United States Department of **Agriculture**

National Agricultural Statistics Service

Research and Development,
Washington, D.C., 20250

RDD Research Report
Number RDD-23-01

January 2023

On Farm Grain Stocks Sample Methodology Review

Yang Cheng
Leslie Smith
Jeff Bailey
Linda J. Young

On Farm Grain Stocks Sample Methodology Review

Yang Cheng, Leslie Smith, Jeff Bailey, and Linda J. Young

Abstract

The Quarterly Agricultural Survey (QAS) is a quarterly survey conducted in June (base), September, December, and March. The survey provides detailed estimates of crop acreage, yield and production, and quantities of grain and oilseeds stored on farms. This report reviews QAS sample design and documents the design in detail. In addition, the report evaluates the efficiency of the NASS area frame, assesses the rotation panel design, defines multivariate probability proportional to size (MPPS) sampling, and derives inclusion probabilities in a dual- or multiple-frames setting. After carefully reviewing the documentation and digesting the information, some potential short-term and medium-term to long-term improvements have been identified.

Key Words: Quarterly Agricultural Survey, Multivariate Probability Proportional to Size Sampling, Rotation Panel Design, Area Frame, Poisson Sampling

1. Introduction

The information on the amount of grain stocks on the farm is derived from the Crops Acreage, Production and Stock (Crops APS) survey, which is also called the Quarterly Agricultural Survey (QAS). This review's narrow scope on estimating on-farm grain stocks was determined at the planning phase. This section of the report focuses on the sample methodology of QAS. We began by collecting sampling documentation, identifying the staff who are experts in QAS sampling, and meeting with them for discussion. After an initial investigation, the following areas were targeted for evaluation: the QAS sample design, the efficiency of the NASS area frame, and the rationale behind the rotation panel design. This document is divided into 3 sections: current sample methodology, some potential opportunities for improvement, and recommendations.

2. Current QAS sample design

The QAS is a quarterly survey conducted in June (base), September, December, and March. The target population, U, for the QAS is all agricultural establishments with cropland or storage capacity. Let F be the NASS list frame. If samples are only drawn from F for the QAS, the agricultural establishments that are not on the NASS list frame will be missed. To completely cover the target population, the sample frame for U has 2 components: (1) a list of "large" units from F, U_L, where U_L is the set of all agricultural establishments on F with 50 acres or more of cropland or 1,000 bushels of grain storage capacity for most states and (2) the complement of U_L, which is denoted by U_L^c. Hence, $U = U_L \cup U_L^c$. Note that U_L is a truncated list frame.

The NASS list frame includes all known agricultural establishments. Only list frame records with positive planted acres or storage capacity of the desired commodities are included in the list frame population for the QAS sample. Thus, U_L is a subset of F, that is to say, $U_L \subset F$. According to the Grain Stocks Methodology (March 2012) draft report, U_L includes approximately 825,000 farms or ranches and covers approximately 92 percent of the acreage for major row crops in the United States. U_L^c consists of "small" units on F and all unknown agricultural establishments. Since U_L^c is not completely known, the NASS area frame, F_A, is used to draw samples covering U_L^c. The sample methodologies differ for U_L and U_L^c.

2.1 Detail sample design on U_L

Current QAS samples from U_L are comprised of 4 components, which are drawn from 4 partially overlapping frames. Each of the four frames targets a specific population: general, row crops, small grains, or specific individual crops. For all quarters of QAS, the three main subpopulations, general, row crops, and small grains, are used in all states. The specific individual crops frame is decided by state and may not be sampled for all quarters. Before 1997, a stratified sample, which did not have panels, was used to collect all items.

2.1.1 Samples from general frame:

The general frame, which is a list of all crop farms known to NASS, ensures every farm has a chance of selection each quarter. Let U_G be the general frame and note that $U_G = U_L$. U_G is partitioned by state, and a Poisson sample for each state is drawn from the corresponding partition of U_G. At NASS, a Poisson sample with the desired first-order inclusion probability of frame unit derived from multiple frames is called a multivariate probability proportional to size (MPPS) sample. The following steps are the key to obtaining the desired inclusion probability for each farm within a state.

1) Define the auxiliary variables for calculating the measure of size (MOS):
   There are 3 auxiliary variables for MOS: the total cropland acres, storage capacity, and calculated land in field crops (control types 300, 305, and 303, respectively). Control type 303 was created in the middle of the 1980s, probably when the original crops and stocks survey began. It is the sum of crop acres for only the crops in the survey. This eliminates farms, such as fruit farms, from being included in the survey. Some states may use both 300 and 303 and some states may only use one.
2) Calculate MOS:
   MOS (M_k) is calculated by the formula: $M_k = \frac{\sum_{i \in U_G} x_{k,i}^p}{n_k}$, where $k = 1$ means the total cropland acres, $k = 2$ means storage capacity, $k = 3$ means calculated land in field crops, $x_{k,i}$ is a known characteristic of the general frame, and n_k is a sample parameter, which is normally determined by survey precision. However, no target coefficient of variation (CV) is shown in **Table 1a** of the Policy and Standards Memorandum No. PSM-ASMS-12 for the general items of cropland or capacity. For this MOS, the distance

measure is defined on the L^p space. $p = \frac{3}{4}$ was determined empirically in the QAS design before 1997. Normally, $p = 1$ indicates a standard probability proportional to size (PPS) sample design.

3) Calculate PPS inclusion probability:
For $i \in U_G$ and *k = 1, 2, 3*, the first-order PPS inclusion probability is

$$\pi_{k,i} = \begin{cases} \frac{x_{k,i}^p}{M_k} & if\ x_{k,i}^p < M_k \\ 1 & Otherwise \end{cases}$$

4) Calculate desired target inclusion probability for Poisson sampling:
For $i \in U_G$, the target inclusion probability for Poisson sampling is $\pi_i = \max_k \pi_{k,i}$.

5) Adjust the inclusion probability in step 4):
For $i \in U_G$, the final Bernoulli trial probability is $\pi_{G,i} = \min\{\pi_i, L\}$, where L is the predetermined limit. For QAS, $L = \frac{1}{3}$. If $L = 1$, the inclusion probability in step 4) is not adjusted, that is, $\pi_{G,i} = \pi_i$.

After step 5), each farm has a target probability, $\pi_{G,i}$. Now, each farm is also assigned a random number (ε_i) from the uniform (0, 1) distribution. If each farm is associated with the same random number permanently over all components or surveys, this random number is called the permanent random number (PRN). Poisson sampling is used to select the sample and determine the selection based on the comparison of $\pi_{G,i}$ and ε_i. If $\varepsilon_i < \pi_{G,i}$, then this farm is selected for the sample. Otherwise, the farm is not included in the sample. Thus, a sample from U_G is generated and referred to as S_G, that is, $S_G = \{i | \varepsilon_i < \pi_{G,i}\ for\ i \in U_G\}$. The sample size n_G, and $n_G = \sum_{i \in U_G} I_{[\varepsilon_i < \pi_{G,i}]}$, where $I_{[\varepsilon_i < \pi_{G,i}]}$ is an indicator function. The expectation of the sample size is $E(n_G) = \sum_{i \in U_G} \pi_{G,i}$. It is easily shown that $\mathrm{E}(n_G) \geq \max_k n_k$. Note that n_G does not directly depend on n_k.

2.1.2 Samples from row crops, small grains, and specific individual crops:

For different states, QAS chooses different commodities as auxiliary variables for calculating the MOS for row crops, small grain crops, and specific individual crops. The random number (ε_i) assigned to each farm on the general frame is also used for the decision of selection for all other components. Here we will use 1997 Minnesota as example to demonstrate the sample selection procedure.

Row crops frame:

The row crops frame, say U_R, contains all farms that produce row crops. U_R is subset of U_G. Two auxiliary variables for MOS are chosen: sunflowers acres ($x_{1,i}$) and dry edible beans acres ($x_{2,i}$). After completing steps 1) - 5) provided for the General frame, a Poisson sample for row crops is drawn from U_R. The inclusion probability for row crops is $\pi_{R,i}$, and a sample set drawn from U_R is S_R.

Small grains frame:
The small grains frame, say U_S, contains all farms that produce small grains. Note that $U_S \subset U_G$. Three auxiliary variables for MOS are chosen: winter wheat acres ($x_{1,i}$), durum wheat acres ($x_{2,i}$), and barley acres ($x_{3,i}$). After completing steps 1) - 5) in the General frame, a Poisson sample for small grains is drawn from U_S. The inclusion probability for small grains is $\pi_{s,i}$, and a sample set drawn from U_S is S_S.

Specific individual crops frame:
The specific individual crops frame (CS) in Minnesota is potatoes. The CS frame, denoted by U_P, contains all farms that produce potatoes. Thus, $U_P \subset U_G$. Apply PPS on U_P, and the inclusion probability is $\pi_i = \begin{cases} \frac{x_i^p}{M} & if\ x_i^p < M \\ 1 & Otherwise \end{cases}$, where the auxiliary variable for MOS is potato acres (x_i) and MOS $M = \frac{\sum_{i=1}^{N} x_i^p}{n_p}$, where n_p is the expected number of potato farms in the Minnesota sample. After completing steps 1) - 5) in the General frame, a Poisson sample for specific individual crops is drawn from U_P. The inclusion probability for specific individual crops is $\pi_{P,i}$, and the sample drawn from U_P is S_P.

Note that all auxiliary variables from row crops, small grains, and specific individual crops are commodities. Sample parameters n_k or n_p in the row crops, small grains, and specific individual crops can be derived based on the precision requirement from **Table 1a** of the Policy and Standards Memorandum.

After samples are drawn from the subpopulations of general, row crops, small grains, and specific individual crops, the overall QAS samples for U_L are $S_L = S_G \cup S_R \cup S_S \cup S_P$. Since the four components overlap, a farm in U_L can be drawn 0, 1, 2, 3, or 4 times. The first-order inclusion probability π_i for $i \in S_L$ should be correctly calculated. The reciprocal of π_i is the survey design weight.

2.1.3 Special rotation panel design:

The QAS survey is a quarterly survey, but the samples are drawn annually from each subpopulation or component before June. Special quarterly rotation panels are designed for each subpopulation. Each sample set $S_k, k = G, R, S, P$, described in the previous section is systematically assigned with sort on probabilities of selection into three panels, that is,

$$S_k = \bigcup_{m=1}^{3} S_k^{(m)}, \qquad when\ k = G, R, S, P$$

$S_k^{(m)}$ means the *m*th panel of samples from frame *k*. Furthermore, panel designs are different among the four subpopulations. The sample panels are displayed in **Table 1**. The main purpose for a rotation panel design is to measure the change in agricultural characteristics between quarters with stable estimates and less variability.

Table 1: Rotation Panel Design for QAS

Frame	General			Row Crops			Small Grain			Potatoes		
Panels	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
June	x			x			x	x		x	x	x
September	x	x	x	x			x	x	x			
December	x	x	x	x	x	x	x	x	x	x		
March		x	x		x	x		x	x			

Rotation panels cover all quarters and overlap quarter by quarter except sample panels in the specific individual crops. However, there are no overlapping panels between March and June. Finally, quarterly QAS samples are derived by combining the four subpopulations and rotation panel design as follows:

QAS June samples from U_L is $S_{Jun} = S_G^{(1)} \cup S_R^{(1)} \cup \left(\cup_{m=1}^{2} S_S^{(m)}\right) \cup \left(\cup_{m=1}^{3} S_P^{(m)}\right)$.

QAS September samples from U_L is $S_{Sep} = \left(\cup_{m=1}^{3} S_G^{(m)}\right) \cup S_R^{(1)} \cup \left(\cup_{m=1}^{3} S_S^{(m)}\right)$.

QAS December samples from U_L is $S_{Dec} = \left(\cup_{m=1}^{3} S_G^{(m)}\right) \cup \left(\cup_{m=1}^{3} S_R^{(m)}\right) \cup \left(\cup_{m=1}^{3} S_S^{(m)}\right) \cup S_P^{(1)}$.

And QAS March samples from U_L is $S_{Mar} = \cup_{k=G,R,S}\left(\cup_{m=2}^{3} S_k^{(m)}\right)$.

From the sample procedures in previous sections, the first-order inclusion probability for farm i from frame k is derived as $\pi_{k,i}$, where $k = G, R, S, P$, and $i \in S_k$. As a part of the sample design, the survey weight or base weight for each farm $i \in S_L$, where S_L is one of S_{Jun}, S_{Sep}, S_{Dec}, and S_{Mar}, needs to be derived and is the reciprocal of π_i. The inclusion probability, π_i, is derived from overlapping samples, which is discussed in **Appendix 2.**

The 2019 Sampling Summary for NASS's National Probability Surveys (Green Book) provided some ideas about overall samples sizes (**Table 2**). NOL samples will be discussed in the next section. The QAS sample design is quite complicated. To better understand panels and components, a sample distribution by components and panels should be developed. This information can provide insights into the differences among 4 quarters. Since the March 2019 is part of the 2018 samples, it makes sense to include March 2020 sample information in the 2019 Green Book.

Table 2: 2019 QAS Sample Size by Quarter

	March	June	Sept.	Dec.
List	75,755	66,456	56,901	71,914
NOL	4,571	10,298	4,458	4,563

2.2 Detail sample design on U_L^c

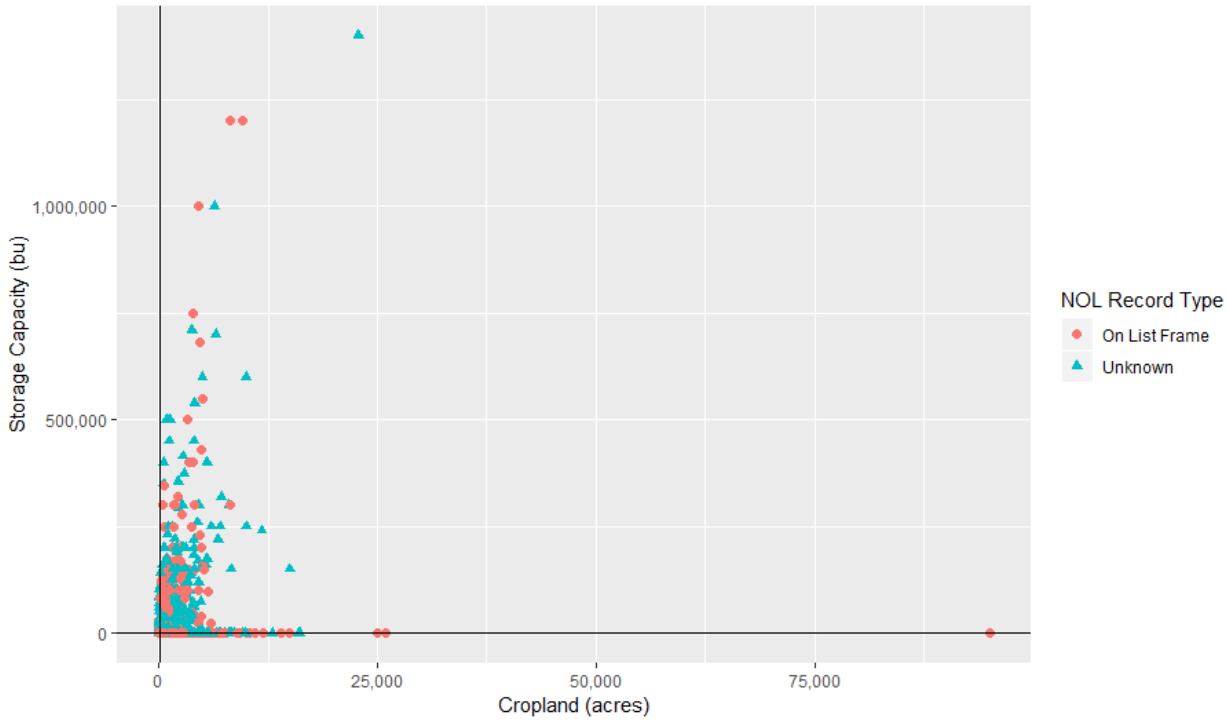
U_L^c consists of "small" crops farms or ranches on the NASS list frame and all agricultural establishments with crops not on the NASS list frame. Most agricultural establishments not on the NASS list frame are also "small" units. Since a list of agricultural establishments to cover U_L^c does not exist, an area sample frame approach is considered. Since area frame covers the whole USA and locations of farms in U_L^c is not well defined, it is difficult to identify a list and map to U_L^c.

NASS conducts the June Area Survey (JAS) annually in June. It utilizes a NASS area sampling frame, F_A, which provides complete coverage of all agriculture activity occurring on the land. Thus, $F_A \supset U$. JAS applies a stratified multi-stage area probability sample within each state. The sampling unit is a segment, which is about a square mile (640 acres). Prior to JAS data collection, field enumerators pre-screen each segment, dividing it into tracts of land. Each tract represents a unique land operating arrangement (see Davis 2009 for more details). Let S_A be JAS samples that are drawn from F_A, and $S_A \subset F_A$. A sample of 9,045 segments was drawn in 2019, approximately 70,795 agricultural and non-agricultural land use tracts were identified within the sampled segments. From that identification, over 31,005 detailed personal interviews were conducted. The JAS has a rotation panel design. A selected segment stays in the sample for five years. Each year, JAS has approximately 20% new segments, which appear in the sample for the first time. JAS also has approximately 20% of segments that appear a second, a third, a fourth, and a fifth time. The following year, approximately 20% of segments, which were interviewed a fifth time, are rotated out. The other 80% of segments remain in the sample and approximately 20% new segments are rotated in.

If S_A is used to sample for QAS, this is a dual frame sampling design. Dual frame sampling designs are a subset of multiple-frame designs in which units within the population of interest are selected via independent probability samples taken from each of two frames. These two frames make up the population (U) of interest, and they typically overlap. In a dual frame survey, two sampling frames together cover the population of interest. Independent probability samples S_A and S_{Jun} are taken, and information from the two samples are combined to estimate the commodities of interest. However, a dual-frame analysis approach is not used for the QAS. Instead, the QAS adopts a coverage improvement idea. Assume that the set of $S_A \cap U_L$ is a sample that can cover U_L (this is not a perfect mapping) and the complement sample of S_A covers subpopulation U_L^c. Let $S^o = S_A - S_A \cap U_L$. Actually, $S^o = S_A \cap U_L^c$ and S^o is used here to project U_L^c. The corresponding inclusion probability, π_i^o, is adopted from the JAS design for a farm $i \in S^o$. The sample size of S^o was 10,298 in 2019 JAS. For QAS, S^o is commonly called the nonoverlap portion of the area frame (NOL), which are the farms found operating in these segments that are not included in the truncated list frame population, U_L, in June.

Numerous units with 50 acres or more of cropland or 1,000 bushels of grain storage capacity that are in the NOL are on the NASS list frame but were not classified as being in the Crops APS population (see **Figure 1**). These records were either inactive or did not have frame data to be included in the population. Furthermore, cropland and storage capacity are not correlated (**Figure 1**). For example, one farm with more than 75,000 acres has nearly zero bushels of on-farm storage capacity.

Figure 1: Scatter Plot by Cropland vs Capacity for 2019 June NOL



To better understand the association between storage capacity and cropland, if any, the scales of cropland and storage capacity used in **Figure 1** are truncated at 1,000 acres and 5,000 bushels, respectively, and two critical lines (cropland = 50 acres and capacity = 1,000 bushels) are added (see **Figure 2**). Clearly, numerous farms with 50 acres or more of cropland or 1,000 bushels of grain storage capacity on the NOL are also on the NASS list frame. In 2019, 1,742 farms with more than 1,000 acres were on both the NOL and the NASS list frame (see **Table 3**). Furthermore, 312 farms with at least 5,000 bushels of on-farm grain storage were on both the NOL and the NASS list frame (see **Table 4**).

In 2018**,** 1,749 of the NOL tracts with more than 50 acres of cropland as reported on the JAS were found on the NASS list frame (**Table 3**). This needs to be investigated. Further, 312 tracts in the NOL with more than 1,000 bushels of capacity captured on the JAS were found on the NASS list frame (see **Table 4**). These 312 units need to be reviewed.

Figure 2: Truncated Scatter Plot by Cropland vs Capacity for 2019 June NOL

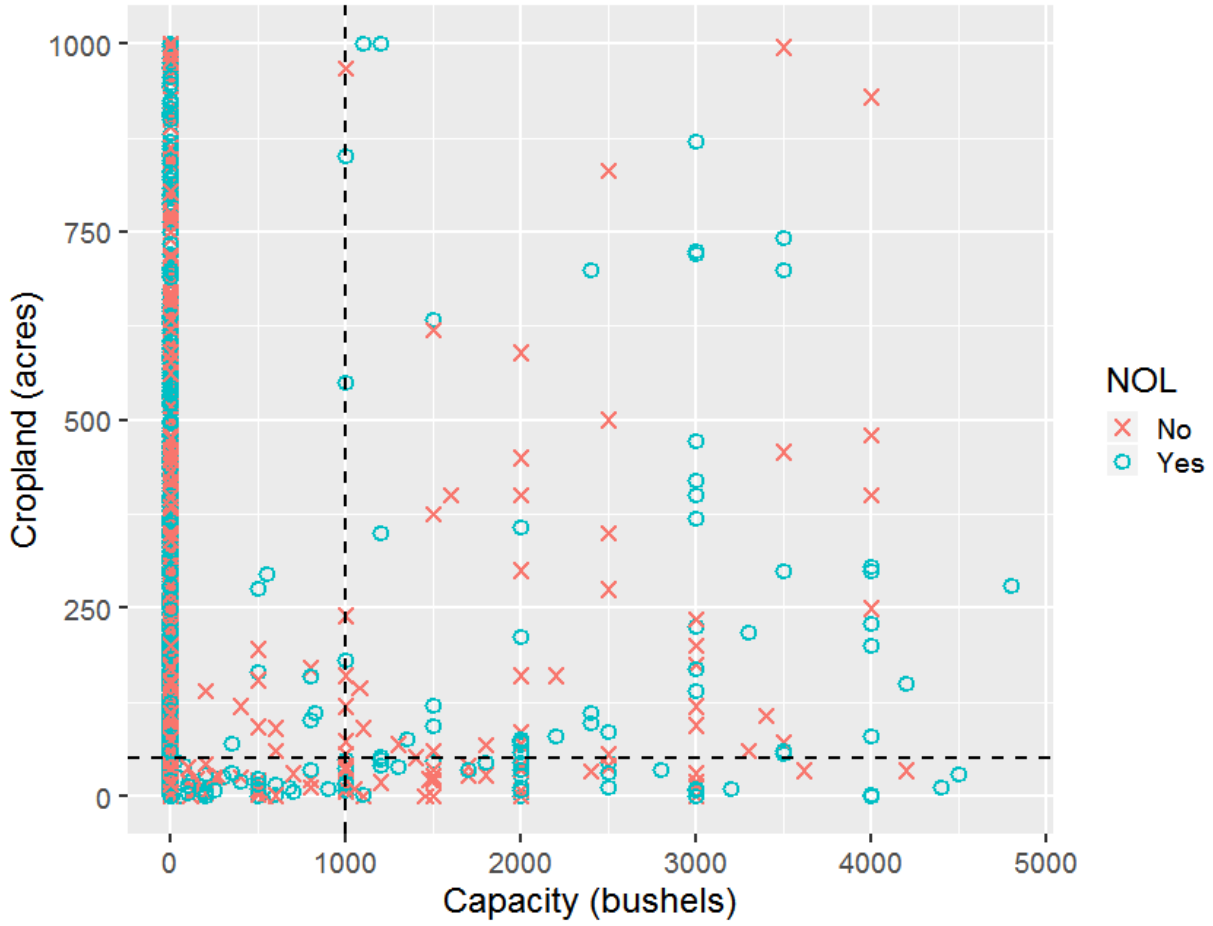


Table 3: Cropland by Frame Type for 2019 NOL Records

	Unknown	On List	Subtotal
Cropland > 50	2,222	1,749	3,971
Cropland <= 50	3,680	2,647	6,327
Subtotal	5,902	4,396	10,298

Table 4: Capacity by Frame Type for 2019 NOL Records

	Unknown	On List	Subtotal
Capacity > 1000	519	312	831
Capacity <= 1000	5,087	3,946	9,033
Capacity Unknown	296	138	434
Subtotal	5,902	4,396	10,298

Finally, the overall samples for the population (U) in June are generated as $S = S_{Jun} \cup S^o$. Since $S_{Jun} \cap S^o = \emptyset$, the first-order inclusion probability is

$$\pi_i^F = \begin{cases} \pi_i & when\ i \in S_{Jun} \\ \pi_i^o & when\ i \in S^o \end{cases}$$

Note that S^o does not have a rotation panel design, which is different from the design of $S_k,\ k = G, R, S, P$. Samples from U_L^c for other quarters, such as September, December, and March, need to be drawn.

In June, information on all the JAS records is available. So, S^o is used as the NOL sample. For the other quarters (September, December, and March), S^o is used as the sampling frame for U_L^c. Within each state, a stratified sample S^{oo} is drawn from S^o (see **Table 5**) for the stratum descriptions). For each stratum, systematic sampling (SYS) is applied with sample interval (SI). Note that, we did not locate any documents to determine SI. SI is a design weight for NOL sample in September, December, and March. S^{oo} is a subset of S^o; that is, $S^{oo} \subset S^o$. In 2019, the sample size of S^{oo} was 4,458 and 4,563 for September and December, respectively (see **Table 2**). The slight difference in sample sizes between September and December is because CT, MA, NV, NH, RI, and VT are not in the September QAS.

Table 5: Stratified Sample Design for other Quarter

Stratum	Description
1	Large cropland or large capacity
2	Small cropland, small or missing capacity, positive or unknown stocks, and positive or unknown small grain intentions
3	Small cropland, small or missing capacity, positive or unknown stocks, and zero cattle/hog/sheep/goat/chicken and no cattle/hog/sheep/goat/chicken intentions
4	Small cropland, small or missing capacity, positive or unknown stocks, and OL cattle/hog/sheep/goats/chickens
5	Small cropland, small or missing capacity, positive or unknown stocks, and positive cattle/hog/sheep/goats/chickens or cattle/hog/sheep/goat/chickens intentions
6	Zero stocks/crops items
7	Non-ag tracts with potential ag

Since S^o is a set of records from U_L^c with inclusion probability π_i^o, and S^{oo} is a set of records drawn from S^o with systematic sampling, the first-order inclusion probability for a unit $i \in S^{oo}$ drawn from U_L^c is

$$\pi_i^{oo} = \frac{\pi_i^o}{SI}$$

The sample design from June to March is a longitudinal design, and due to the panel design of the JAS, there is approximately an 80% overlap in the March to June samples. This is different from the sample design for U_L, which does not have overlapping panels from March to June.

Finally, quarterly QAS samples are derived by combining samples from the two frames as follows:

QAS June sample is $S_{Jun} \cup S^o$.
QAS September sample is $S_{Sep} \cup S^{oo}$.
QAS December sample is $S_{Dec} \cup S^{oo}$.
And QAS March sample is $S_{Mar} \cup S^{oo}$.

Since $S_{Jun} \cap S^o = \emptyset$ and $S_k \cap S^{oo} = \emptyset$ when *k = Jun, Sep, Dec*, the first-order inclusion probability $\pi_{k,i}$ for farm *i* in the sample does not need to be adjusted.

$$\pi_i^F = \begin{cases} \pi_i & when\ i \in S_k, \quad k = Sep, Dec, Mar \\ \pi_i^{oo} & when\ i \in S^{oo} \end{cases}$$

3. Some potential areas for improvement

After carefully reviewing the documentation and gathering information from QAS sampling experts, some potential areas for improvement have been identified. The following could improve the quality of the QAS methodology.

1. Technical documentation that addresses the details of the sample design, estimation, and variance estimator for QAS should be developed.
2. The undercoverage of the NASS list frame is a contributing factor to the complexity of the QAS sample design. Efforts should continue to improve the coverage of the NASS list frame.
3. The sample design differs between the two frames. How important is the design for different panels in the different components? Rationale should be provided for the design inconsistencies between quarters.
4. The sample size is a key element of survey design. It is related to the budget and affects the precision of survey estimates. Documentation on sample size to address sample parameters, samples in components by panels, sample size vs. design sample size, and sample adjustment does not exist and should be developed.
5. The NASS Green Book provides survey CVs and compares them to their targets from the Policy and Standards Memo. This is a quote from QAS sampling statistician: "**If CVs are well below CV targets, for example, we would reduce samples appropriately. If CVs are not meeting CV targets, we would increase samples if possible."** An action plan for the procedures to be followed to address CVs that fail to meet their target should be developed.
6. MPPS is applied to many surveys at NASS, but there is no clear definition of MPPS. This is a time to clearly describe the MPPS design, including the purpose of design.

7. The MPPS sample parameter p was set to $p = \frac{3}{4}$ based on an empirical study conducted before 1997. This sample parameter should be evaluated periodically.
8. Update the documentation to use standard statistical terminology. For example, NASS Acronyms define NOL = No Overlap or Not On List. However, in many QAS documents, NOL implies U_L^c. As another illustration, stratum has a specific meaning in survey methodology, but it represents an imputation cell in QAS and a nonresponse adjustment cell in the Off-Farm Grain Stocks survey.
9. Every rotation panel design has its own motivation, and the rationale should be documented. However, this documentation could not be found for the QAS. It should be noted that, in general, random assignment is not good practice for panel designs.
10. Original SAS programs used for sample selection were coded before 1997 without flow charts. Coding techniques have been improved and numerous new SAS Procs have been developed in past 20 years for efficiency and accuracy. These programs should be updated.
11. During the review, some changes were noticed. For example, samples among components changed from non-overlapping samples and sequence in selection to overlapping samples, and the number of auxiliary variables increased from two to three for sample selection from the general frame. No documentation reflecting when and why these changes were made or whether testing was conducted prior to implementation could be found. Changes should be clearly documented.
12. Update the documentation to show how the inclusion probabilities are calculated. For example, the inclusion probabilities of the sample from the truncated list frame U_L need to be modified.
13. To implement the MPPS method for the truncated list frame U_L, several sample parameters need to be derived and adjusted. For example, n_k or n_p involves different crops (i.e., corn, soybeans, rice, etc.) for different components, and predetermined limit L.
14. Each sampled farm in the JAS represents all other farms in the stratum, which includes "large" farms in the truncated list frame U_L, "small" farms, and unknown to NASS farms in the U_L^c. So, for farm $i \in S^o$, farm i represents farms in both U_L and U_L^c. Thus, S^o may not cover U_L^c well. An evaluation of the efficiencies of S^o needs to be conducted, and a weight adjustment may need to be added.
15. In the JAS, grain stocks information is mostly collected from S^o only. For farms in $S_A \cap U_L$, if it is known before data collection that the tract is the overlap with QAS, the stocks section of the June Area questionnaire is skipped. This reduces respondent burden and could cause bias the JAS estimates. This process can be modified such that the stocks questions are asked of S_A, allowing a more efficient dual frame estimate approach to be adopted. For other quarters, the sample design also needs to be modified to implement a dual frame approach to estimates.
16. Note: No Quality Assurance (QA) procedure is in place for sample review. After QAS samples are drawn, staff from regional offices review the sample. This may not be enough for QA. Sample statisticians need to develop a check list for key statistical parameters and precision targets.
17. All sample design variables are known but may not be accurate when the sample is conducted. As common practice, nonresponse, coverage, and calibration adjustments are applied to correct slight differences. For the QAS two-frames approach, it may be difficult to separate auxiliary variables between the two frames.

18. A truncated list frame, U_L, is a population with 50 acres or more of cropland or 1,000 bushels of grain storage capacity for most states. But some "large" units from the NASS list frame appear in U_L^c, the NOL frame (see **Table 3** and **Table 4**).

4. Recommendations

Based on this review, some recommendations are provided for improving the existing QAS sampling procedures. The items that need to be addressed immediately or that can be fixed easily are identified for short-term improvement. Other problems that may be difficult or require more resources are designated for medium-term and long-term improvement. The short-term recommendations should be addressed within 2 years, and medium-term and long-term recommendations should be addressed within two to five years.

4.1 Short-term improvement

1. Produce a QAS methodology technical paper that addresses all technical details including:
    1. A guideline to determine and adjust sample parameters.
    2. Description of derived sample size and relationships with survey precision requirements.
    3. Rationale for panel design and creation of rotation panels.
    4. Description of target population and frame population.
    5. Sample methodologies for each subpopulation and stratum.
    6. Inclusion probability for each sampled unit.
2. Develop a sample adjustment plan based on the CV results from Green Book to address CVs that differ substantially from their targets.
3. Write a technical paper fully describing MPPS.
4. Evaluate sample parameter ($p = \frac{3}{4}$) and adjust, if needed.
5. Create new, more efficient SAS code for QAS sample selection with flow charts.
6. Evaluate using PPS method instead of MPPS method for sample selection in the specific individual crops component.
7. Update the documentation to clarify how the inclusion probability formula for the samples in S_{Jun}, S_{Sep}, S_{Dec}, and S_{Mar} Are determined.
8. Evaluate applying JAS weights to the NOL component of the QAS. Consider adjusting QAS weights based on the evaluation of weights from U_L vs. weights from $S_A \cap U_L$.
9. Develop a Quality Assurance (QA) procedure for sample review.
10. Identify proper auxiliary variables for combining samples from two frames and develop a comprehensive weighting adjustment procedure, which includes nonresponse, coverage, and calibration adjustments.
11. Since the March 2020 list sample is drawn from the 2019 frozen frame and the NOL sample is drawn from the 2019 June Area NOL, consider including the March 2020 sample information in the 2019 Green Book in addition to the March 2019 data.
12. Investigate why so many "large" units fall in the NOL.

13. Since QAS uses a complicated sample design, a design effect should be calculated and monitored.

4.2 Medium-term and long-term improvement

1. Develop a new sample design. A sample design based on the list frame only with a cut-off sampling technique should be considered.
2. Evaluate the NASS list frame coverage and continue to explore methods to improve that coverage. Evaluate whether QAS a list frame sample design without an NOL component should be implemented.
3. Develop dual frame estimates based on the data from JAS and QAS data drawn from U_L in June. Evaluate a modified sample design for NOL samples for other quarters that is similar to the approach in June.

References

- Agricultural Statistics Board (2021), "Grain Stocks Methodology and Quality Measures". National Agricultural Statistics Service.
- Amrhein, J. F. (1999). "Multivariate Probability Proportional to Size Sampling: Its Description, Development and Implementation in NASS Surveys". NASS SSB Report Number SSB-99-01.
- Amrhein, J. F., Hicks, S. D, and Kott, P. S. (1996). "Methods to Control Selection When Sampling from Multiple List Frames". ASA Proceedings of the Section on Survey Research Methods.
- Amrhein, J. F. and Bailey, J. T. (1998). "Sampling Villages for Multipurpose Surveys". Proceedings of the Joint IASS/IAOS Conference: Statistics for Economic and Social Development.
- Bailey, J. T. and Kott, P. S. (1997). "An application of multiple list frame sampling for multi-purpose surveys". ASA Proceedings of the Section on Survey Research Methods.
- Davis, C., (2009). "Area Frame Design for Agriculture Surveys". RDD Research Report, U.S. Department of Agriculture, National Agricultural Statistics Service. Washington, D.C.
- Field Offices and Headquarters Units (2018). "Policy and Standards Memoranda: No. PSM-ASMS-12". National Agricultural Statistics Service (http://nassportal/NASSdocs/Documents/PSM-ASMS-12.pdf).
- Hicks, S. D, Amrhein, J. F., and Kott, P. S. (1996). "Methods to Meet Target Sample Sizes Under A Multivariate PPS Sampling Strategy". ASA Proceedings of the Section on Survey Research Methods.
- Kott, P. S., Amrhein, J. F. and Hicks, S. D. (1998). "Sampling and estimation from multiple list frames". Survey Methodology, 24(1).
- Kott, P. S. and Bailey, J. T. (2000). "The Theory and Practice of Maximal Brewer Selection with Poisson PRN Sampling". ASA Proceedings of the Section on Survey Research Methods.

- National Agricultural Statistics Service internal draft report (2012). "Grain Stocks Methodology".
- National Agricultural Statistics Service (2018). Sample Design and Statistical Precision. Policy and Standards Memorandum - No. PSM-ASMS-12.
- National Agricultural Statistics Service (2022). Agricultural Surveys Interviewer's Manual (http://nassportal.nassad.nass.usda.gov/NASSdocs/Documents/2022_Ag_Surveys_Interviewers_Manual.pdf)
- Sampling and Frame Development Section (2020). "2019 Sampling Summary for NASS's National Probability Surveys (Green Book)", NASS Methodology Division SFDS Report Number SEIMB 19-01.

Appendix 1: Multivariate Probability Proportional to Size Sampling

For a survey with multiple ($K: K > 1$) study variables, some units in the population may not contain the information for all *K* study variables. Let *U* be the population. *S* is a sample drawn from *U*. Define $y_i = (y_{1,i}, \cdots, y_{K,i}),\ i \in U$ as a study survey variable where $y_{k,i}$ is the information on the *k*th ($k = 1, \cdots, K$) study variable for the *i*th unit and *N* is the size of *U*. Furthermore, let U_k be the population that contains all units for which the *k*th study variable exists if it has a value. Thus, U_k is a subpopulation and $U_k \subset U$. The size of subpopulation U_k is denoted by N_k, where $N_k = \sum_{i \in U} I_{[y_{k,i}]}$, $I_{[y_{k,i}]}$ is an indicator function for the *k*th study variable in the *i*th unit, that is, $I_{[y_{k,i}]} = \begin{cases} 1 & if\ y_{k,i}\ exists\ with\ a\ value \\ 0 & otherwise \end{cases}$. For some *k*, $N_k < N$. A special case is that $y_{k,i}$ exists for all units in *U*, in this case, $N_k = N$.

The same scenario can be defined for *S*. Let S_k be a set of samples with for which the *k*th study variable exists if it has a value. S_k is a subset of *S*, and $S_k = \left\{ i \in S: I_{[y_{k,i}]} = 1 \right\}$. Also, let *n* be the sample size of *S*. The size of S_k is denoted by n_k, where $n_k = \sum_{i \in S} I_{[y_{k,i}]}$, and $n_k \leq n$. For some *k*, $n_k < n$. This may cause a problem that n_k is too small to meet survey precision requirements even if *n* is large enough.

What is a Poisson sampling? In survey methodology, Poisson (PO) sampling is a sampling process where each unit of the population is subjected to an independent Bernoulli trial that determines whether the unit becomes part of the sample. First, put units in frame in some order and generate a random number (ε_i) from the uniform distribution from zero to one for every population unit ($\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$). Then, unit *i* is selected for the sample if $\varepsilon_i \leq \pi_i$ where $0 < \pi_i < 1$ is the desired inclusion probability for each unit. π_i is predetermined and may have differ with unit depending on the sample design. The probability of being included in a sample during the drawing of a single sample is denoted as the first-order inclusion probability of that unit.

What is a Multivariate Probability Proportional to Size (MPPS) sampling? MPPS sampling is a Poisson sampling with the desired first-order inclusion probability derived from multiple frames

based on the Probability Proportional to Size (PPS) measurement. So, to implement MPPS sampling, the key is to identify the desired inclusion probability for each unit through multiple frames on PPS measurement. The detail for deriving π_i follows:

Step 1: Construct *K* frames U_k from population *U*.
Step 2: Determine the target sample size n_k^t to meet survey precision requirement. Normally, n_k^t is determined as though U_k is the only frame and is a function of survey precision, population size, and auxiliary data.
Step 3: Identify an auxiliary variable $x_{k,i}$, which is used to calculate measure the size.
Step 4: Calculate the inclusion probability based on PPS setting, $\pi_{k,i} = n_k^t \frac{x_{k,i}}{\sum_{i \in U_k} x_{k,i}}$, where $x_{k,i}$ is an auxiliary variable. Furthermore, the PPS design can be extended by using MOS $x_{k,i}^p$ in place of $x_{k,i}$. Then $\pi_{k,i} = n_k^t \frac{x_{k,i}^p}{\sum_{i \in U_k} x_{k,i}^p}$, where $0 < p \leq 1$.
Step 5: Take the maximum of PPS inclusion probability $(\pi_{k,i})$ across all *K* frames, $\pi_i = \max_{1 \leq k \leq K} \pi_{k,i}$. If there exists a unit *k* such that $\pi_{k,i} \geq 1$, then $\pi_{k,i} = 1$.
Step 6: Set π_i as a predetermined probability for Poisson sample selection based on the sample design.

After π_i is calculated for $i \in U$, a random number ε_i should be generated before implementing the Poisson scheme. Sample selection is determined on the comparison of π_i and ε_i. Sample indicator function $I_i = \begin{cases} 1 & if\ \varepsilon_i \leq \pi_i \\ 0 & otherwise \end{cases}$, *i* = 1, 2, …, N, is an independent Bernoulli(π_i). The probabilities of selection, which may be unequal, are $P(I_i = 1) = \pi_i = 1 - P(I_i = 0)$. In MPPS sampling, selections are independent, and the sample design is $P(s) = \prod_{i \in S} \pi_i \prod_{i \in U/S} (1 - \pi_i)$. The total number of possible samples is 2^N. The size of MPPS sampling is $n_s = \sum_{i \in U} I_{[\varepsilon_i \leq \pi_i]}$. n_s is random. The expectation of the MPPS sample size, E(n_s), is $\sum_{i \in U} \pi_i$, which provides an indication of what the final sample size will be. And the variance of the MPPS sample size, $V(n_s) = \sum_{i \in U} \pi_i (1 - \pi_i)$. It is easy to show that $\mathrm{E}(n_s) \geq \max_k n_k^t$. In MPPS sampling, $S_k = S \cap U_k$. The size of S_k is $n_k = \sum_{i \in U} I_{[\varepsilon_i \leq \pi_i] \cap [y_{k,i}]}$. Thus, n_k is not related to n_k^t.

Since $\pi_{ii} = \pi_i$, $\pi_{ij} = \pi_i \pi_j\ for\ i \neq j$, $\Delta_{ij} = 0$, and $\Delta_{ii} = \pi_i(1 - \pi_i)$, the π-estimator of the population total under MPPS sampling is

$$\hat{t}_\pi = \sum_{i \in S} y_i / \pi_i$$

Its variance is

$$V_{MPPS}(\hat{t}_\pi) = \sum_{i \in U} \pi_i (1 - \pi_i) \check{y}_i^2 = \sum_{i \in U} \frac{1 - \pi_i}{\pi_i} y_i^2$$

An unbiased variance estimator is

$$\hat{V}_{MPPS}(\hat{t}_\pi) = \sum_{i \in S} \frac{1 - \pi_i}{\pi_i^2} y_i^2$$

Appendix 2: Inclusion Probabilities in a Dual- or Multiple-Frames Setting

In design-based sampling, the indices and frames are deterministic mathematical objects. We begin by assuming that U_1 and U_2 are overlapping sets of indices, where $U = U_1 \cup U_2$ is the study population.

Let the index *i* be the indicated units in the population and *k* = 1, 2 as the index for the two subpopulations U_1 and U_2. A sampling design that simultaneously draws samples from subpopulation U_k is denoted by S_k. The inclusion probabilities on S_k and $S_1 \cap S_2$ are defined as follows:

$$\pi_i^{(k)} \equiv P(i \in S_k) \text{ and } \pi_i^{(1,2)} \equiv P(i \in S_1 \cap S_2).$$

In QAS design, the samples S_k for *k* = 1, 2 are drawn dependently because they share the same random number ε_i. For the Poisson sampling setting, $S_k = \left\{i|\varepsilon_i < \pi_i^{(k)} \; for \; i \in U_k\right\}$ and $S_1 \cap S_2 = \left\{i|\varepsilon_i < \pi_i^{(1)} \; and \; \varepsilon_i < \pi_i^{(2)} for \; i \in U_1 \cap U_2\right\} = \left\{i|\varepsilon_i < \min\left\{\pi_i^{(1)}, \pi_i^{(2)}\right\} for \; i \in U_1 \cap U_2\right\}$. So, the separate notation $\pi_i^{(1,2)}$ is not needed since for all $i \in U_1 \cap U_2$,

$$\pi_i^{(1,2)} \equiv P(i \in S_1 \cap S_2) = \min\left\{\pi_i^{(1)}, \pi_i^{(2)}\right\}.$$

Note that the inclusion probability notation can be summarized efficiently, for all possible $i \in U_1 \cup U_2 = U$, in the form

$$P(i \in S_1) = I_{[i \in U_1]} * \pi_i^{(1)}, \; P(i \in S_2) = I_{[i \in U_2]} * \pi_i^{(2)}.$$

Therefore, by the well-known *Inclusion-Exclusion Formula* with $S = S_1 \cup S_2$,

$$P(i \in S) = P(i \in S_1) + P(i \in S_2) - P(i \in S_1 \cap S_2) = I_{[i \in U_1]} * \pi_i^{(1)} + I_{[i \in U_2]} * \pi_i^{(2)} - I_{[i \in U_1 \cap U_2]} * \min\left\{\pi_i^{(1)}, \pi_i^{(2)}\right\}.$$

In the special case *k* = 1, 2, when unit *i* is selected for sample from both population U_1 and U_2,

$$P(i \in S) = P(i \in S_1) + P(i \in S_2) - P(i \in S_1 \cap S_2) = \pi_i^{(1)} + \pi_i^{(2)} - \min\left\{\pi_i^{(1)}, \pi_i^{(2)}\right\} = \max\left\{\pi_i^{(1)}, \pi_i^{(2)}\right\}$$

Under a similar assumption of Poisson sampling setting with sharing the same random number ε_i and samples S_k with respective inclusion probabilities $P(i \in S_k) = \pi_i^{(k)}$ for $i \in U_k$ and $k = 1, \cdots, 4$, it is easy to generalize the formula to obtain:

$$\begin{aligned}\pi_i = P(i \in S) &= P\left(i \in \bigcup_{k=1}^{4} S_k\right) \\ &= \sum_{k=1}^{4} I_{[i \in U_k]} * \pi_i^{(k)} - \sum_{1 \le k_1 < k_2 \le 4} I_{[i \in U_{k_1} \cap U_{k_2}]} * \min\left\{\pi_i^{(k_1)}, \pi_i^{(k_2)}\right\} \\ &+ \sum_{i1 \le k_1 < k_2 < k_3 \le 4} I_{[i \in U_{k_1} \cap U_{k_2} \cap U_{k_3}]} * \min\left\{\pi_i^{(k_1)}, \pi_i^{(k_2)}, \pi_i^{(k_3)}\right\}\end{aligned}$$

$$- I_{[i \in \cap_{k=1}^{4} U_k]} * \min \{\pi_i^{(1)}, \pi_i^{(2)}, \pi_i^{(3)}, \pi_i^{(4)}\}$$

This version of the four-way inclusion formula is correct for all indices $i \in U = \bigcup_{k=1}^{4} U_k$.