12,759	26,190	73,020	121,315	207,094
4-digit Option 2	11,338	13,048	25,758	71,319	122,727	207,094
No SSN/EIN	11,085	12,763	26,035	71,264	123,067	207,094

Table C4: Totals by Model for SC *After* Removing SSNs from List Frame File

Model	Definite Matched Link Groups	FSA Records Definite Matches to List Frame Records	Possible Matched Link Groups	FSA Records Possible Matches to List Frame Records	FSA Records Not Matching to List Frame Records/Link Groups[5]	Totals FSA Records
Standard (9-digit)	11,152	12,571	26,658	75,119	119,404	207,094
4-digit Option 1	10,830	12,302	26,411	76,013	118,779	207,094
4-digit Option 2	10,764	12,106	26,353	75,117	119,871	207,094
No SSN/EIN	10,604	11,936	26,847	74,271	120,887	207,094

Appendix D

Number of Records Processed by Pass During 2009 FTI Review to Set Cutoffs with 9-digit SSN/EIN, 4-digit SSN/EIN with Placename and without SSN/EIN – TX

PASS DESCRIPTION	# pairs w/out SSN/EIN	# pairs with 9-digit SSN/EIN	# pairs with 4-digit SSN/EIN & placename	Difference 9-digit vs No Digit	Difference 4-digit vs No Digit	Difference 9-digit vs 4-Digit
Phone (phn)	2,040	563	585	1,477	1,455	-22
Operation (Op) phn and phn	279	118	125	161	154	-7
Phone and phone other	340	156	163	184	177	-7
Surname, street, house #, and city	30,081	7,114	7,042	22,967	23,039	72
Opsur/psurname, street, house #, & city	2,639	1,486	1,481	1153	1158	5
Pfirst, street, house #, and city	25,701	3,692	3,624	22,009	22,077	68
Ofirst/psurname, street, house #, and city	295	185	185	110	110	0
Pfirst/psurname, street, house #, and city	141	76	76	65	65	0
Psurname/opsur, street, house #, and city	8,351	3,214	3,200	5,137	5,151	14
Op surname, street, house #, and city	1,485	793	792	692	693	1
Ppartfirst/pfirst, street, house #, and city	2,395	2,008	2,006	387	389	2
Op surname/pfirst, street, house #, and city	220	142	142	78	78	0
surname, box #, and city	12,874	3,501	3,465	9,373	9,409	36
Op/psurname, box #, and city	1,761	973	970	788	791	3
psurname/opsurname, box #, and city	4,914	1,829	1,809	3,085	3,105	20
nysiis of surname, first, mid, sfx, and city	14,848	1,785	1,739	13,063	13,109	46
wholename and city	8,385	2,437	2,416	5,948	5,969	21
opername and city	4,106	1,057	1,041	3,049	3,065	16
wholename/opername and city	673	233	233	440	440	0
Opername/ wholename and city	47	21	21	26	26	0
last, house #, city	32,099	7,579	7,506	24,520	24,593	73
Street, house # and city	32,824	8,655	8,574	24,169	24,250	81
house # and city	34,442	9,178	9,096	25,264	25,346	82
Surname, Street, house #	30,294	7,177	7,146	23,117	23,148	31
Pfirst, house #, city	27,404	3991	3,922	23,413	23,482	69
box # and city	15,874	5,092	5,042	10,782	10,832	50
surname, first, middle, sfx	20,658	2,380	2,374	18,278	18,284	6
nysiis of surname, first, sfx, and city	39,578	5,979	5,872	33,599	33,706	107
House #	224	70	68	154	156	2
Pfirst/ofirst, psur/osur city	4,355	1,498	1,487	2,857	2,868	11
oopername and city	480	240	237	240	243	3
nysiis of surname and city	50,846	13,840	13,722	37,006	37,124	118
op surname /surname and city	5,387	3,170	3,162	2,217	2,225	8
Popkey/oopkey, city	4,927	1,652	1,639	3,275	3,288	13
o/p first, o/p mid, o/p sur, o/p suf, cty	1,117	667	667	450	450	0
Totals	**422,084**	**102,551**	**101,629**	**319,533**	**320,455**	**922**