United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Methodology Division

SEDMB Staff Report
Number SEDMB 23-01

May 2012
Revised January 2023

THE YIELD FORECASTING PROGRAM OF NASS

The Summary, Estimation, and Disclosure Methodology Branch

THE YIELD FORECASTING AND ESTIMATING PROGRAM OF NASS, by the Summary, Estimation, and Disclosure Methodology Branch, Methodology Division, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C., January 2023. NASS Staff Report No. SEDMB 23-01.

ABSTRACT

The National Agricultural Statistics Service (NASS) is responsible for estimating production of most crops grown in the United States. Additionally, early season forecasts are prepared for the major crops. NASS conducts several surveys to obtain the basic data needed to fulfill this obligation. These surveys are a mix of grower interviews and objective field visits employing sophisticated survey sample designs and statistical methodology.

Large surveys designed to measure acreages are used to define prescreened subsampling populations for the yield surveys. These surveys and the subsampling techniques are described, and the data collection procedures are also outlined. Summary formulas are given, and regression techniques employed in the forecasting process are discussed in detail.

Each survey produces indications of prospective yield which commodity specialists must interpret to arrive at the official forecast or estimate of NASS and the USDA. This paper discusses in detail the process of producing these indications by the Summary, Estimation, and Disclosure Methodology Branch and outlines the review process used by the commodity specialists in the Crops Branch. A brief discussion of acreage estimates is included to the extent that they impact sampling and the calculation of production.

KEY WORDS

Yield, Forecast, Estimate, Regression, Outlier.

TABLE OF CONTENTS

THE YIELD FORECASTING PROGRAM OF NASS

CHAPTER 1 OVERVIEW

Introduction

Each month, the U.S. Department of Agriculture publishes crop supply and demand estimates for the Nation and the world. These estimates are used as benchmarks in world commodity markets because of their comprehensive nature, objectivity, and timeliness. The statistics that USDA releases affect decisions made by businesses and governments by defining the fundamental conditions in commodity markets. When using USDA statistics, it is helpful to understand the forecasting and estimating procedures used and the nature and limitations of crop estimates.

Several agencies within USDA are responsible for preparing world crop statistics. The National Agricultural Statistics Service (NASS) and the World Agricultural Outlook Board (WAOB) have crop statistics among their primary focus. NASS forecasts and estimates U.S. crop production based on data collected from farm operations and field observations. The WAOB is responsible for monthly forecasts of supply and demand for major crops, both for the United States and the world, and follows a balance sheet approach to account for supplies and utilization. The major components of the supply and demand balance sheet are beginning stocks, production, domestic use, trade, and end-of-season carry-out stocks. Forecasts and estimates of U.S. crop production are independently prepared by NASS, while U.S. and foreign supply and demand forecasts are developed jointly by several USDA agencies with WAOB coordinating.

This paper is dedicated to the crop production estimating program of NASS. A brief discussion of acreage estimation is followed by a detailed presentation of yield forecasting and estimating. This paper examines the NASS process from sample design to data collection to summarization and data interpretation.

Definitions

Several variables, key to forecasting and estimating crop production, are defined below:

Planted acreage: Acreage planted for all purposes includes: (a) acreage planted that has been or will be harvested; (b) acreage planted and replanted to the same crop (only the first planting is included); (c) acreage planted and later plowed down, grazed, or abandoned; (d) volunteer acreage, only if the acres will be harvested; and (e) acreage planted on land enrolled in Government diversion programs.

Harvested acreage: Acreage harvested includes: (a) all acres already harvested or intended for harvest and (b) the same crop acres (such as hay) harvested two or more times for the same utilization from the same planting are included only once.

Biological Yield: The gross or total amount of a crop produced by plants expressed as a rate per unit; for example, bushels per acre.

Net Harvested Yield: The portion of total crop production removed from the field, expressed as a quantity per unit of area, and derived by deducting harvesting and other losses from the biological yield.

Production: The total quantity of an agricultural commodity recovered or removed from the field. In other words, net harvested production computed as harvested acres times net harvested yield.

Preparing NASS Production Forecasts

Crop production forecasts and estimates have two components -- acres to be harvested and yield per acre. A full program of forecasts and estimates includes determining acres planted at the beginning of the growing season, estimates of acres to be harvested for grain, forecasts of yield during the season, and final acres and yield after harvest. For example, corn and soybean planted acreage estimates are made using data obtained from a survey of farmers conducted during the first two weeks in June. Expected corn and soybean yields are obtained monthly using the Ag Yield survey in August and two different types of yield surveys September through November. Acres to be harvested for grain are measured in June and monitored through the season. Final acreage and yield are measured in December.

Two types of crop forecast surveys are conducted, a grower-reported survey and objective measurement surveys. The survey of growers, the Agricultural Yield Survey (AY), covers all major field crops included in the NASS estimating program. Growers in the sample are asked, monthly, to provide their assessment of yield prospects for the crops they grow. The objective measurement surveys, known as Objective Yield (OY) Surveys, cover wheat, corn, soybeans, and cotton. The OY surveys consist of a sample of fields in which counts and measurements are made of plants in random plots laid out in each field.

Data collected from the yield surveys reflect seasonal growing conditions and weather events as of the first of the month. An historical accumulation of monthly OY data collected under a variety of growing conditions is an invaluable forecasting asset. The implicit relationship between OY data and seasonal growing conditions is also explicitly evaluated using temperature and precipitation relative to “normal”. Departures from normal are evaluated not only for the current year but for the range of historic years under consideration. An assumption of "normal" conditions is always held for the remainder of the growing season. Data collected from AY surveys also reflects seasonal growing conditions and weather events up to the first of the month.

AY datasets have also been accumulated over time and form an integral part of yield forecasting. In the context of AY surveys, the influence growing conditions and weather events have had upon this year’s yield is given by the respondent’s collective perception, judgment, and experience gathered over some given period of time.

NASS does not attempt to predict future weather conditions or events. Long-range weather forecasts are not used in any forecast models and growing conditions and weather events after the first day of the month are evaluated in the following months forecast. A significant change in conditions or a weather event between the survey period and the report release date such as a killing freeze, serious heat wave, beneficial rains, etc., will not alter the forecasted values based on conditions existing on the first day of the month. NASS policy requires forecasts to be based on conditions as of the first of the month, the period which corresponds to data collected in the OY and AY surveys.

CHAPTER 2 SAMPLE DESIGNS

Acreage and final production estimates for the major field crops are based on data collected from a set of quarterly surveys designed to measure these items. The two yield forecasting surveys documented in this manual use a subsample of operations and fields identified during these quarterly surveys. Grower-reported yield surveys cover most major field crops included in the NASS estimating program and are referred to as the Agricultural Yield Survey (AY). Objective measurement surveys are conducted for corn, cotton, soybeans, and winter wheat and are referred to as Objective Yield Surveys (OY).

Sampling Frames

The sample designs for these surveys utilize two different sampling frames. The area frame is defined as the entire land mass of the United States and ensures complete coverage of the U.S. farm population. The list frame is a roster of known farmers and ranchers and includes a profile of each operation indicating the size of the operation and what commodities have historically been produced. The main strengths of the area frame are its completeness and stability. The weaknesses are its inefficiency for crops grown in small regions and its cost to build and collect data. The list frame can be sampled more efficiently (commodity specific, if necessary) and data can be collected using less expensive methods (e.g., mail, telephone, or web). The list frame does not provide complete coverage of all farms and is not stable since farming arrangements are constantly changing.

The area frame is stratified by land use for efficient sampling. All land in each State is classified into land use categories by intensity of cultivation using a variety of map products, satellite imagery, and computer software packages. These land use classifications range from intensely cultivated areas to marginally cultivated grazing areas to urban areas. The land in each use category is further divided into segments ranging in size from about 1 square mile in cultivated areas to 0.1 square mile in urban areas. Different sampling rates are applied to different strata with intensely cultivated land segments selected with a greater frequency than those in less intensely cultivated areas.

All AY samples are selected from respondents to the March or June Agricultural Survey. Objective Yield survey samples for corn, cotton, and soybeans are selected from the June Area Survey (JAS) tracts having the commodity of interest. Objective Yield survey samples for winter wheat are selected from March Agricultural Survey respondents with winter wheat planted for harvest as grain.

Objective Yield samples are selected as soon as possible following the final summary of the March Agricultural Survey or JAS. For geographic representation of the samples, the records are first sorted by state, district, county, segment, tract, and crop. The sample select programs use probabilities proportional to size to select a systematic random sample of acres from the reported acres (multiplied by the inverse of the sampling fraction) of the parent survey. The selected acres

are used to determine sample fields. Two counting areas (plots) are then randomly selected in each field.

The following example displays the area design for Nebraska. This frame was built with seven land-use strata covering all 77,157 square miles. Each stratum is mutually exclusive and independent. The stratum labeled commercial is made up of urban areas and the non-agricultural contains mostly protected forest land. The expansion factors are inverses of the sampling fractions. Note the allocation of the samples favors the more intensely cultivated areas with 200 of 207 segments falling in the first three strata. Strata with little or no agriculture are lightly sampled.

Nebraska – 2019 JAS samples

Stratum	Square Miles	Segment Size	Segments in Frame	Sample Size	Exp Factor	Stratum Definition
11	22,878	1	22,867	200	114	>75% Cultivated
12	9,480	1	9,452	70	135	51-75% Cultivated
20	12,084	1	12,037	63	191	15-50% Cultivated
31	465	0.25	1,850	2	925	Agri-Urban
32	66	0.1	657	2	329	Commercial
40	5,147	2	2,567	6	428	0-15% Cultivated
42	26,775	4	6,690	12	558	<15% Cultivated
50	**262**	pps	16	2	8	Non-Agricultural
Total	**77,157**		**56,136**	**357**		

The Agricultural Survey list frame sample is selected using a multivariate probability proportional to size (MPPS) sampling scheme. Each list frame record is assigned a measure of size based on historic data for multiple specified commodities. The MPPS design makes it very easy to target samples sizes for the commodities of interest. The desired number of samples for each commodity can be controlled with a minimum overall sample size. The MPPS design makes it easier to change samples to meet the needs of the crops program changes. The MPPS design is a more efficient design because operations will have a more optimal probability of selection based upon their individual commodities and size.

The strata from a previously used stratified design are not being used for sampling, however they are still used for nonresponse adjustments and item level imputation. Stratification for the Agricultural survey sample is based on the frame data for total cropland, on-farm grain storage capacity, and some rare or specialty crops (for each State). An example of the nonresponse stratification for Illinois Agricultural Survey is shown below.

ILLINOIS
Agricultural Survey Strata

Strata	Boundaries
97	Capacity 500K+
95	Cropland 7,000+
79	Cropland 2,500-7,499
78	Capacity 50K-499,999
73	Sorghum 1+
72	Cropland 600-2,499
66	Capacity 10K-49,999
65	Cropland 100-599
62	Capacity 4K-9,999

Multiple frame statistical methodology has been developed that captures the efficiency of the list frame and uses the area frame to measure incompleteness. This methodology was developed jointly by NASS and Iowa State University with provisions to account for each farm or land area once and only once. The survey process requires a check of all operations found in the area sample against the entire list frame. Area operations not found on the list comprise the sample from which incompleteness is measured.

Acreage and Final Production Surveys

The basic data for all NASS acreage estimates and final production estimates are collected on the quarterly Agricultural Surveys. These surveys also cover the quarterly grain stocks data requirements. NASS views the annual cycle of these surveys beginning in June with September, December, and March completing the cycle. Each survey employs multiple frame methodology.

All surveys have list samples of about 70,000 operations. These samples are replicated, and

replicates are rotated from quarter to quarter with about 60 percent of the sample retained from one quarter to the next. This scheme allows for response burden management while keeping the ability to measure quarter to quarter change using matched reports.

The June Area survey (JAS) features complete enumeration of an area sample of about 9,000 segments. The June area sample forms a stand-alone survey from which a set of unbiased indications are generated. They can also be married to the respective list samples to provide another set of unbiased multiple frame indications. The September, December and March area frame samples include only tracts that were not eligible to be included in the list frame crops population. These tracts are referred to as non-overlap (NOL) tracts. Approximately 5,000 area non-overlap tracts are sampled in these follow-on quarters.

Survey content differs each quarter to meet the varying requirements of the estimating program. The June and March surveys also define subsampling populations for the yield forecasting surveys. The following table outlines key data items collected on each survey, the yield surveys subsampled from them, and from which frame the subsample is drawn.

Survey	Items Measured	Surveys Subsampled
June	Planted acres of spring planted crops. Acres harvested and to be harvested for spring crops and winter wheat.	Ag Yield (Aug. - Nov.) (list) Corn Objective Yield (area) Soybean Objective Yield (area) Cotton Objective Yield (area)
September	Final harvested acres and yield of small grains.	None
December	Seeded acres of winter wheat (new crop). Final harvested acres and yield for spring crops.	None
March	Planting intentions for spring planted crops. Winter Wheat acres for harvest as grain.	Ag Yield (May - August) (list) Winter Wheat OY (multiple frame)

Estimates of planted acres, made at the beginning of the season, include some acres left to be planted at the time of the survey. Generally, these fields do get planted and planted acreage estimates are not changed during the crop season. Occasionally, the planting season runs extremely late causing abnormally large intentions in the data or some weather event alters grower plans after the data are collected. When this happens, NASS may re-visit these farms during late July to determine what was actually planted. If necessary, planted and harvested

acreage estimates are revised and published in the *August Crop Production* report.

Yield Forecast Surveys

As noted previously, there are two types of crop yield surveys conducted to obtain data for yield forecasting, the grower-reported yield surveys or the Agricultural Yield Survey, and objective measurement surveys, or the Objective Yield Surveys.

Agricultural Yield Surveys

Grower reported surveys, called the Agricultural Yield Surveys (AY), cover most of the field crops estimating program. The survey covers most crop yield data needs for each State. The AY program begins in May using a sample drawn from the list portion of the March Agricultural Survey. This sample is used each month through August and focuses on the small grains: wheat, oats, barley, and rye. The second AY sample is drawn from the list portion of the June Agricultural Survey. This survey is conducted monthly from August through November and includes numerous row crops like corn, cotton, and soybeans, as well as hay, tobacco, and oilseeds.

The AY survey uses a multivariate probability proportionate to size (MPPS) sample design, with list frame data used to determine a unit's selection probability. A more detailed description of this sample design is provided in "Chapter 3 - Agricultural Yield Surveys".

Sample size targets are set for each commodity in AY. The overall sample size varies, depending upon the month, from a maximum of 20,000 in August, to a minimum of 3,500 in June. In AY, targeting is especially important for the commodities that are considered rare commodities or specialty crops.

Objective Yield Surveys

Objective measurement surveys, called Objective Yield (OY), are conducted for corn, cotton, soybeans, and winter wheat. These surveys are very costly and are conducted only in the top producing States. The States in the OY program usually produce more than 75 percent of the U.S. total. For each commodity, a series of monthly net yield forecasts culminates in a final net yield at maturity

As noted in the previous table, all OY samples except winter wheat are drawn from an area frame parent survey. JAS data are collected and recorded at the tract level, multiplied by the inverse of the sampling fraction, and summed to obtain State totals. OY fields are selected systematically from the acres of the crop of interest. In other words, OY samples are selected with probability proportional to size, making them self-weighting samples. The detail of the

recorded area data allows sample selection at the tract level. During the interview with the farm operator, enumerators randomly select fields within the selected tract. Tracts with large acreages or expansion factors may be selected for more than one sample. Separate plots are laid out for each sample within a field up to four samples.

Winter wheat acres are collected at the farm level on the March Agricultural Survey questionnaire, multiplied by the inverse of the sampling fraction adjusted for nonresponse, and totaled in the summary program. Farms are selected probability proportional to size (expanded acres). Fields are selected proportional to size within farm by the enumerator during an interview with the farm operator making this, too, a self-weighting sample. Farms and fields within farms may be selected more than once.

CHAPTER 3 AGRICULTURAL YIELD SURVEYS

The Agricultural Yield Survey (AY) collects farmer assessments of yield prospects monthly, through the growing season. A sample of farmers who reported planting the crops of interest on a parent survey (March or June Agricultural Surveys) are asked to predict their final yield for those crops. The AY fills the yield forecasting needs of most field crops in the NASS estimating program and provides data for all individually published State forecasts. In other words, this survey provides yield indications to ensure the entire program is covered.

The AY survey uses a multivariate probability proportionate to size (MPPS) sample design, with list frame data used to determine a unit's probability of selection. A more basic PPS sample design has their units selected by size depending on the proportion of the commodity of interest the operation has in comparison with other operations on the list frame. The MPPS sample design is similar to a traditional PPS sample design, but there are multiple commodities or frame items used to determine a unit's probability of selection. The MPPS design makes targeting of samples to the desired commodities much easier, improves the sampling efficiency over the traditional stratified design, and simplifies sample designs when there are multiple commodities. In the MPPS sample design, a sample size is targeted for each commodity of interest that has frame data available. A unit's resulting probability of selection is determined by the commodity that has the largest proportion of the total and the sample size for that commodity.

The AY samples are drawn from respondents in list strata in the March and June Agricultural Surveys. A small grains (SG) sample, to be used May through August, is drawn from the March Agricultural Survey from respondents who reported having a small grain crop of interest. A row crops (RC) sample, to be used August through November, is drawn from the June Agricultural Surveys from respondents who reported having a row crop of interest. All records to be included in the AY SG sample were used only in the March quarter of the Agricultural Surveys (with respect to June through March survey year). In a similar fashion, all records to be included in the AY RC sample were used only in the June quarter of the Agricultural Surveys (with respect to the June through March survey year). Excluded from the AY sampling are operations in the largest (preselect) list strata, as well as the non-overlap area tracts – farms identified through our area frame that were not on the list frame.

Since in some months (August in most states, for example) the AY sample includes operations from both samples, a composite weighting methodology was developed. Using such an approach allows maximum use of the information obtained from AY responses. That is, information about SG crops that was obtained from AY RC only sample records can have that SG information used in AY SG survey indications. Similarly, information about RC crops that was obtained from AY SG only sample records can have that RC information used in AY RC survey indications. Under the MPPS sample design, stratification is not used at all as an underlying component. The strata are used, however, in computing nonresponse adjustment weights.

Data Collection

The reference date of every AY survey is the first of the month. The data collection period must span the first, and States are expected to collect data as close to this reference date as reasonable. In practice, the data collection period begins around the 25th of the previous month and ends no later than the 7th of the survey month. This amounts to about seven working days with allowances for weekends.

Survey instruments are prepared in paper and electronic forms. Most data are collected in the electronic form using Computer Assisted Telephone Interviewing (CATI) techniques. Many States will collect some data by mail; however, the short data collection period limits this activity. A small number of samples are interviewed face to face due to special reporting arrangements or other considerations. Computer Assisted Self Interviewing (CASI) via the internet began with the 2006 crop year.

The complete questionnaire for AY includes acres for harvest and yield for each crop of interest. For most crops, planted acres will also be asked in either the initial month of the survey - May for SG and August for RC, the first month the crop appears if it is not asked in the base month, or the first time an operator is interviewed. Actual survey instruments are customized for each month in each State. Some differences may also exist between the paper and CATI version. Acreage responses are retained in the dataset from month to month and these items are not asked on later interviews. The acreage questions are printed on all paper versions of the questionnaire and enumerators are responsible for managing the flow of the interview and recognizing when to ask the acreage questions and when to skip them. The CATI software easily tracks previously reported data and manages the flow of the interview accordingly. The paper and CATI versions also include a question on whether each crop has been harvested and the reported yield is final. Once harvested, the yields for those crops are not asked on subsequent interviews but are brought forward and used in subsequent month's summaries. The AY survey has an additional ability to measure harvested acreage changes during the crop year due to extreme weather conditions when necessary. This distressed acres sub-survey estimates the current acres for harvest versus previously reported acres for harvest and can be targeted at specific crops and States.

The small grains are surveyed May-August and the row crops August-November. The small grains and row crops probability samples overlap in August. A single survey instrument is prepared, and respondents are asked all questions regardless of the sampling base.

States are expected to achieve a minimum response rate of 80 percent. In order to meet this minimum level, States are expected to conduct a telephone follow-up for operations that have not responded. States must also monitor response by crop to determine the amount of follow-up necessary to achieve 50 complete reports for major crops.

Analysis

All AY data are processed through an interactive edit program as the first step in data review. This edit performs all within-record data (micro level data) checks. Data from paper versions are manually reviewed, key entered, and merged with CATI data. Any CASI data is also merged with the CATI data prior to editing. The edit program checks that reported data are within absolute limits, compares acres reported on AY and the parent survey, and compares yields reported by the same reporting unit in consecutive months. This ensures data review is consistent across States. States provide customized edit limits for their data. Most of the edit checks are also performed by the CATI software during the interview which allows enumerators to probe for additional information and correct errors when suspect values are recorded.

The next step is an across-record (macro level data) review of the raw data via the Interactive Data Analysis and Review Tool (iDART). Reported data for all responses are listed for each crop as well as data expanded by probability weights. This allows statisticians to examine data distributions and to identify extreme values that may overly influence the summary results. Data are displayed graphically by district so statisticians can also analyze yield relationships geographically within their State. Statisticians re-examine these values before allowing them to pass to the summary. Some follow-up may be necessary to validate a response. If the data are deemed correct, no action is taken.

iDART displays extreme differences between surveys. In May and August, AY acreage values are compared to acres reported on the parent survey for review. Similarly, month-to-month yield differences are displayed beginning in June for small grains and in September for row crops.

The output is displayed in graphical and tabular form. The graphs show the frequency distribution of the positive data while the tabulations show actual record level data

Summarization

The AY summary program is really two summaries combined. The first part is a probability summary that applies a combination of the appropriate MPPS sample weight and a nonresponse adjustment to produce an indication with associated measurable statistical error. The second part is a non-probability indication which pools the complete reports from all respondents for each crop within an Agricultural Statistics District (ASD). These respondents' yields are then reweighted based on their harvested acres and the acres in their respective ASD for that specific crop. This produces an indication but, because of a lack of probability sampling design, has no measurable error associated with it. The probability and non-probability indications are both sorted by ASD prior to output for comparative purposes.

Probability Summary

The probability summary computes three types of indications:

1. Average expected yield.
2. The ratio of yields reported on consecutive AY surveys.
3. The ratio of any two acreage items from the AY, the parent survey, or both.

Average expected yield

Average expected yield is defined as the expected total production divided by the total acres standing for harvest. For an individual report, production is acres for harvest times expected yield per acre.

The nonresponse weight is calculated using the Agricultural survey stratum each respondent is associated with and is based on the total number of expected respondents within the stratum divided by the total number of complete respondents. The k^{th} stratum nonresponse weight would be calculated as:

$$w_{nr_k} = \frac{\sum_{i=1}^{N_k} w_{MPPS_i}}{\sum_{i=1}^{n_k} w_{MPPS_u}}$$

where

w_{nr_k} = nonresponse weight for stratum *k*
N_k = total sample size for stratum *k*
n_k = number of complete responses within stratum *k*
w_{MPPS_i} = MPPS weight for sample *i* within stratum *k*
w_{MPPS_u} = MPPS weight for complete responses *u* within stratum *k*

The total probability weight for the i^{th} individual record within stratum *k* would then be the product of the MPPS weight and the nonresponse weight:

$$w_i = w_{MPPS_i} w_{nr_k}$$

where

w_i = total weight for record *i*
w_{MPPS_i} = MPPS weight for record *i*
w_{nr_k} = nonresponse weight for stratum *k*

Production is the product of reported current month yield and harvested acreage:

$$p_i = Y_i h_i$$

where

p_i = production for record *i*
Y_i = record *i* yield
h_i = acres of harvested or to be harvested for record *i*

Production is calculated for each ASD using the total probability weight and production for each record, then summed to the State using the following formula:

$$P_S = \sum_{j=1}^{N_D} \sum_{i=1}^{N_{iD}} p_{iD} w_i u_i$$

where

P_S = production for State *S*
N_D = number of ASD in State *S*
N_{iD} = total number of sampled records in ASD *D*
p_{iD} = production for record *i* in ASD *D*
w_i = total weight for record *i*
u_i = 1 if sample *i* is complete, else 0

Similarly, the total number of acres harvested or to be harvested is determined for each ASD and summed to the State:

$$H_S = \sum_{j=1}^{N_D} \sum_{i=1}^{N_{iD}} h_{iD} w_i u_i$$

where

H_S = acres harvested or to be harvested for State *S*
N_D = number of ASD in State *S*
N_{iD} = total number of sampled records in ASD *D*
h_{iD} = acres harvested or to be harvested for record *i* in ASD *D*
w_i = total weight for record *i*
u_i = 1 if sample *i* is complete, else 0

The average expected yield for the State would be:

$$\hat{Y}_S = \frac{P_S}{H_S}$$

where

$\hat{Y}_S$ = average expected yield for State *S*

P_S *and* H_S were previously described

Ratio of Yields

The ratio of yields reported on consecutive AY surveys quantifies the change in the collective judgment of the respondents. Computations are made by converting the current and previous month's sample reported yields to a sample level production value using the last reported acres for harvest and the total weight w_i . ASD and State totals are obtained for each month and the ratio of current over previous provides the measure of change. The b^{th} complete value is1 if complete for current and previous month, and 0 if not complete for both months. The equation for State *S*:

$$R_S = \frac{\sum_{j=1}^{N_D} \sum_{i=1}^{N_{iD}} y_{iD_c} h_{iD_c} w_i b_i}{\sum_{j=1}^{N_D} \sum_{i=1}^{N_{iD}} y_{iD_p} h_{iD_p} w_i b_i}$$

where

R_S = ratio of current month's expected yield to previous month's expected yield for State *S*

N_D = number of ASD in State *S*

N_{iD} = total number of sampled records in ASD *D*

y_{iD_c} = current expected yield for record *i* in ASD *D*

h_{iD_c} = current acres harvested or to be harvested for record *i* in ASD *D*

w_i = total weight for record *i*

b_i = 1 if sample *i* is complete for current and previous month, else 0

y_{iD_p} = previous expected yield for record *i* in ASD *D*

h_{iD_p} = previous acres harvested or to be harvested for record *i* in ASD *D*

Ratio of two acreage items

The ratio of any two acreage items from AY and the parent survey offers several key indicators. Any ratio of AY reported acres to the acres reported on the parent survey provides a link between the AY and the parent survey. Every AY probability sample unit matches a parent survey response. This affords the opportunity to calculate ratios of acres reported on both surveys. The acres used vary between crops. For most crops ratios of harvested acres are computed, for a few crops planted acres are used, and a few crops have both. Acreage ratios provide an assessment of the presence or absence of reporting errors in the data used to determine the AY subsampling population. This ratio is expected to be near 1.0. Deviation from 1.0 can indicate a reporting error and/or changes in growing conditions such as delayed planting.

In a drought year, harvested to previous harvested ratios provide an indication of increasing abandonment. Under these conditions, reporting errors become confounded with true change in the acreage level. The AY survey is not designed to provide clean measurement of current year acreage changes; however, it can provide limited information about current acreage changes.

The general formula for the ratio between surveys for State *S* is shown below. Note that the current AY weights apply in all instances.

$$A_S = \frac{\sum_{j=1}^{N_D} \sum_{i=1}^{N_{iD}} a_{iD_c} w_i b_i}{\sum_{j=1}^{N_D} \sum_{i=1}^{N_{iD}} a_{iD_p} w_i b_i}$$

where

A_S = ratio of current month's acreage to the parent survey acreage in State *S*
N_D = number of ASD in State *S*
N_{iD} = total number of sampled records in ASD *D*
a_{iD_c} = current acreage for record *i* in ASD *D*
w_i = total weight for record *i*
b_i = 1 if sample *i* is complete for current and previous month, else 0
a_{iD_p} = parent survey acreage for record *i* in ASD *D*

The formula for the harvested to planted acreage ratio for State *S* is:

$$\frac{H_S}{P_S} = \frac{\sum_{j=1}^{N_D} \sum_{i=1}^{N_{iD}} h_{iD} w_i b_i}{\sum_{j=1}^{N_D} \sum_{i=1}^{N_{iD}} p_{iD} w_i b_i}$$

where

$\frac{H_S}{P_S}$ = ratio of acres harvested to acres planted in State *S*
N_D = number of ASD in State *S*
N_{iD} = total number of sampled records in ASD *D*
h_{iD_c} = acres harvested for record *i* in ASD *D*
w_i = total weight for record *i*
b_i = 1 if sample *i* is complete, else 0
p_{iD_p} = acres planted for record *i* in ASD *D*

Although all calculations in the probability summary are performed within design stratum, the printed output presents the results by ASD. This facilitates the interpretive process by making it

easier to compare AY results to other data sources reported by ASD. Temperature, precipitation, and crop progress data provide additional evidence to support the AY indications to build a complete picture of current conditions. An additional analytical benefit is gained from the ability to see a geographic breakdown.

The summary program derives yields and ratios by expanding the sample level data and grouping the samples by ASD. Key variables are summed to obtain ASD totals and the yields and ratios are computed. It is important for commodity analysts to remember that even though the summary shows the results by ASD, the calculations performed at the sample level follow the design strata.

Non-Probability Summary

The non-probability portion of the summary treats the pooled dataset as a simple random sample. The data are partitioned by ASD. Sample weights are 1.0 for all reporting units. Reported yields are weighted by acres for harvest when computing ASD means. ASD means and ratios are weighted to the State level using externally provided historical district harvested acreage estimates.

The same three types of indications are calculated. The non-probability formulas similar to the probability at the ASD level, with design weights (w_i) eliminated. State level formulas use the ASD level calculations and then applies external weights.

Average expected yield

Average expected yield is a weighted average of reported yields with harvest acres serving as weights. The yield for ASD *D* would be calculated as follows:

$$\hat{Y}_D = \frac{\sum_{i=1}^{N_{iD}} y_{iD}\, h_{iD} k_i}{\sum_{i=1}^{N_{iD}} h_{iD} k_i}$$

where

$\hat{Y}_D$ = expected yield for ASD *D*
N_{iD} = total number of sampled records in ASD *D*
y_{iD} = yield for record *i* in ASD *D*
h_{iD} = parent survey acreage for record *i* in ASD *D*
k_i = 1 if sample *i* is complete, else 0

Ratio of Yields

Similarly, the ratio of yields reported on consecutive AY surveys is:

$$R_D = \frac{\sum_{i=1}^{N_{iD}} y_{iD_c} h_{iD} k_i}{\sum_{i=1}^{N_{iD}} y_{iD_p} h_{iD} k_i}$$

where

R_D = ratio of current to expected yield to previous month's expected yield for ASD *D*
N_{iD}, h_{iD}, and k_i were previously described
y_{iD_c} = current yield for record *i* in ASD *D*
y_{iD_p} = previous yield for record *i* in ASD *D*

Ratio of two acreage items

The ratio of acreage items between AY and the parent survey and the harvested to planted ratio for ASD *D* are derived as:

$$A_D = \frac{\sum_{i=1}^{N_{iD}} a_{iD_c} k_i}{\sum_{i=1}^{N_{iD}} a_{iD_p} k_i}$$

where

A_D = ratio of current month's acreage to the parent survey acreage in ASD *D*
N_{iD} = total number of sampled records in ASD *D*
a_{iD_c} = current acreage for record *i* in ASD *D*
k_i = 1 if sample *i* is complete for current and previous month, else 0
a_{iD_p} = parent survey acreage for record *i* in ASD *D*

The formula for the harvested to planted acreage ratio for ASD *D* is:

$$\frac{H_D}{P_D} = \frac{\sum_{i=1}^{N_{iD}} h_{iD}\, k_i}{\sum_{i=1}^{N_{iD}} p_{iD}\, k_i}$$

where

$\frac{H_D}{P_D}$ = ratio of acres harvested to acres planted in ASD *D*
N_{iD} = total number of sampled records in ASD *D*
h_{iD_c} = acres harvested for record *i* in ASD *D*
k_i = 1 if sample *i* is complete, else 0
p_{iD_p} = acres planted for record *i* in ASD *D*

State Level Estimates

The ASD level non-probability estimates shown above are weighted to the State level using external weights. These external weights are derived from historical district harvested acreage estimates. Let E_D denote any ASD level non-probability estimate shown above. The State *S* level estimate, E_w is:

$$E_w = \frac{\sum_{i=1}^{N_D} w_D E_D}{\sum_{i=1}^{N_D} w_D}$$

where

E_w = weighted State *S* estimate E_D
N_D = total number of ASD in State *S*
w_D = external weight for ASD *D*
E_D = ASD *D* level non-probability estimate

Survey Bias

Note that AY data are the respondents' expected yield. The crop may not be mature and ready for harvest for another several weeks after the respondent has been contacted. Experience has shown these responses tend to be conservative (biased down). Under drought conditions, this bias gets much larger as respondents' perceptions of a crop are influenced by current weather conditions. Therefore, the interpretation phase of the review must recognize this tendency and factor it into the final deliberations.

CHAPTER 4 OBJECTIVE YIELD SURVEY AND GENERAL YIELD FORECASTING PROCEDURES

This chapter presents basic sampling, data collection, and mathematical methods commonly employed to forecast the yield and production of any economically valuable agricultural crop. The sampling review will emphasize the reliance objective yield survey forecasts have to a well-defined population and the traditional sampling methods employed in objective yield studies. Data collection will focus on the plant characteristics and measures commonly collected for objective yield programs. Finally, the mathematical methods review will center on mathematical/statistical estimation principles common to forecasting any crop's yield and production.

Sampling

NASS currently expends federal funds to forecast crop yields in 24 of the lower 48 states. NASS also supports yield forecasting for fruits and nuts in Florida and California. For most crops, yield forecasting begins with partitioning each states' land area into uniquely identifiable pieces called segments. The resulting collection of segments is a complete population of all land area in each state and is referred to as an "*Area Frame*". This painstaking segmentation of land area into uniquely identifiable segments makes possible the selection of a probability-based sample of crop acres for objective yield (OY) field enumeration.

The most important statistical result of area frame construction is that any crop acre can be assigned a known inclusion probability for OY studies. For annual crops like corn, NASS's *Area Frame* explicitly identifies and separates each corn acre from every other possible utilization and makes probability calculations possible. For perennial crops like Florida citrus, a *Citrus Area Frame* allows separation of citrus acres from other possible utilizations allowing probabilities of selection to be established. The effort required to identify and separate each type of utilization from another is a common feature of agricultural area frames. In return for the effort, statistically defensible estimates of crop yields are obtainable.

Accurate inferences from probability-based surveys require the construction of a sampling population and, in addition, a method to select a representative sample from that population. Probability proportional to size (PPS) is the most common statistical method employed by NASS for selecting crop acres to be included in objective yield studies. Since a given crop acre is most often embedded within an individual *field* composed of multiple acres, PPS sampling ensures that a proper accounting is made for the size of field. In agriculture a *field* is one, continuous acreage of land devoted to the same use. Since *fields* are the unit of selection for objective yield studies and because *fields* vary in acreage, PPS sampling is ideally suited to selecting a sample that perfectly reflects the population characteristics of the particular crop. For example, employing PPS sampling implies a crop field accounting for twice the percentage of total state corn acres relative to a neighboring corn field, will also carry twice the probability of selection as

the neighbor. The representative *acre* selected from a sampled *field* is finally accomplished with simple random sampling. Since a selected *acre* is too large for complete enumeration, a selected *sample* will be enumerated and expanded up to the one-acre level.

Data Collection

A full OY survey collects data at different times during the growing season. The following paragraphs describe the data collected and the how the data are used in the forecasting and estimating process.

During the initial OY survey, the operator is asked to verify the acreage reported in the parent survey. This is accomplished on a field-by-field basis. Changes may be due to recording or reporting errors in the parent survey, failure to fulfill planting intentions, or switching to other utilizations. Other data concerning the crop, for example planting date and use of genetically modified seed, are collected at this time. The final question asks for permission to enter the sample field and make counts and measurements through the growing season.

The initial visit to the sampled acre involves precise selection of plants from which fruit counts, measurements, and maturity determinations are to be made for the remainder of the growing season. Each sample is identical in its dimensions according to crop. Extra precautions during sample layouts ensure the exact same plants are revisited in subsequent months. Plant counts and a variety of plant characteristics collected from these samples will be used to forecast gross yield and the components of yield such as number of fruit and weight per fruit. The final visit obtains all harvestable yield, which will determine final gross yield. The counts and measurements from all visits are added to a five-year historical database which will be used to forecast gross yields the following season.

After the farmer has harvested the sample field, postharvest gleaning data are collected. All un-harvested fruit and loose grains are gleaned from specially prepared plots that are separate from the original sample plot. The calculations from these plots will be used as a deduction from gross yield which provides a net yield number. Before gleanings data become available, a five-year average harvest loss is assumed for the calculation of net yield.

Models

Linear statistical models are used extensively in forecasting and estimating crop yields and production. Other methods employed are variable plots, scattergrams, and tables. The basic agronomic model used to conceptualize final net yield per acre is:

$$Y = (F * W) - L$$

where

Y = the population net yield per acre,
F = the population average number of fruit per acre,
W = the population average net fruit weight per unit, in industry standard moisture, and
L = the population average harvest loss per acre.

This model can be estimated as follows:

First, forecast the average number of fruit per sample as

$$f_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Where f_i is the estimate for average number of fruit for sample i, β_0 and β_1 are historic coefficients estimated by ordinary least squares from five years of data, x_i is some characteristic such as plant count for sample i, and ε_i is an error term exhibiting classical properties. Each f_i is expanded to the acre level after the forecast is made.

Secondly, forecast the average weight of fruit per unit as:

$$w_i = \beta_2 + \beta_3 z_i + \varepsilon_i$$

Where w_i is the estimate for average weight of fruit per unit, e.g., weight of corn per ear, weight of soybeans per pod, weight of cotton per boll, weight of wheat per head, etc., β_2 and β_3 are historic coefficients estimated by ordinary least squares from five years of data, z_i is some characteristic such as average fruit size from sample i, and ε_i is an error term exhibiting classical properties. Each w_i is converted to industry standards for moisture when required.

Third, forecast the average yield loss per sample as:

$$\bar{L} = \frac{1}{N_L} \sum_{i}^{N_L} L_i$$

Where L_i is the historic sample harvest loss in the i^{th} sample expanded to one acre, then summed from the first sample to the N_L sample and divided by N_L which gives $\bar{L}$, the average historic loss per acre. After the current years' final harvest, each sample's harvest loss reflects this years' loss statistics.

Finally, substitute each component back into the agronomic model:

$$Equation\ 1. \quad \bar{y} = \frac{1}{N}\sum_{i}^{N}(f_i * w_i - \bar{L})$$

Where $\bar{y}$ is now an estimate of the population average net yield per acre in industry standard moisture and volume.

The models used to forecast fruit count and fruit weight above may also be independently evaluated from the general agronomic model as in:

$$Equation\ 2. \qquad \bar{f} = \frac{1}{N}\sum_{i}^{N} f_i$$

Where $\bar{f}$ is now an estimate of the population average number of fruit per acre, or:

$$Equation\ 3. \qquad \bar{w} = \frac{1}{N}\sum_{i}^{N} w_i$$

Where $\bar{w}$ is now an estimate of the population average fruit weight per unit in industry standard moisture.

Component forecasts, represented by equations 2 and 3, are evaluated with statistical plots, charts, and scatter grams to assist statisticians forecasting yields. Analysts will investigate component fruit counts and weights relative to the current years growing conditions and compare to previous years' conditions. An in-depth examination is made relative to potential impacts of outliers on component forecast and in turn on net yield forecasts. In addition to component forecasting, other more esoteric relationships are examined during the course of yield forecasting.

Regressed to Board Indications

Time series models are also estimated for each aggregate using the expression:

$$Y_t = \beta_4 + \beta_5\bar{y}_t + \varepsilon_t$$

Where Y_t, is the known population final net yield per acre in time period *t*, β_4, and β_5 are coefficients to be estimated using fifteen years of data, $\bar{y}_t$ is the average net yield aggregate from *Equation* 1 for time period *t* , and ε_t is an error term exhibiting classical properties. Individual years' that have absolute studentized deleted residuals (see Neter, Wasserman, and Kutner [2], page 406) greater than three are not used for estimating β_4 and β_5. Similarly, other models are

estimated, such as:

$$F_t = \beta_6 + \beta_7 \bar{f}_t + \varepsilon_t$$

Where F_t, is the known population final fruit count per acre in time period *t*, β_6, and β_7 are historic coefficients to be estimated from fifteen years of data, $\bar{f}_t$ is the aggregate average number of fruit per acre from *Equation* 2 for time period *t* , and ε_t is an error term exhibiting classical properties.

Strengths and Weaknesses of Each Model

The strength of the sample level models is that each model is estimated at every level of maturity, for example the sample level model estimating fruit count per acre, as given above was:

$$f_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

which is estimated for each crop maturity classification. For example, corn has six forecasting classes, thus up to six equations are estimated for every month and geographic area represented in the survey. A similar set of equations are estimated for weight of fruit per unit by maturity class. If x_i is one plant characteristic used to forecast fruit count per acre, but r_i is an alternative plant characteristic, which can also predict fruit count per acre, the process repeats. This approach to estimation allows for a high degree of complexity. The weaknesses of sample level models are that plant measurements are very sensitive in nature and require extra caution in their collection and the data typically exhibit large variation within and across years.

A statistical benefit of aggregate level models like *Equations* 1, 2, and 3 is the reduction in mean variation of net yield, or fruit count, or fruit weight as sample sizes increases across all maturity classes. Additionally, aggregated survey means are well correlated with general changes in agricultural technology across time. Disaggregated sample level models cannot capture technological change from relatively short five-year time periods, limited geographic area, and single maturity classifications. Another advantage for aggregate models is that their dependent variables have very small measurement error due to end of year administrative data. A noted weakness of aggregated models is that maturity levels across many years are never equal as of a date certain on the calendar. By their nature, aggregated models also have few degrees of freedom.

Acreage Indications

Acreage adjustment ratios are another byproduct of objective yield surveys. Data collected in the initial interview may be compared to data obtained in the parent survey. Acreage adjustments are not the main purpose of OY surveys and thus are not designed for great precision but to detect

gross changes in acreage.

There are three acreage adjustment ratios:

1. The R1 ratio is a harvest intentions ratio. For crops sampled from the area frame, it is the ratio of total acres intended for harvest in the tract to total acres planted in the tract. Acres intended for harvest are reported on the initial OY interview and acres planted are reported on the base survey.

2. For cotton, the R2 is the ratio of planted acres to the planted acreage from the JAS. Thus, the R2 measures the change in planting intentions for cotton since the JAS.

3. The abandonment ratio is used each month to adjust for samples destroyed or abandoned, that is, "lost" samples. The numerator of the ratio equals the total number of lost samples, and the denominator is the total number of samples, including lost samples. An active sample is where harvest has occurred or is expected to occur.

Model Based Yield Indication

The Agricultural Statistics Board (ASB) examines each indication to decide the best forecast in any given month. One such indication introduced in 2015 is a model-based yield indication that combines the farmer reported survey, the objective yield survey, and a linear covariate model with crop condition, weather data, and trend as predictor variables. This additional indication has proved to be a valuable piece of information when forecasting crop yields. The ASB's procedures for selecting the official forecast has not changed; however, this new indication provides additional insight in considering the official forecast for any given month. Currently, winter wheat, corn, soybeans, and cotton have this model-based yield indication. There are no plans to introduce any other crops.

Three components are used to generate this indication: farmer reported survey, objective yield survey, and a linear regression model using crop condition, weather data, and trend as predictor variables. The linear regression will be hitherto referred as the covariate model. The farmer reported survey and the objective yield survey are adjusted for bias for each state and for each month. In most cases, the farmer reported survey is adjusted upward while the objective yield survey is adjusted downward. This bias is an average bias comparing the final ASB yield to those respective indications for that month.

The covariate model has four predictor variables that are regressed to the final ASB yield. The simplest of these is a trend variable. Because crop yields have shown a strong linear trend upward, this trend variable is simply the forecasting year minus a constant for easy calculation. Monthly average temperature and precipitation observed in the prior month

are also used as predictors. For corn, only July weather data are used for this covariate model because July is the most crucial month for corn development. For soybeans, July weather data are used for the August forecast whereas August weather data are used for the September and later forecasts. For wheat, the weather data used are specific to the state because weather across the wheat speculative region is quite different. The final predictor variable in this model is the crop condition measures from the NASS weekly Crop Condition report. For corn, only the percentage of good to excellent corn in week 30 is used for all forecasting months. For soybeans, week 34 is used for each month. Wheat varies by state. The following table shows a timeline for updating the covariate model for wheat.

		May Covariates		**June Covariates**		**July Covariates**		**August Covariates**	
State/		*Week*	*Weather*	*Week*	*Weather*	*Week*	*Weather*	*Week*	*Weather*
FIPS		*#*	*Month*	*#*	*Month*	*#*	*Month*	*#*	*Month*
CO	8	15	April	21	May	21	May	21	May
IL	17	15	April	19	May	19	May	19	May
KS	20	15	April	19	May	19	May	19	May
MO	29	15	April	19	May	19	May	19	May
MT	30	15	April	19	May	24	June	24	June
NE	31	15	April	21	May	21	May	21	May
OH	39	15	April	21	May	21	May	21	May
OK	40	15	April	17	April	17	April	17	April
TX	48	15	April	17	April	17	April	17	April
WA	53	15	April	22	May	22	May	22	May

A model is also calculated at the speculative region level. The covariate component is simply a weighted average of the state level covariates with weights being the harvested acreage. The survey components are directly inputted from the summaries of those respective surveys which generate a speculative regional yield. Because the speculative region has less variability, each state's model-based indication must be adjusted to match the speculative region as a whole.

Given all three indications (the bias adjusted farmer reported survey, the bias adjusted objective yield survey, and the covariate model), weights are then assigned to each indication to calculate a single model-based yield indication. These weights are determined by how well each indication does at predicting the final ASB yield. The weights are determined in proportion to the inverse of the variance for each indication according to the formula below:

$$w_k = \frac{1/V_k}{1/V_{OY} + 1/V_{AY} + 1/V_{COVAR}}$$

Where *k* in {*OY, AY, COVAR*}.

Generally, the weights for the covariates and farmer reported survey are the greatest early in the forecasting season. As the season progresses, more weight is assigned to the farmer reported survey and the objective yield survey as less weight is attributed to the covariates. The covariate model all but drops out in the last two months of the season as more weight is given to each of the two survey indications.

For more information on this indication please refer to the paper "A Bayesian Hierarchical Model for Combining Several Crop Yield Indications" by Cruze (2016).

Remotely Sensed Empirical Model Indication

The latest indication considered by the ASB, introduced in 2019, is a within-season, remotely sensed-based corn and soybean yield estimation. This indication is made available to the ASB from August through October and uses a linear regression model for prediction as a function of the Normalized Difference Vegetation Index (NDVI) values accumulated over the growing season. The NDVI measurements come from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) which provides a landscape-level snapshot of vegetation conditions every eight days. The imagery is refined further by constraining it to crop-specific areas via an ancillary bitmask image derived from the NASS Cropland Data Layer. Ultimately, the NDVI data are averaged to the state or regional area of interest annually and related to the NASS historically published yields going back to 2002, the start of the MODIS era, to build the empirical model. The image handling is currently performed through the USDA supported NASA Global Agricultural Monitoring website and statistical aspects undertaken within NASS.

CHAPTER 5 CORN OBJECTIVE YIELD METHODS

This chapter presents basic sampling, data collection, and mathematical methods utilized to estimate U.S. corn yield and production. The sampling review will emphasize NASS's unique *Area Frame* and its suitability for sample selections intended to measure U.S. corn yields. Data collection will focus on the variety of data collected and how data collection changes as corn crop maturity changes. The mathematical methods review will center on mathematical/statistical estimation of final corn yields utilizing measurable plant characteristics observed at specific points during the growing season. The end results of careful sampling, data collection, and mathematical methods are numeric indications suitable to establishment, and/or revision, of U.S. corn acreage, yield, and production.

Sample Design

Corn Objective Yield (COY) surveys are conducted in Illinois, Indiana, Iowa, Kansas, Minnesota, Missouri, Nebraska, Ohio, South Dakota, and Wisconsin. On average, over the last three years, these ten states produced more than 80 percent of the U.S. corn crop. As described in Chapter 2, NASS partitions each of these states' land area into approximately one square mile pieces and calls the pieces segments. The resulting collection of segments is referred to as the "*Area Frame*". This painstaking segmentation of land area into uniquely identifiable segments makes possible the selection of a probability-based sample of U.S. corn acres.

Probability based surveys require the explicit identification and separation of all elements contained in a population of interest. For yield surveys like COY, the most important result of constructing a population of corn acres is that ***any*** acre of corn, in a given state, can be assigned a known probability of selection. This result allows statistical samples to be extracted, population parameters to be estimated, hypotheses to be tested, and inferences to be made. NASS's *Area Frame* is the vehicle that transports us from a very large population of U.S. corn acres to a representative sample of U.S. corn acres. The sample is both statistically defensible and economically feasible to enumerate. The procedure for identifying every U.S. corn acre is to first, select a subsample of *Area Frame* segments, and secondly, to account for all agricultural activity occurring within those segments. Each selected segment will have all its acres inspected by and accounted for by professional enumerators with assistance from the owner/operator. In this manner, a new listing of corn acres is constructed each year. This listing is the COY sample population. No other purveyor of U.S. corn estimates can construct such a sample population.

Accurate inferences from a probability-based survey require the construction of a sampling population and, in addition, a means to select a representative sample from that population. The statistical method used to select corn acres to be included in COY is probability proportional to size (PPS). In statistics, this method ensures a representative sample is selected when sample elements vary in their size. A simple explanation should be given as to how sample elements in the corn acres population can differ in size. During constructing of the corn acres sample

population, the explicit identification and separation of corn acres, from all other acres, is made according to *fields*. In agriculture a field is one, continuous acreage of land devoted to the same use. Therefore, the sample elements in the corn population are *fields*. Since *fields* of corn vary in acreage, the sample selection method of PPS ensures the probability of selection is adjusted for field size. For example, employing PPS sampling, a corn field that is twice the size of its neighbor will also be twice as likely to be included in the sample. After the determination of which fields are included in the sample, the representative *acre* is selected with simple random sampling designed to give every acre in the selected field, an equal chance of selection. Since a selected *acre* is too large for complete enumeration, a selected *sample* will be enumerated and expanded up to the one-acre level.

In mid-July of each year, professional statisticians in the ten COY states train field enumerators to properly identify and prepare selected samples for data collection. Generally, enumerators will have many years of experience in COY data collection and are well qualified to conduct their assignments. Practical field training is also available through relationships developed among individual enumerators and their supervisors. Overall, the preparation for data collection is rigorous and includes quality control processes that continue through the corn growing season. In late July, field enumerators will begin visiting and enumerating samples and will continue personal visits at monthly intervals throughout the season until final harvest.

Each sample consists of two units (or plots) to be utilized in forecasting final net yield per acre. Each unit consists of two parallel 15-foot sections of row. During each visit, enumerators count or measure each required plant characteristic, the required characteristics being determined according to the units' maturity level. For example, at early maturities enumerators will count plants and measure row spacing. At later maturities, ear lengths, diameters, and weights are measured and recorded. After each data collection period, a forecast of each unit's final net yield is constructed. Essentially forecasts of corn yields depend on two items: One, the current year's measurable plant characteristics and, two, the historic relationship between plant characteristics measured in the past, at the same level of maturity, and final plant characteristics that result in harvestable yield. Each item is needed to forecast a samples final net yield per acre. In cases where one or the other of the two required items is missing, a five-year average of the final plant characteristic is substituted. During some early forecasts there may be no plant characteristic that NASS measures that can be related the final plant characteristic. Ear weight is the prime example. In these cases, historic averages are used to forecast final net yield.

Data Collected

Field enumerators count and measure several items within or near the units. The size of the unit (square feet), the number of ears, grain weight, and harvest loss. The following lists the data items collected and what it is used to measure.

Data items used to measure the size of each unit:

Distance between two rows (one row middle)
Distance between five rows (four row middles)

Data items used to forecast or estimate the number of ears:
Number of stalks in each row
Number of stalks with ears or silked ear shoots in each row
Number of ears and silked ear shoots in each row
Number of ears with kernel formation

Data items used to forecast or estimate grain weight:
Kernel row length from the first five ears beyond the unit in a specified row
Ear diameter at a point one inch from the butt on the same five ears beyond the unit
Weight of the first five ears in the dent stage (when the sample reaches this maturity)
Weight of shelled grain from the five dent stage ears
Moisture content of grain from the five dent stage ears
Field weight of all ears in the two units at maturity (crop cutting)
Lab weight of sample of four mature ears harvested
Weight of grain shelled from the four mature ears
Moisture content of shelled grain from the four mature ears

Data items used to estimate harvest loss:
Distance between two rows (one row middle)
Distance between five rows (four row middles)
Grain weight of ears between Row 1 and Row 3
Grain weight of loose kernels between Row 1 and Row 2

Maturity Categories

At each visit, the enumerator makes maturity assessments within the units and a maturity category is established for the sample. If necessary, ears outside the unit may be husked to make this determination. Forecast equations are derived using data collected during the previous 5-years for each maturity in each month. The maturity definitions used by the enumerators are:

Maturity	Definition
1 - no ear shoots	No ears or ear shoots are present.
2 - pre-blister	Ear shoots are present with some silks showing. Most silks are yellow to white in color. Spikelets contain little or no liquid.

3 - blister	Most silks protruding from husks are beginning to turn brown. Spikelets have swollen and contain clear to white liquid.
4 - milk	Silks protruding from husks have turned brown and dry. Plant or shuck is green. Ears are erect. Kernels contain a milk-like substance and show little or no denting.
5 - dough	Shucks starting to take on a light rust color. Ears beginning to lean away from stalks. About half the kernels are dented and contain a milk or dough-like substance. Maturity line has not moved halfway to the cob on a majority of the kernels.
6 - dent	Shucks are dry but not opening up. Nearly all kernels are dented. Maturity line on kernels has not reached the cob.
7 - mature	Shucks are dry and opening up. Ears are hanging down from the stalk. Maturity line on kernels has reached the cob.

Analysis of Raw Data

In the following sections estimation at the sample level is discussed. All sample level regression equations are derived from the previous 5 years' survey data using multiple regression techniques. Certain influential data points (i.e., "outliers") are excluded from the dataset prior to deriving the coefficients. These influential data points are identified using a "deleted residual" analysis or the "Cook's D" statistic (Belsley, Kuh, and Welsch, [4]). There is usually very little change in the regression equations from year-to-year because roughly 80 percent of the data for each class were used in the analysis the previous year. Classes that do change significantly from one year to the next are usually those with very few observations. If a class has little data and a plausible forecast equation cannot be derived, the equation from the previous year is used.

Forecasting and Estimating Number of Ears for Sample Fields

The mathematical expression for calculating the number of ears per acre from a sample plot is straightforward, i.e., a specific count of ears per 60 linear feet and a measurement of average row spacing for the sample results in a unique value of ears per acre. The formula for calculating the number of ears per acre is:

$$\text{ears per acre} = \frac{\text{(ears in sample plots) (43,560)}}{\text{(60) (average row space)}}$$

Where 43,560 is the number of square feet in an acre, 60 is the total length of rows in two units, and average row space is the sum of the two 4-row space measurements divided by 8.

In the equation above, forecasts for ears per acre are dependent on the number of ears in the sample plots. Should the ears in a sample plot not be observable due to immaturity, then a forecast of the final number of ears before farmer harvest is required. NASS bases this prediction on the historic relationship between observable plant characteristics and the eventual outcome for number of ears. NASS carefully considers which plant characteristics are used to make ear count forecasts based on the historic performance of the particular plant characteristic in question.

When samples are in maturities categories 1-4, two models are used to forecast the number of ears in each sample. The first model uses five years of historic data to estimate the relationship between final ears per sample and the historic stalk count from the same month.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where x_i is the number of stalks in sample *i*, y_i is the final number of ears for sample *i*, and ε_i are random departures from the relationship. The coefficients β_0 and β_1 are estimated by ordinary least squares from data collected during the previous five years. This model produces historic results with R^2's typically in the 80 to 90 percent range for all states. The second model uses five years of historic data to estimate the relationship between final ears per sample and the ratio of stalks with ears to total stalk counts per sample. This model may be represented by,

$$y_i = \beta_2 + \beta_3 z_i + \epsilon_i$$

Where z_i is the number of stalks with ears or silked ear shoots (maturity category 1-4) divided by the total number of stalks in the sample, y_i is the final ratio, measured at or before maturity category 4, of stalks with ears to total stalks and ϵ_i are random departures from the relationship. The coefficients β_2 and β_3 are estimated by ordinary least squares and a forecasted number of ears per sample is formulated by dividing the current count of stalks with ears by the predicted value of the ratio. This model produces R^2's that are much lower than the simple stalks model and, in addition, are generally variable across states and within states across maturity categories.

The forecast models for ears per sample are a weighted combination of the two regressions.

Model 1: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Model 2: $y_i = \beta_2 + \beta_3 z_i + \epsilon_i$

A composite is constructed for forecasting ears per sample as follows:

$$\hat{Y}_i = \frac{R_1^2\ (\hat{y}_{i1}) + \ R_2^2(\hat{y}_{i2})}{R_1^2 + \ R_2^2}$$

where, $\hat{Y}_i$ is the weighted forecast of ears per sample from Model 1 and 2, and R_1^2 and R_2^2 are the multiple correlation coefficients for Model 1 and 2 respectively. Typically, the performance of Model 1 completely dominates Model 2.

Samples classified in dough stage or higher use the actual count of ears with evidence of kernel formation as the forecasted number of ears in the sample. Also, for the final visit to the sample, the actual ear count is used, regardless of the maturity category.

Forecasting and Estimating Grain Weight Per Ear for Sample Fields

Corn is no exception to the general rule of objective yield surveying, i.e., successfully forecasting fruit weight is more difficult than forecasting fruit count. Therefore, it should be no surprise that over the course of time and program development, one model has been successfully employed for ear count predictions whereas multiple models are utilized forecast final weight per ear. The sensitivity of final net yield to grain weight per ear necessitates a careful consideration of all plant characteristic which may contribute to successfully forecasting final grain weight per ear.

When samples are in maturity categories 1 and 2 the grain weight per ear will be forecasted to the final utilizing the 5-year historical average grain weight per ear. For the September 1 survey, this average is computed from all samples with final lab grain weights during the last 5-years. For the October 1 and later surveys, this average is computed using only samples that were in maturity category 1 or 2 (no ear shoots and pre-blister stage) for October 1 or later surveys from the last 5-years. The October 1 historic average is rarely used.

Corn samples in maturity categories 3 through 6 use historic regression models to forecast final grain weight per ear. There are currently three models calculated that are based on the following plant characteristics: kernel row length, ear volume, and maturity code 6 harvested ears. Kernel row length measurements are taken just beyond the 15-foot count section of each unit. Kernel row length measures the average length of five cobs from one inch above the butt to the end of the cob. These measurements, collected over a series of years, can be utilized to forecast future sample grain weights. Ear volume measures are calculated by combining kernel row length measures with cob diameter measurements. These too are historically related to final grain weights. Lastly, maturity code 6 harvested ears are collected just beyond the 15-foot count

section and are then laboratory weighed and adjusted to 15.5 percent moisture. These MC6 weights are also related to final grain weights by means of regression.

In general, the form of the three models discussed above is:

$$wt_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Where x_i is the average kernel row length for sample *i* or the computed volume measurement of the ears in sample *i*, or the MC6 grain weight for sample *i*. wt_i is the final grain weight for sample *i* and ε_i is a random departure from the relationship. The coefficients β_0 and β_1 are estimated by ordinary least squares from the previous five years of data.

Parameter estimates (β_0 and β_1) are calculated for each maturity category in each month for each State. Each of these models exhibits significant variation but, in general, each improves in predictive ability over the course of a growing season. R^2's are in the 20's, 30's and 40's for ear volume and slightly less for kernel row length. The R^2 for MC6 weights does show the expected improvement given the later maturity, however, by statistician standards, has considerable remaining variation. Ultimately, it is difficult to make both an early and confident prediction of final grain weights per ear utilizing kernel row length, ear volume, and MC6 weights.

A sample is enumerator harvested when three or more ears of the first ears beyond a unit are mature or the farmer intends to harvest the sample within three days. If a final harvest sample is missed by an enumerator, its most recent forecast is perpetuated. The average field weight per ear is an average of the combined ear weight (cob and kernels) from all ears harvested in sample *i*:

$$Field\ Wt = \left[\frac{Total\ wt\ of\ ears}{count\ of\ ears}\right]$$

A conversion must be made to adjust this field weight to a shelled ear weight at 15.5 percent moisture. The conversion factor is calculated in one of two ways:

1. When lab data are available, the adjusted weight per ear is calculated by:

$$\frac{Wt}{Ear} = Field\ Wt * \left[\frac{w_s}{w_4 - b}\right] * \left[\frac{\frac{1 - m_i}{100}}{.845}\right]$$

where, for sample *i*, w_s is the weight of all grain shelled from four ears, w_4 is the weight of four ears (including cob), plastic bags and rubber bands (as mailed), b is the weight of plastic bags and rubber bands, m_i is the moisture content of the shelled grain, and .845 = (100 - 15.5/100) adjusts to standard moisture.

2. If lab data are not available, a 5-year historical average shelling fraction and moisture adjustment is applied to the average field weight.

Independent Variables used in Sample Level Forecasts and Estimates

The following table summarizes the data items used to estimate or forecast the number of ears, weight per ear and harvest loss for each of the 7 maturities:

Maturity	Number of Ears	Weight per Ear	Harvest Loss
1 No ears or ear shoots	Stalks	5-year average	5-year average
2 Pre-blister	Stalks	5-year average	5-year average
3 Blister	Stalks	Kernel row length/ Volume	5-year average
4 Milk	Stalks	Kernel row length/ Volume	5-year average
5 Dough	Ears with kernels	Kernel row length/ Volume	5-year average
6 Dent	Ears with kernels	Kernel row length Grain weight per ear	5-year average
7 Mature	Ears with kernels	Grain weight per ear	Realized Harvest Loss
Final	Ears with kernels	Grain weight per ear	Realized Harvest loss

Forecasting Yield for Sample Fields

The gross yield for sample i is calculated by:

$$\widehat{GY}_i = \frac{\hat{y}_i * \widehat{wt}_i}{56}$$

Where $\hat{y}_i$ is the forecasted or final estimate for ears per acre, $\widehat{wt}_i$ is the forecasted or final average grain weight per ear in pounds at 15.5 percent moisture, and 56 converts the numerator to bushel per acre

State Average Forecasts and Estimates

The sample level gross yield forecasts are averaged to the State level. Since the sample is self-weighting, the simple mean of the sample forecasts is an unbiased estimate of the State gross yield. Therefore, gross yield for a State is given as $\bar{G}$,

$$\bar{G} = \frac{1}{N_G}\sum_{i}^{N_G} G_i$$

where

$\bar{G}$ = State mean gross yield
N_G = number of samples with gross yield forecasts (estimates)
G_i = gross yield for sample i

The standard error of $\bar{G}$ is:

$$S_{\bar{G}} = \sqrt{\sum_{i}^{N_G} \frac{(G_i - \bar{G})^2}{N_G(N_G - 1)}}$$

Simple means are also appropriate for Stalks per Acre, Ears per Acre, and Harvest Loss. No weighting is required when calculating State level averages for these items:

State Average Stalks per Acre = Σ (Sample Field Stalks per Acre) / N_G

State Average Ears per Acre = Σ (Sample Field Ears per Acre) / N_G

State Average Harvest Loss = Σ (Sample Field Harvest Loss) / N_L

The State average grain weight per ear is calculated using a weighted mean. The weighting variable is the sample field Ears per Acre.

State Average Grain Wt per Ear = Σ (sample field grain weight * sample field ears per acre) / Σ (sample field ears per acre)

Harvest Loss

Harvest loss data are collected from every fourth sample. If less than 10 samples have current harvest loss data then harvest loss, L, is the 5-year average harvest loss, expressed as a percentage of gross yield. This 5-year average is used during the early months of the forecast season.

AVG. PERCENT LOSS = 100* (1/5) * Σ (Avg Loss in bu. / Avg Gross Yield in bu.)

The percentage loss is applied to the current year gross yield indication to calculate an indicated loss per acre.

INDICATED HARVEST LOSS = AVG. PERCENT LOSS * INDICATED GROSS YIELD

Later in the season, when 10 or more samples have harvest loss data, sample harvest loss calculations are made with the following expression:

$$L_i = \frac{\left(w_e + 2w_g\right)\left(1 - \left(\frac{m_i}{100}\right)\right) * 43{,}560}{(\bar{R}_i)(453.6)(60)(56)(.845)}$$

Where L_i is the forecasted harvest loss of sample i, w_e is the weight of ears between Row 1 and Row 3 for sample i, w_g is the weight of grain between Row 1 and Row 2 for sample i, $\bar{R}_i$ is the average row spacing, sample i, (sum of the two 4-row space measurements divided by 8), m_i is the moisture content of the shelled grain for sample i, and the following conversion factors:

453.6= conversion of grams to pounds
43,560 = square feet per acre
60 = row feet in 2 units
56 = pounds of corn in a bushel
.845 = converts to standard moisture (15.5) percent.

State average harvest loss is calculated using data from the individual samples as:

$$\bar{L} = \frac{1}{N_L}\sum_{i}^{N_L} L_i$$

Where L_i is the harvest loss in sample i and N_L is the number of harvest loss samples and standard error of harvest loss can be understood as:

$$S_{\bar{L}} = \sqrt{\sum_{i}^{N_L} \frac{(L_i - \bar{L})^2}{N_L(N_L - 1)}}$$

State Level Net Yield

Net yield for the State is computed by subtracting the estimated State level harvest loss from the mean of all sample level gross yield forecasts and estimates. Thus, estimated average net yield is:

$$\bar{Y} = \bar{G} - \bar{L}$$

Where $\bar{G}$ and $\bar{L}$ are State level gross yield and state level harvest loss

The standard error of the estimate is:

$$S_{\bar{Y}} = \sqrt{S_{\bar{G}}^2 + S_{\bar{L}}^2 - \frac{2}{N_G} COV(G, L)}$$

Where $S_{\bar{L}}$ and $S_{\bar{G}}$ were defined previously, and

$$COV(G,L) = \sum_{1}^{N_L} \frac{(G_i - \overline{G})(L_i - \overline{L})}{N_L - 1}$$

When less than 10 Form E's are completed, and historical average loss is used, the standard error is:

$$S_{\bar{Y}} = S_{\bar{G}}$$

Production for the State

Production (P) for the State is the product of estimated State level net yield and acres to be harvested for grain:

$$P = (A_{harv})\,(\bar{Y}),$$

with standard error:

$$S_P = \sqrt{(A_{harv}^2)\left(S_{\bar{Y}}^2\right) + (\bar{Y}^2)\left(S_{A_{harv}}^2\right) + \left(S_{\bar{Y}}^2\right)(S_{A_{harv}}^2)}$$

Gross Yield for Samples with Incomplete Data

Gross yield is forecasted from the current month's survey data. In some cases, current data are unavailable and data from a previous month may be used to forecast gross yield, or no gross yield is forecasted for the sample. The specific cases are discussed below.

Refusals

If the farmer refuses permission to enter the field, the sample is lost for the season. In this case the yield for this sample is left missing. Consequently, the refused sample contributes nothing to the State level average yield. Stated another way, the assumption is made that if the sample had not been a refusal, its gross yield would have been equal to the State's average gross yield.

Inaccessible Samples

Occasionally, some samples are inaccessible due to scheduling or field conditions. If data from a previous visit are available, the previous forecast is carried forward. Otherwise, the sample is excluded from gross yield calculations. The sample must still be intended for harvest as grain.

Early Farmer Harvest

If a sample is harvested by the farmer before current data can be collected, the previous month's predicted yield is brought forward.

Lost, Abandoned, Destroyed Samples

If a sample is lost, abandoned, destroyed, and so forth, no gross yield is computed for the sample. The sample contributes nothing to the sample-level yield indication.

Removing Bias from State Level Calculations

Forecasts of state or regional yields are inherently subject to differ from the final, administratively-determined, yield. The difference is due to weather and crop conditions yet to be encountered at the time the forecast is made, the difference between weather and crop conditions in the current year and the historic weather and crop conditions utilized to predict current yields, and systematic non-sampling errors which contribute to forecast error. NASS makes every attempt to minimize the impact of these three sources of error by means of administratively determine final values. These final, sometimes referred to as *official,* values are used as dependent variables to estimate an ordinary least squares equation with the averages calculated from objective yield samples as the independent variable. For example, each state will have an administratively-determined, *official* corn yield for all previous years. To forecast the *official* corn yield for the current year, NASS regresses the objectively determined State level average corn yields from previous years to the corresponding *official* corn yields. This process provides a historical context for weather, biases, and non-sampling errors.

The regression equation is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where y_i is the official state yield and x_i is the calculated State average net yield, β_0 and β_1 are historic least squares estimates, and ε_i is a random departure from the relationship.

Computational Examples

Sample Field Yield Examples

Suppose data have been collected for the following four samples. Calculations of gross yield will be demonstrated. The maturity categories are defined earlier in this chapter.

1. Sample 1

Maturity category	1, no ear shoots
Stalk count	79
8-row space width (ft.)	20.3
Historical 5-year avg grain weight per ear (lbs.)	0.29

Suppose regression models for samples with no ear shoots are:

Ears	=	8.6 + 0.86 (stalk count)
Grain Wt	=	Historical average grain weight

Then

Ears	=	8.6 + 0.86 (79) = 76.54 ears
Grain Wt	=	0.29 pounds

Then forecasted gross yield for sample 1 is:

Gross Yield = [(76.54)(0.29)(43560)] / [(56)(15)(20.3)/(2)] = 113.4 bu/acre

2. Sample 2

Maturity category	3, blister
Stalk count	81
8-row space width (ft)	20.3
Average kernel length (in.)	5.8

Ears Model = 7.2 + 0.87 (stalk count)
Grain Wt = 0.23 + 0.02 (kernel row length)
Therefore,

Ears	=	7.2 + 0.87 (81) = 77.67 ears
Ears Model	=	89 / [1.2 + 0.09 * (76 / 81)] = 69.29
Grain Wt	=	0.23 + 0.02 (5.8) = 0.346 lbs

and Gross Yield = [(73.65)(0.346)(43560)] / [(56)(15)(20.3)/(2)] = 130.2 bu/acre

3. Sample 3

Maturity category	5, dough
Ears with kernel formation	70
8-row space width (ft)	19.5
Average kernel length (in.)	5.2

Suppose regression models for samples in dough stage are:

Ears	=	count of ears with kernel formation
Grain Wt	=	0.10 + 0.04 (kernel row length)

Therefore,

Ears	=	70 ears
Grain Wt	=	0.10 + 0.04 (5.2) = 0.308 lbs.

and

Gross Yield	=	[(70)(0.308)(43560)] / [(56)(15)(19.5)/(2)] = 114.7 bu/acre

4. Sample 4

Maturity category	6, dent
Ears with kernel formation	50
8-row space width	20.3
Ears husked with grain	22
Field weight of husked ears (lbs.)	12.1
Wt. of ears in sealed bags (grams)	1042.2
Wt. of bags and rubber bands (grams)	45.2
Wt. of grain at moisture test(grams)	758.9
Moisture content (percent)	25.0

Suppose models for enumerator harvested samples are:

Ears	=	count of ears with kernel formation
Grain Wt	=	(field wt per ear)(fraction dry grain wt of field wt) / (0.845)

Therefore,

Ears	=	50 ears
Field weight per ear	=	12.1 / 22 = 0.55 lbs
Fraction dry weight of field weight	=	[(758.9)(1-(25.0/100))] / [(1042.2 - 45.2)]
	=	0.571
Grain Wt	=	[(0.55)(0.571)] / 0.845 = 0.372 lbs.

And,

Gross Yield = [(50)(0.372)(43560)] / [(56)(15)(20.3)/(2)] = 95.0 bu/acre

CHAPTER 6 SOYBEAN OBJECTIVE YIELD METHODS

This chapter presents basic sampling, data collection, and mathematical methods utilized to estimate U.S. soybean yield and production. The sampling review will emphasize NASS's unique *Area Frame* and its suitability for sample selections intended to measure U.S. soybean yields. Data collection will focus on the variety of data collected and how data collection changes as soybean crop maturity changes. The mathematical methods review will center on mathematical/statistical estimation of final soybean yields utilizing measurable plant characteristics observed at specific points during the growing season. The end results of careful sampling, data collection, and mathematical methods are numeric indications suitable to establishment, and/or revision, of U.S. soybean acreage, yield, and production.

Sample Design

Soybean Objective Yield (SOY) surveys are conducted in Arkansas, Illinois, Indiana, Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, Ohio, and South Dakota. On average, over the last three years, these eleven states produced more than 80 percent of the U.S. soybean crop. As described in Chapter 2, NASS partitions each of these states' land area into approximately one square mile pieces and calls the pieces segments. The resulting collection of segments is referred to as the "*Area Frame*". This painstaking segmentation of land area into uniquely identifiable segments makes possible the selection of a probability-based sample of U.S. soybean acres.

Probability based surveys require the explicit identification and separation of all elements contained in a population of interest. For yield surveys like SOY, the most important result of constructing a population of soybean acres is that *any* acre of soybeans, in a given state, can be assigned a known probability of selection. This result allows statistical samples to be extracted, population parameters to be estimated, hypotheses to be tested, and inferences to be made. NASS's *Area Frame* is the vehicle that transports us from a very large population of U.S. soybean acres to a representative sample of U.S. soybean acres. The sample is both statistically defensible and economically feasible to enumerate. The procedure for identifying every U.S. soybean acre is to first, select a subsample of *Area Frame* segments, and secondly, to account for all agricultural activity occurring within those segments. Each selected segment will have all its acres inspected by and accounted-for by professional enumerators with assistance from the owner/operator. In this manner, a new listing of soybean acres is constructed each year. This listing is the SOY sample population. No other purveyor of U.S. soybean estimates can construct such a sample population.

Accurate inferences from a probability-based survey require the construction of a sampling population and, in addition, a means to select a representative sample from that population. The statistical method used to select soybean acres to be included in SOY is probability proportional to size (PPS). In statistics, this method ensures a representative sample is selected when sample

elements vary in their size. A simple explanation should be given as to how sample elements in the soybean acres population can differ in size. During constructing of the soybean acres sample population, the explicit identification and separation of soybean acres, from all other acres, is made according to *fields*. In agriculture a field is one, continuous acreage of land devoted to the same use. Therefore, the sample elements in the soybean population are *fields*. Since *fields* of soybeans vary in acreage, the sample selection method of PPS ensures the probability of selection is adjusted for field size. For example, employing PPS sampling, a soybean field that is twice the size of its neighbor will also be twice as likely to be included in the sample. After the determination of which fields are included in the sample, the representative *acre* is selected with simple random sampling designed to give every acre in the selected field, an equal chance of selection. Since a selected *acre* is too large for complete enumeration, a selected *sample* will be enumerated and expanded up to the one-acre level.

In mid-July of each year, professional statisticians in the eleven SOY states train field enumerators to properly identify and prepare selected samples for data collection. Generally, enumerators will have many years of experience in SOY data collection and are well qualified to conduct their assignments. Practical field training is also available through relationships developed among individual enumerators and their supervisors. Overall, the preparation for data collection is rigorous and includes quality control processes that continue through the soybean growing season. In late July, field enumerators will begin visiting and enumerating samples and will continue personal visits at monthly intervals throughout the season until final harvest.

Each sample consists of two units (or plots) to be utilized in forecasting final net yield per acre. Each unit consists of two parallel 3.5-foot sections of row partitioned into a 3-foot section and a 6-inch section. During each visit, enumerators count or measure each required plant characteristic, the required characteristics being determined according to the units' maturity level. For example, at early maturities enumerators will count plants and measure row spacing. At later maturities, lateral branches, flowers, and pod weights are measured and recorded. After each data collection period, a forecast of each unit's final net yield is constructed. Essentially forecasts of soybean yields depend on two items: One, the current year's measurable plant characteristics and, two, the historic relationship between plant characteristics measured in the past, at the same level of maturity, and final plant characteristics that result in harvestable yield. Each item is needed to forecast a samples final net yield per acre. In cases where one or the other of the two required items is missing, a five-year average of the final plant characteristic is substituted. During some early forecasts there may be no plant characteristic that NASS measures that can be related the final plant characteristic. Pod weight is the prime example. In these cases, historic averages are used to forecast final net yield.

Data Collected

Field enumerators count and measure several items within or near the units. The size of the unit (square feet), the number of pods, pod weight, and harvest loss. The following lists the data items collected and what it is used to measure.

Data items used to measure the size of each unit:
- Distance between two rows (one row middle)
- Distance between five rows (four row middles)

Data items used to forecast or estimate the number of pods:
- Number of plants in each section of each row
- Number of main stem nodes in the 6-inch section
- Number of lateral branches in the 6-in section
- Number of dried flowers and pods in the 6-inch section
- Number of pods with beans in the 6-inch section

Data items used to forecast or estimate bean weight per pod:
- Weight of beans harvested by enumerator
- Moisture content of beans harvested

Data items used to estimate harvest loss:
- Distance between two rows (one row middle)
- Distance between five rows (four row middles)
- Weight of beans gleaned from harvest loss units
- Moisture content of beans gleaned

Maturity Categories

At each visit, the enumerator makes maturity assessments within the units and a maturity category is established for the sample. Forecast equations are derived using data collected during the previous 5-years for each maturity in each month. The maturity definitions used by the enumerators are:

Maturity 2,	Pods set, leaves still green, or earlier
Maturity 3,	Pods filled, leaves turning yellow
Maturity 4,	Pods turning color, leaves shedding
Maturity 5,	Pods brown, almost mature or mature

In analysis, these categories are further refined into 10 forecasting categories, based on the counts made by the enumerators. The 10 forecasting categories form a more homogeneous grouping and are defined as:

0 No plants present in either row of the 6-inch section
1 Field maturity 2, no pods with beans in 6-inch section and the ratio of total fruit to main stem nodes is less than .20.
2 Field maturity 2, no pods with beans in 6-inch section and the ratio of total fruit to main stem nodes is between 0.20 and 1.75 inclusive.
3 Field maturity 2, no pods with beans in 6-inch section and the ratio of total fruit to main stem nodes is greater than 1.75.
4 Field maturity 2, pods with beans are present in the 6-inch section and the ratio of pods with beans to total fruit is less than 0.05.
5 Field maturity 2 and the ratio of pods with beans to total fruit is at least 0.05 but less than 0.20.
6 Field maturity 2 and the ratio of pods with beans to total fruit is at least 0.20 but less than 0.65.
7 Field maturity 2 and the ratio of pods with beans to total fruit is at least 0.65 but at most 0.85.
8 Field maturity 2 and the ratio of pods with beans to total fruit is greater than 0.85, or Field maturity 3 (and plants present in the 6-inch section).
9 Field maturity 4 and plants present in the 6-inch section.
10 Field maturity 5, regardless of whether there are plants in the 6-inch section.

Analysis of Raw Data

In the following sections estimation at the sample level is discussed. All sample level regression equations are derived from the previous five years' survey data using multiple regression techniques. Certain influential data points (i.e., "outliers") are excluded from the dataset prior to deriving the coefficients. These influential data points are identified using a "deleted residual" analysis or the "Cook's D" statistic (Belsley, Kuh, and Welsch, [4]). There is usually very little change in the regression equations from year-to-year because roughly 80 percent of the data for each class were used in the analysis the previous year. Classes that do change significantly from one year to the next are usually those with very few observations. If a class has little data and a plausible forecast equation cannot be derived, the equation from the previous year is used.

Forecasting Gross Yield

Forecasting soybean yields is somewhat different, but follows similar patterns, to corn and cotton yield forecasting. The difference results from two features of soybeans: one, their typical row

space pattern and two, their fruit size and count. Soybean producers exhibit more variation with respect to row spacing than either corn or cotton producers. This variation is a direct result of soybeans flexibility to be used as either a single or double crop. Another unique feature of soybeans that alter yield forecasting patterns is fruit size and counts. Corn ears are large and most often singular to individual plants. Even cotton bolls are large compared to soybean pods and cotton's fruit counts are not unmanageable. On the other hand, individual soybeans pods are much smaller, more prolific, and therefore, more difficult to count. In addition, other plant characteristics for soybeans are almost unmanageable due to prolific counts, e.g., flowers. These two differences alter the methods NASS employs to count and measure soybeans. The most important change is a reduction in the size of the count area for soybean samples. Another change is the conversion of all soybean sample measures to a common denominator of 18 square feet.

The forecast for soybean yield per acre is expressed as follows:

$$\text{Gross yield per acre} = \frac{(\text{pods per 18 square feet}) * (\text{weight per pod}) * (43{,}560)}{(18)(453.6)(60)}$$

Where 43,560 is the number of square feet in an acre, 60 converts pound of beans to bushels, and 453.6 converts grams to pounds.

In the equation above, forecasts for gross yield per acre depends on the number of pods per 18 square feet and the weight per pod. Should the pods per 18 square feet in a sample plot be unobservable due to immaturity, then stochastic models are employed to "predict" the number of future pods per 18 square feet. The prediction is based on the historic relationship between past observable plant characteristics and the eventual outcome for pods. These historic models are stochastic because a range of potential pods per 18 square feet exists for each value of the observable plant characteristic. NASS gives careful consideration to plant characteristics that are proven performers in terms of prediction and forecasting. Table 1 illustrates the plant characteristics utilized for predicting gross yields at alternative time and maturity levels:

Table 1. Soybean Objective Yield Models

Sample Level Models

models based on previous 5 years

Forecasting Category		maturity 2							maturity 3,4	maturity 5	
		1 Fruit/ Node 0 to .2	2 Fruit/ Node .2 to 1.75	3 Fruit/ Node 1.75+	4 Pods/ Fruit 0 to .05	5 Pods/ Fruit .05 to .2	6 Pods/ Fruit .2 to .65	7 Pods/ Fruit .65 to .85	8 yellow	9 brown	10 enumerator harvested
Pods per Plant	Sep	Plants, Nodes	Plants, Laterals	Laterals, Fruit	Plants, Laterals	Laterals, Fruit	Laterals, Fruit, Pods	Pods	Pods	Pods	Pods
	Oct							Pods	Pods	Pods	
	Nov										
Weight per pod		5-year average									Lab Data

fruit = blooms + dried flowers + pods

To form a complete gross yield expression, a value is needed for pods per 18 square feet which is only observable at or just before harvest. Therefore, yield forecast made before final harvest require some method to forecast final pods per 18 square feet. NASS forecasts pods per 18 square feet with the following expression:

Pods per 18 square feet = Final Plants per 18 square feet * Final Pods per Plant

But 'Final Plants per 18 square feet' is also unobservable before harvest, as well as 'Final Pods per Plant'. NASS forecasts these two subcomponents in the following manners.

1. Forecasting Final Plant Count per 18 Square Feet

A simple linear (one variable) regression model is used to forecast the final number of plants per 18 square feet. The form of the model is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

The independent variable, x_i, is the current period, total plant count in the 3-foot and 6-inch sections of unit i and y_i is the final plant count from the same area. These counts are available from previous years' data collections and therefore are available to calculate parameters β_0 and β_1 by ordinary least squares. As might be expected, the R^2's from this model, in any state, is rarely below 90 percent.

If the forecasted number of plants exceeds the number obtained during the monthly visit, the forecast is replaced with the monthly visit value. A negative forecast is replaced with zero.

2. Forecasting Final Pods per Plant

The final number of pods per plant is forecasted using one or two variable regression models. The independent variables used to predict pods per plant depend upon the forecasting category of the unit and are shown in Table 1. There are five possible forecasting variables:

V1 = Plants per 18 square feet (the same variable used to forecast the final number of plants per 18 square feet)
V2 = Main stem nodes per plant
V3 = Lateral branches with blooms, dried flowers, or pods per plant
V4 = Blooms, dried flowers, and pods per plant
V5 = Pods with beans per plant

Thus, the general form of the model is:

$$y_i = \beta_0 + \beta_1 V_1 + \beta_2 V_2 + \beta_3 V_3 + \beta_4 V_4 + \beta_5 V_5 + \varepsilon_i$$

where three or four of the coefficients ($\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$) are zero. y is the final (at-harvest) number of pods per plant in the 6-inch section for sample *i*.

If a unit is classified as forecasting category 0, no counts are possible in the 6-inch sections so there are no forecasting variables. The average number of pods with beans per plant in all other forecasting categories (1-10) is substituted for units in category zero. In all States separate averages are computed for "wide" row units (row width at least 1.5 feet, broadcast, or blank) and narrow row units (row width less than 1.5 feet) for category zero substitutions.

Recall the expression for final gross yield given above was:

$$\text{Gross yield per acre} = \frac{(\text{pods per 18 square feet})*(\text{weight per pod})*(43{,}560)}{(18)(453.6)(60)}$$

Together, the two sub-component forecasted values for final plants per 18 square feet and final pods per plant may be used to forecast a final pods per 18 square feet. Now, if a final weight per pod is also known, the yield may be calculated. Unfortunately, the final weight per pod is unknown until just before, or at, harvest. As Table 1 shows, weight per pod is forecasted at a 5-year average over all forecasting categories, allowing for shifts in the five-year average according to whether the sample has wide or narrow row spacing. The average is carried until laboratory data make the weight per pod final.

State Average Forecasts and Estimates

The sample level gross yield forecasts are averaged to the State level. Since the sample is self-weighting, the simple mean of the sample forecasts is an unbiased estimate of the State gross yield. Therefore, gross yield for a State is given as G,

$$\bar{G} = \frac{1}{N_G} \sum_{i}^{N_G} G_i$$

where

$\bar{G}$ = State mean gross yield
N_G = number of samples with gross yield forecasts (estimates)
G_i = gross yield for sample *i*

And the standard error of $\bar{G}$ is:

$$S_{\bar{G}} = \sqrt{\sum_{i}^{N_G} \frac{(G_i - \bar{G})^2}{N_G(N_G - 1)}}$$

Simple means are also appropriate for Pods per 18 Square Feet, Plants per 18 Square Feet, Harvest Loss, etc. No weighting is required when calculating State level averages for these items:

Harvest Loss

For one quarter of the samples, an additional plot is laid out near each unit and gleaned after farmer harvest of the field. If less than 10 harvest loss samples have been completed for a State, a 5-year historical average (bu/acre) is the State-level estimate of harvest loss. When a sampling gleaning has been completed, harvest loss (bu/acre) is computed for each sample as follows:

$$L_i = \frac{(w_{loose/thres})\left(1 - \left(\frac{m_i}{100}\right)\right) * 43{,}560}{(3)(\bar{R}_i)(453.6)(60)(.875)}$$

Where $w_{loose/thres}$ is the weight of loose and threshed beans, m_i is moisture, and $\bar{R}_i$ is the 4-row spacing for unit 1 plus the 4-row spacing for unit 2, divided by 2. If a unit is broadcast, 6.0 is used for its 4-row space width.

These sample-level harvest loss estimates are averaged to the State level, with mean

$$\bar{L} = \frac{1}{N_L}\sum_{i}^{N_L} L_i$$

Where L_i = harvest loss in sample i, and N_L = number of harvest loss samples.

And the standard error is given by $S_{\bar{L}} = \sqrt{\sum_{i}^{N_L} \frac{(L_i - \bar{L})^2}{N_L(N_L - 1)}}$

State Level Net Yield

Net yield for the State is computed by subtracting the estimated State level harvest loss from the mean of all sample level gross yield forecasts and estimates. Thus, estimated average net yield is:

$$\bar{Y} = \bar{G} - \bar{L}$$

Where $\bar{G}$ and $\bar{L}$ are the State level gross yield and state level harvest loss respectively.

The standard error of the estimate is:

$$S_{\bar{Y}} = \sqrt{S_{\bar{G}}^2 + S_{\bar{L}}^2 - \frac{2}{N_G} COV(G,L)}$$

Where $S_{\bar{L}}$ was defined previously, and

$$S_{\bar{G}} = \sqrt{\sum_{i}^{N_G} \frac{(G_i - \bar{G})^2}{N_G(N_G - 1)}}$$

$$COV(G,L) = \sum_{1}^{N_L} \frac{(G_i - \overline{G})(L_i - \overline{L})}{N_L - 1}$$

When less than 10 Form E's are completed, and historical average loss is used, the standard error is:

$$S_{\bar{Y}} = S_{\bar{G}}$$

Production for the State

Production (P) for the State is the product of estimated State level net yield and acres to be harvested for grain:

$$P = (A_{HARV})(\bar{Y})$$

with standard error:

$$S_P = \sqrt{(A_{harv}^2)(S_{\bar{Y}}^2) + (\bar{Y}^2)(S_{A_{harv}}^2) + (S_{\bar{Y}}^2)(S_{A_{harv}}^2)}$$

Removing Bias from State Level Calculations

Forecasts of State level yields are inherently subject to differ from the final, and in most cases administratively-determined, yield. The difference is due to weather and crop conditions yet to be encountered at the time the forecast is produced, the difference between weather and crop conditions in the current year and the historic weather and crop conditions utilized to predict current yields, and also systematic non-sampling errors which contribute to forecast error. NASS makes every attempt to minimize the impact of these sources of error by means of administratively determine final values. These final, sometimes referred to as *official,* values are used as dependent variables and are related to corresponding State level averages of net yield to estimate historic linear regression equations. For example, each state will have an administratively-determined, *official* soybean yield and a State level soybean yield for all previous years. To forecast the *official* soybean yield for the current year, NASS takes the State level average soybean yields from previous years surveys and regresses them to the corresponding *official* soybean yields. This process provides a historical context for weather, biases, and non-sampling errors.

The following table shows the variety of independent and dependent variables for which historic biases are evaluated:

Dependent variable	Independent variable
Official Final Yield	September - average number of pods per acre October - December - average net yield per acre
Final Number of pods per acre	September - December - average number of pods per acre
Final weight per pod	September - average number of pods per acre October - December - average weight per pod
Final Harvest Loss	September - November - average of previous 5 years harvest loss December - Harvest Loss for current year

Laboratory Weight Measures

Bean weight per pod is calculated using the harvested data from the sample. The weights from both units are combined, so only one weight is calculated for the sample.

$$\left[\frac{W_C}{N_C}\right]\left[\frac{W_B}{W_{12}}\right]\left[\frac{1.0-(\text{moisture content }/100)}{0.875}\right]$$

where

W_c = weight of the pods and beans from Row 1 of the 3-foot section of Unit 1. If there are no plants in Row 1 of Unit 1, then Row 2 is used. If that is also blank, then the same process is applied to Unit 2.
N_c = the number of pods with beans from the row counted above.
W_{12} = weight of pods and beans from Row 1 of the 3-foot sections of Units 1 and 2.
W_B = weight of the threshed beans form Row 1 of the 3-foot sections of Units 1 and 2.
0.875 = conversion to 12.5 percent moisture (1.0 - .125).

Number of Pods with Beans per 18 Square Feet is computed for each unit from the harvested data:

Unit 1: $$\frac{(W_1)(N_c)(18)}{(W_c)(3)(4\text{-row space width})/(4)}$$

Unit 2: $$\frac{(W_2)(N_c)(18)}{(W_c)(3)(4\text{-row space width})/(4)}$$

Where

W_i = weight of pods and beans from Row 1 of the 3-foot section of Unit i (i = 1 or 2),
N_c and W_c were defined previously, and the expression "(3)(4-row space width)/(4)" is the area of the rectangular unit formed by Row 1 of the 3-foot section and its row middle. If the unit is broadcast, a 4-row space of 6.0 is used.

Example

Suppose Unit 1's pods are counted in the lab, and the following data are obtained:

W_c = 103.2 grams = W_1
N_c = 221
W_B = 134.8 grams
W_{12} = 236.4 grams
moisture content = 10.6 percent

4-row space width = 11.0 feet

Then, the estimated weight of beans per pod is:

$$\frac{(103.2)(134.8)\ (1-(10.6/100))}{(221)(236.4)\quad(0.875)} = 0.272 \text{ grams.}$$

The estimated number of pods per 18 square feet is:

$$\frac{(103.2)(221)(18)}{(103.2)(3)(11.0)/(4)} = 482.18 \text{ pods/18 square feet.}$$

Then the estimate of gross yield for the unit is:

$$\frac{(482.18)(0.272)(43560)}{(18)(453.6)(60)} = 11.66 \text{ bu/acre}$$

Gross Yield for Units with Incomplete Data

Gross yield is forecasted or estimated from the current month's survey data. In some cases, current data are unavailable and data from a previous month may be used to compute gross yield, or no gross yield may be computed for the unit. The different cases are discussed below.

Refusals

If the farmer refuses permission to enter the field, the sample is lost for the season. In this case, the yield for this sample is left missing. Consequently, the refused sample contributes nothing to the State-level average yield. Stated another way, the assumption is made that if the sample had not been a refusal, its gross yield would have been equal to the State's average gross yield.

Inaccessible Samples and Units

Occasionally, some or both units are inaccessible due to scheduling or field conditions. If data from a previous visit are available, the previous forecast is carried forward. Otherwise, the sample is excluded from gross yield calculations. The sample must still be intended for harvest as beans.

Early Farmer Harvest

If a previously laid out unit is harvested by the farmer before current data can be collected, the previous month's predicted yield is brought forward.

Lost, Abandoned, Destroyed Units

If a unit is lost, abandoned, destroyed, and so forth, no gross yield is computed for the unit. The unit contributes nothing to the sample-level yield indication.

Computational Examples

An example will now be given showing how gross yield per acre is forecasted for a sample. Assume that the following data were obtained for a sample.

Sample Data

	Unit 1	Unit 2
Field maturity	2	2
Four-row space measurement (ft.)	12.8	12.5
Plants in the 2 3-foot row sections	41	40
Plants in the 2 6-inch row sections	11	9
Nodes on the main stems of the plants	96	74
Lateral branches with blooms, dried flowers, or pods	5	2
Blooms, dried flowers, and pods	50	37
Pods with beans	0	0

Before gross yield is computed, a forecasting category is computed for each unit. In this example, both units would be category 2 (no pods with beans in the 6-inch section, fruit/nodes ratio between 0.20 and 1.75 inclusive).

To forecast plants per 18 square feet, the current number of plants is scaled to the standard 18 square feet:

(plants in the 3-foot and 6-inch sections)(18)

$$X = \frac{18}{(3.5)(4\text{-row space width})/(2)}$$

Where 18 is standard area, (3.5) is the length of row counted, and (4-row space width)/(2) is the width of a 2-row unit. If the unit is broadcast, the 4-row space width is 6 feet.

The current plant count per 18 square feet for each unit in the example is:

Unit 1:
$$X = \frac{(41+11)(18)}{(3.5)(12.8)/(2)} = 41.8$$

Unit 2:
$$X = \frac{(40+9)(18)}{(3.5)(12.5)/(2)} = 40.3$$

Suppose that the values for b_o and b_1 in the forecasting equation are 1.2 and 0.92, respectively. Then the forecasted number of plants per 18 square feet for each unit is:

Unit 1: $P_1 = 1.2+(0.92)(41.8)=39.656$

Unit 2: $P_2 = 1.2+(0.92)(40.3)=38.276$

To forecast pods with beans per plant, a two-variable regression model is used for forecasting category 2 (see previous table), containing the following variables:

V1 = current month's plant count expanded to 18 square feet (x)

V3 = lateral branches with blooms, dried flowers, or pods per plant for the 6-inch section

so, the model is:

$Y = b_o+b_1V1+b_3V3.$
Given, the following forecast equation:

$Y = 42.2 - (0.6)\ V1 + (4.8)V3,$

the forecast of pods per plant for each unit is:

Unit 1: $42.2 - (0.6)(41.8) + (4.8)(5/11) = 19.30$

Unit 2: 42.2 - (0.6)(40.3) + (4.8)(2/9) = 19.09

To forecast bean weight per pod, a 5-year historical average weight is used. Assume that the 5-year historical average weight is 0.437 grams for the wide row samples for this State. The expression for each unit's yield per acre is:

Unit 1:

$$Y_1 = \frac{(39.656)(19.30)(.437)(43560)}{(18)(453.6)(60)} = 29.74 \text{ bu/acre}$$

Unit 2:

$$Y_2 = \frac{(38.276)(19.09)(.437)(43560)}{(18)(453.6)(60)} = 28.39 \text{ bu/acre}$$

And the gross yield forecast for the sample is:

(29.74+28.39)/2 = 29.06 bu/acre

CHAPTER 7 COTTON OBJECTIVE YIELD METHODS

This chapter presents the procedures and formulae used to calculate cotton yield indications. The scope of the Cotton Objective Yield Survey, sample plots, and data collected are briefly described. More detail is given to the formulae that use the data to forecast and estimate yield.

Early in the growing season, some or all of the three components of net yield (number of bolls, average boll weight, and harvest loss) cannot be obtained directly and must be forecast. The procedures used to forecast these components are described in the following sections.

Sample Design

Cotton Objective Yield (CtOY) surveys are conducted in major cotton producing States: Arkansas, Georgia, Mississippi, and Texas. As described in Chapter 2, NASS partitions each of these states' land area into approximately one square mile pieces and calls the pieces segments. The resulting collection of segments is referred to as the "*Area Frame*". This painstaking segmentation of land area into uniquely identifiable segments makes possible the selection of a probability-based sample of U.S. cotton acres.

A probability-based survey requires the explicit identification and separation of all elements contained in the population of interest. NASS's *Area Frame* is the vehicle for explicitly identifying and separating each acre of cotton produced in the U.S. Every year, during late spring, all four CtOY states screen a statistically selected sample of *Area Frame* segments for all agricultural activity, including cotton acreage planted. Each acre in every selected segment will be visually inspected and/or accounted-for by the owner/operator with respect to agricultural activity. In other words, every year a new listing of every cotton acre is constructed. This listing is the CtOY sample population. This listing is an important step towards employing probability theory to randomly select a sample of cotton acres from which inferences will be made about all cotton acres. The most important result of constructing a sample population is that ***any*** acre of cotton, planted in each state, has a chance to be selected for the CtOY survey. No other purveyor of U.S. cotton estimates can make an equivalent statement.

Accurate inferences from a probability-based survey require the construction of a sampling population and, in addition, a means to select a representative sample from that population. The statistical method used to select cotton acres to be included in CtOY is probability proportional to size (PPS). In statistics, this method ensures a representative sample is selected when sample elements vary in their size. During constructing of the cotton acres sample population, the explicit identification and separation of cotton acres, from other acres, is made according to *fields*. In agriculture a field is one, continuous acreage of land devoted to the same use. Since *fields* of cotton are the unit of selection for CtOY samples and because cotton *fields* vary in acreage, probability proportional to size is ideally suited to selecting a sample that perfectly reflects the population characteristics of cotton. For example, employing PPS sampling implies a cotton field accounting for twice the percentage of total state cotton acres relative to a

neighboring cotton field, will also carry twice the probability of selection as the neighbor. The representative *acre* selected from a sampled *field* is finally accomplished with simple random sampling. Since a selected *acre* is too large for complete enumeration, a selected *sample* will be enumerated and expanded up to the one-acre level.

In mid-July of each year, professional statisticians in the four CtOY states train field enumerators to properly identify and prepare selected samples for data collection. Generally, enumerators will have many years of experience in CtOY data collection and are well qualified to conduct their assignments. Practical field training is also available through relationships developed among individual enumerators and their supervisors. Overall, the preparation for data collection is rigorous and includes quality control processes that continue through the cotton growing season. Data are collected from each sample at monthly intervals starting in late July and continuing through December or until the sample has been harvested. Each month during the Objective Yield Survey, data collected from the sample fields are used to produce indications of planted acres (September only), acres for harvest, and yield.

A sample consists of two independently located units (or plots), each of which consists of two parallel 10-foot sections of row. An additional 3-foot section is appended to one row of each unit. This extra section is used when making detailed fruit counts. Field enumerators use a random number of rows along the edge of the field and a random number of paces into the field to locate each unit. At each visit, enumerators count all fruit and fruiting positions. Any mature bolls found in the 10-foot sections of the sample plots are picked and sent to a NASS lab where boll weight is determined. The count of bolls picked, and the weight of these bolls are accumulated through the season. Just before farmer harvest, all remaining open bolls are picked and weighed to establish gross yield. The yield is measured as pounds of lint per acre at 5 percent moisture. Harvest loss is measured in separate units located near the monthly yield plots.

Data Collected

Field enumerators count and measure several items within or near the units. Data items are used to measure the size of the unit, number of bolls, weight per boll, and harvest loss. The following lists the data items collected and objective of these measurements:

Data items used to measure the size of each unit:

- Distance between two rows (one row middle)
- Distance between five rows (four row middles)

Data items used to forecast or estimate the number of bolls:

- Number of plants in each row (all sections)
- Number of squares (3-foot sections)
- Number of small bolls and blooms (3-foot sections)
- Number of large unopen bolls (10-foot sections)
- Number of open bolls (10-foot sections)

Data items used to estimate weight per boll:
- Weight of lint harvested by enumerators
- Weight of lint dried to zero moisture

Data items used to estimate harvest loss:
- Distance between two rows (one row middle)
- Distance between five rows (four row middles)
- Number of unopen bolls left in the field
- Weight of lint gleaned from harvest loss units
- Weight of dried lint

Maturity Categories

To forecast each sample's yield per acre, regression models are developed by maturity category for each survey month. For cotton, the maturity categories are defined by the raw counts obtained in the sample. These categories are:

	In 10-foot sections	In 3-foot sections
1	No fruit present	No fruit present
2	No fruit present	Squares only
3	$0 \leq$ RATIO < 0.5	Blooms or Bolls
4	$0.5 \leq$ RATIO < 2.0	----
5	$2.0 \leq$ RATIO	----
6	Sample field has been harvested or harvest immanent.	

RATIO is the ratio of large bolls counted to plants counted in the 10-foot sections of the sample. Large bolls include burrs, open bolls, partially open bolls, and large unopened bolls.

Sample Level Yield Forecasts

Forecasting the Number of Large Bolls

The expected number of large bolls for each sample is forecast using a regression model:

$$\hat{Y}_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

where:

$\hat{Y}_i$ = forecasted number of large bolls in *i*th unit
X_1 = observed number of burrs, open bolls, partially open bolls, and large unopened bolls (40-foot equivalent) in *i*th unit
X_2 = observed number of small bolls and blooms (40-foot equivalent) in *i*th unit
X_3 = observed number of squares (40-foot equivalent) in ith unit
$\beta_0 - \beta_3$ = least squares regression coefficients

Small bolls are defined as boll less than one inch in diameter. Enumerators use a gauge with a one-inch hole to determine whether a boll is small or a large unopened boll. A square is an observable fruiting position that has not reached the bloom stage.

Forecast equations for each model are derived for each maturity category for each month for each district for each State. Not all possible independent variables are used in each model. For instance, for maturity category one only the intercept is fit. For later maturities and or months, squares and small bolls are excluded from the models. Data from the previous 5 years are used to estimate the regression coefficients. If a unique set of coefficients cannot be determined for a given class (due to insufficient data), the previous month's coefficients are used.

The actual count of large bolls is used for any sample in maturity category six in any month, and for all samples in December and later months. All samples in maturity category one use a 5-year historical average.

Analysis of Raw Data

The regression equations are derived from the previous 5 years' survey data using multiple regression techniques. Certain influential data points (i.e., "outliers") are excluded from the dataset prior to deriving the coefficients. These influential data points are identified using a "deleted residual" analysis or the "Cook's D" statistic (Belsley, Kuh, and Welsch, [4]). There is usually very little change in the regression equations from year-to-year because roughly 80 percent of the data for each class were used in the analysis the previous year. Classes that do change significantly from one year to the next are usually those with very few observations. If a class has little data and a plausible forecast equation cannot be derived, the equation from the previous year is used.

Forecasting Boll Weight

Early in the season, until 20 percent of the projected number of large bolls have been picked and weighed by the enumerator, a 5-year historical average is used. When 20 to 85 percent of the projected number of large bolls has been picked, one model is used to forecast boll weight for all maturity categories in a district in a State:

$$\widehat{W_{sd}} = w_{sd}(\beta_0 + \beta_1 X_{sd})$$

where:

$\widehat{W_{sd}}$ = forecast boll weight for the s^{th} State and d^{th} district
w_{sd} = observed boll weight
X_{sd} = ratio of bolls picked and weighed to large bolls forecasted
$\beta_0 - \beta_1$ = regression coefficients

When more than 85 percent of the projected number of large bolls has been picked and weighed by the enumerator, actual boll weight is used.

The following table shows the independent and dependent variables for the State level indication models used during the 1996 growing season.

Dependent variable	Independent variable
Official Final Yield	Average estimated net yield per acre over all samples
Final Number of Bolls	August - Average small bolls and blooms per acre over all samples September - Average small bolls and blooms plus cumulative large bolls per acre over all samples October - January - Average cumulative large bolls per acre
Final boll weight	August - September - Weight derived from average estimated final gross yield and average estimated final large bolls per acre October - January - average cumulative net weight per boll
Final Harvest Loss	August - November - average of previous 5 years harvest loss December - January - OY B Harvest Loss for current year

Cotton Objective Yield Models
Sample Level Models
models based on previous 5 years

Forecast Category		1 no fruit present	2 squares present	3 ratio 0 to 0.5	4 ratio 0.5 to 2.0	5 ratio 2.0+	6 harvested or soon to be harvested
Number of Bolls	August	5-year average	squares	cumulative large bolls small bolls & blooms squares			cumulative large bolls
	September	5-year average	squares	cumulative large bolls small bolls & blooms squares			
	October			cumulative large bolls small bolls & blooms			
	November			cumulative large bolls			
	December			cumulative large bolls			
Weight per boll	<20% picked	5-year average					
	20-85% picked	cumulative net weight x smoothing parameter					
	>85% picked	cumulative net weight					

ratio = cumulative large bolls / plants in 10-foot units
large bolls = burrs + large opened bolls + large partially opened bolls + large unopened bolls
smoothing parameter = value <1 that approaches 1 as percent picked approaches 85 percent

Forecasting Directly to State Level

The discussion in the previous sections centers on processing data at the sample level. Modeling and yield calculations are done at the sample level and averaging is done as the last step. Additionally, averages of the raw counts and component forecasts can be computed for supporting analysis.

A second approach to forecasting State yield, using the same data, can be applied by doing the averaging first and the modeling last. Averages per acre at the State level can be calculated for each of the count variables (plants, squares, small bolls and blooms, large unopen bolls, and open bolls. Average weight per boll can also be calculated, weighting the average weight per boll in each sample by the number of bolls in that sample. This process creates State level independent variables and leads to State and regional level models. The State and regional level independent variables can be regressed to final official yield, final bolls per acre, and final weight per boll. The distinction is State and regional averages are used as independent variables in regression models that predict State and regional level final yields, bolls per acre, and weight per boll. In these models, one year and month represents one observation, so instead of partitioning thousands of sample level points into forecasting categories, we have one data point per month per year. A 15-year dataset is used for these models. The models are simple one variable regression models and are called the State level models, referring to the fact that they are State and regional level models, not sample level models as described in the previous sections.

Gross Yield

The estimate of final gross yield is computed by multiplying the forecasted number of large bolls at harvest by the forecasted average weight per boll, expanding to a per acre basis, and converting to a standard unit. The standard unit for cotton is pounds of lint at 5 percent moisture. Production is reported in 480-pound bales.

The formula for computing gross yield is:

$$\widehat{G_i} = \frac{(2.401)ZB_iW_i}{R_i}$$

where

$\widehat{G_i}$ = estimated gross yield (lbs. of lint per acre) for sample *i*
Z = lint/seed ratio (3-year average)

B_i = number of large bolls at harvest in 40 feet for sample *i*
W_i = average boll weight for sample *i* (grams at 5 percent moisture, gin equivalent)
R_i = average row spacing for sample *i*
2.401 = 43,560 / (40 * 453.59) = which converts grams of seed cotton per 40 feet of row

to pounds of seed cotton per acre.

The Objective Yield samples are selected in such a way that each acre has equal probability of selection within districts. Therefore, the average of the sample level yields across all samples in a district provides a forecast of mean gross yield per acre for the district.

Mean Gross Yield for State

The sample level gross yield forecasts (estimates) are averaged to the State level. Since the sample is self-weighting, the simple mean of the sample forecasts (estimates) is an unbiased estimate of the State gross yield. Therefore,

$$\bar{G} = \frac{1}{N_G} \sum_{i}^{N_G} G_i$$

where

$\bar{G}$ = State mean gross yield
N_G = number of samples with gross yield forecasts (estimates)
G_i = gross yield for sample i

The standard error of the estimate is:

$$S_{\bar{G}} = \sqrt{\sum_{i}^{N_G} \frac{(G_i - \bar{G})^2}{N_G(N_G - 1)}}$$

Gross Yield for Units with Incomplete Data

Gross yield is forecasted or estimated from the current month's survey data. In some cases, current data are unavailable and data from a previous month may be used to compute gross yield, or no gross yield may be computed for the unit. The different cases are discussed below.

Refusals

If the farmer refuses permission to enter the field, the sample is lost for the season. In this case the yield for this sample is left missing. Consequently, the refused sample contributes nothing to the State-level average yield. Stated another way, the assumption is made that if the sample had not been a refusal, its gross yield would have been equal to the State's average gross yield.

Inaccessible Samples and Units

Occasionally, some or both units are inaccessible due to scheduling or field conditions. If data from a previous visit are available, the previous forecast is carried forward. Otherwise, the sample is excluded from gross yield calculations. The sample must still be intended for harvest as cotton.

Early Farmer Harvest

If a previously laid out unit is harvested by the farmer before current data can be collected, the previous month's predicted yield is brought forward.

Lost, Abandoned, Destroyed Units

If a unit is lost, abandoned, destroyed, and so forth, no gross yield is computed for the unit. The unit contributes nothing to the sample-level yield indication.

Harvest Loss

The harvest loss is computed from gleanings obtained from one quarter of the samples. The sample level harvest loss is found by determining the total weight of seed cotton gleaned, expanding to a "per acre" basis, and converting to standard units.

The formula for harvest loss is:

$$L_i = \frac{(2.401)W_i Z}{R_i}$$

where

L_i = harvest loss (lbs. of lint per acre) for the *i*th sample
W_i = weight of cotton left in units for sample *i* which is computed as: (partially opened and large unopened bolls left in the units) * (average net weight per boll) + (weight of cotton gleaned adjusted to 5 percent moisture)
Z = lint/seed ratio (3-year average)
R_i = row space measurement for sample *i*
2.401 = conversion factor (defined above)

For each month, if fewer than 10 harvest loss samples have been completed within a district, a 5-year average harvest loss is used as an estimate.

These sample-level harvest loss estimates are averaged to the State level, with mean:

$$\bar{L} = \frac{1}{N_L}\sum_{i}^{N_L} L_i$$

where

$\bar{L}$ = State level mean harvest loss
L_i = harvest loss in sample *i*
N_L = number of samples with Form E data

The standard error of the estimate is:

$$S_{\bar{L}} = \sqrt{\sum_{i}^{N_L} \frac{(L_i - \bar{L})^2}{N_L(N_L - 1)}}$$

Net Yield for the State

Net yield for the State is computed by subtracting the estimated State-level harvest loss from the mean of all sample-level gross yield forecasts and estimates. Thus, estimated average net yield is:

$$\hat{Y} = \bar{G} - \bar{L}$$

where

$\hat{Y}$= estimated average State net yield
$\bar{G}$ = State mean gross yield
$\bar{L}$ = State level mean harvest loss

The standard error of the estimate is:

$$S_{\hat{Y}} = \sqrt{S_{\bar{G}}^2 + S_{\bar{L}}^2 - \frac{2}{N_G} COV(G, L)}$$

where

$S_{\bar{G}}$ and $S_{\bar{L}}$ were previously defined, and

$$COV(G,L) = \sum_{i}^{N_L} \frac{(G_i - \bar{G})(L_i - \bar{L})}{N_L - 1}$$

When less than 10 gleanings are completed, and historical average loss is used, the standard error is:

$$S_{\bar{Y}} = S_{\bar{G}}$$

Production for the State

Production for the State is the product of estimated State-level net yield and acres harvested:

$$P = \hat{Y}A$$

where

P = State production
$\hat{Y}$= State level net yield
A = acres harvested

with standard error:

$$S_P = \sqrt{A^2 S_{\hat{Y}}^2 + \hat{Y}^2 S_A^2 + S_{\hat{Y}}^2 S_A^2}$$

Removing Bias from State Level Calculations

Forecasts of State level yields are inherently subject to differ from the final, administratively-determined, yield. The difference is due to the three possible reasons: weather and crop conditions yet to be encountered at the time the forecast is produced, the difference between weather and crop conditions in the current year and the historic weather and crop conditions utilized to predict current yields, or systematic non-sampling errors which contribute to forecast error. NASS makes every attempt to minimize the impact of these three sources of error by means of administratively determine final values. These final, sometimes referred to as *official,* values are used as dependent variables to estimate an ordinary least squares equation with the averages calculated from objective yield samples as the independent variable. For example, each state will have an administratively-determined, *official* cotton yield for all previous years. To forecast the *official* cotton yield for the current year, NASS regresses the objectively determined State level average cotton yields from previous years to the corresponding *official* cotton yields. This process provides a historical context for weather, biases, and non-sampling errors.

The regression equation is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where

y_i = Official state yield
x_i = calculated State average net yield
β_0 and β_1 = ordinary least squares estimates
ε_i=random departure from the relationship

Computational Yield Example

Yield (computed for a single sample)

September 1 Data

8-row space measurement	25.8

Counts Within 10-foot Units

Number of plants (4 rows)	87
Number of burrs (2 units)	113
Total open bolls (4 rows)	130
Weight of seed cotton picked (2 units)	650
Number of partially open bolls (4 rows)	48
Number of large unopened bolls (4 rows)	121

3-foot Tag Section Beyond Unit 1

Number of plants	11
Number of burrs and open bolls	33
Number of large unopened bolls	14
Number of small bolls and blooms	4
Number of squares	2

3-foot Count Section Beyond Unit 2

Number of plants	8
Number of burrs and open bolls	27
Number of large unopened bolls	11
Number of small bolls and blooms	6
Number of squares	1

Current Month Lab Form

Weight of seed cotton before drying	56
Weight of seed cotton after drying	52

Previous Months= Data Brought Forward

Accumulated burrs within unit	20
Accumulated bolls picked within unit	50
Accumulated adjusted weight seed cotton	257

Maturity Category Determination

$$ratio = \frac{(B_c + B_a) + (T_c + X_a) + V_c + U_c}{P_c}$$

where

B_c = current burrs within 10-foot units
B_a = accumulated burrs within 10-foot units
T_c = current total open bolls within 10-foot units
X_a = accumulated bolls picked within unit
V_c = current partially open bolls in 10 foot-units
U_c = current large unopened bolls within 10-foot units
P_c = current number of plants within 10-foot units

Using the example data $ratio = \frac{(113+20)+(130+50)+48+121}{87} = 5.54$. The ratio is greater than 2.0, thus the maturity category is 5.

Forecast Number of Large Bolls

Multiple Regression Model

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Where

$\hat{Y}$ = forecasted number of large bolls
X_1 = number of burrs, open bolls, partially open bolls, and large unopened bolls (40-foot equivalent)
X_2 = number of small bolls and blooms (40-foot equivalent)

X_3 = number of squares (40-foot equivalent)
$\beta_0 - \beta_3$ = least squares regression coefficients derived from the previous 5 years of sample level data

The multiple regression model uses 40-foot equivalents for the number of bolls. Each boll measurement must first be converted to this standard.

Burrs and open bolls, partially open bolls, and large unopened bolls are counted in a total of 46 feet of row (four 10-foot units and two 3-foot units).

$$X_1 = \frac{40}{46}[(B_c + B_a) + (T_c + X_a) + V_c + U_c + (C_1 + U_1) + (C_2 + U_2)]$$

where
X_1 = number of burrs, open bolls, partially open bolls, and large unopened bolls (40-foot equivalent)
B_c = current burrs within 10-foot units
B_a = accumulated burrs within 10-foot units
T_c = current total open bolls within 10-foot units
X_a = accumulated bolls picked within unit
V_c = current partially open bolls in 10 foot-units
U_c = current large unopened bolls within 10-foot units
C_1 = burrs and open bolls within 3-foot Unit 1
U_1 = large unopened bolls within 3-foot Unit 1
C_2 = burrs and open bolls within 3-foot Unit 2
U_2 = large unopened bolls within 3-foot Unit 2

Using the example data:
$$X_1 = \lfloor(113 + 20) + (130 + 50) + 48 + 121 + (33 + 14) + (27 + 11)\rfloor = 493.043$$

Small bolls and blooms are counted in six feet of row (both 3-foot units)

$$X_2 = \frac{40}{6}(D_1 + D_2)$$

where
X_2 = number of small bolls and blooms (40-foot equivalent
D_1 = small bolls and blooms within 3-foot Unit 1
D_2 = small bolls and blooms within 3-foot Unit 2

Using the example data: $X_2 = \frac{40}{6}(4 + 6) = 66.667$

Squares are counted in six feet of row (both 3-foot units)

$$X_3 = \frac{40}{6}(Q_1 + Q_2)$$

where

X_3 = number of squares (40-foot equivalent)
Q_1 = squares within 3-foot Unit 1
Q_2 = squares within 3-foot Unit 2

Using example data: $X_3 = \frac{40}{6}(2 + 1) = 20.000$

Least squares regression coefficients ($\beta_0 - \beta_3$) are derived from the previous 5 years of sample level data, for this example let:

β_0 = 14
β_1 = 0.933
β_2 = 0.300
β_3 = 0.110

The estimate of number of bolls using the regression model for this sample is:

$$\hat{Y} = 14 + (0.933 * 493.043) + (0.300 * 66.667) + (0.110 * 20.000) = 496.209$$

Forecast Boll Weight

$$\widehat{W} = w(\beta_0 + \beta_1 X)$$

where:

$\widehat{W}$ = forecast boll weight
w = observed boll weight at 5% moisture (gin equivalent)
X = ratio of bolls picked and weighed to large bolls forecasted
$\beta_0 - \beta_1$ = regression coefficients

Determine the observed boll weight at 5% moisture:

$$w = \frac{\left[r\left(\frac{j_a}{j_b}\right)(1.0526) + K_a\right]}{(X_a + T_c)}$$

where

w = observed boll weights at 5% moisture (gin equivalent)
r = weight of seed cotton picked in 10-foot section
j_a = weight of seed cotton after drying
j_b = weight of seed cotton before drying
1.0526 = conversion factor to 5% moisture (gin equivalent)
K_a = accumulated adjusted weight of seed cotton
X_a = accumulated bolls picked within unit
T_c = current total open bolls within 10-foot units

Using the example data:

$$w = \frac{\left[650\left(\frac{52}{56}\right)(1.0526) + 257\right]}{(50 + 130)} = 4.957$$

Determine the ratio of picked to forecasted large bolls:

$$X = \frac{(T_c + X_a)}{\hat{Y}}$$

where

X = ratio of bolls picked and weighed to large bolls forecasted
T_c = current total open bolls within 10-foot units
X_a = accumulated bolls picked within unit
$\hat{Y}$ = forecasted number of large bolls

Using the example data:

$$X = \frac{(130 + 50)}{496.209} = 0.363$$

For this example, let:

β_0 = 0.882
β_1 = 0.131

The forecasted boll weight using the calculated inputs:

$$\widehat{W} = 4.957(0.882 + 0.131 * 0.363) = 4.608\ grams\ per\ boll$$

Forecast Gross Yield per Acre

The estimated gross yield for this example sample is:

$$\hat{G} = \frac{(2.401)ZBW}{R}$$

where

$\hat{G}$ = estimated gross yield (lbs. of lint per acre)
Z = lint/seed ratio (3-year average)
B = number of large bolls at harvest in 40 feet
W = average boll weight (grams at 5 percent moisture, gin equivalent)
R = average row spacing
2.401 = 43,560 / (40 * 453.59)
which converts grams of seed cotton per 40 feet of row to pounds of seed cotton per acre.

For this example, let:

Z = 0.368

$$\hat{G} = \frac{(2.401) * 0.368 * 496.209 * 4.608}{(25.8/8)} = 626.45 \; pounds \; of \; lint \; per \; acre$$

CHAPTER 8 WHEAT OBJECTIVE YIELD METHODS

This chapter presents basic sampling, data collection, and mathematical methods utilized to estimate U.S. winter wheat yield and production. The sampling review will emphasize the technical differences in winter wheat sampling relative to techniques employed for corn, soybeans, or cotton. Data collection will focus on the variety of data collected and how data collection changes as the wheat crop matures. The mathematical methods review will center on mathematical/statistical estimation of final winter wheat yields utilizing measurable plant characteristics observed at specific points during the growing season. The end results of careful sampling, data collection, and mathematical methods are numeric indications suitable to establishment, and/or revision, of U.S. wheat acreage, yield, and production.

Sample Design

Wheat Objective Yield (WOY) surveys are conducted in the 10 major winter wheat producing States: Colorado, Illinois, Kansas, Missouri, Montana, Nebraska, Ohio, Oklahoma, Texas, and Washington. These ten states, on average over the last three years, have accounted for more than 65 percent of U.S. production. A sample of wheat producers to be included in WOY is selected each year from among all producers with positive wheat acreage reported during the March Agricultural survey. Thus, the major sampling difference in WOY compared to other crops is that producers are the sample element for WOY, whereas fields are the sample element for corn, cotton, and soybeans. The difference is necessary since *Area Frame* records are unavailable for the current years' winter wheat plantings. In the past, NASS sampled and enumerated segments from the *Area Frame* during late November to early December specifically for winter wheat plantings but discontinued the practice for budgetary reasons. A reasonable alternative to *Area Frame* records for winter wheat yield sampling are the *Multiple Frame* records collected during the March Agricultural survey. The *Multiple Frame* combines an exhaustive listing of agricultural producers, called the *List Frame*, with producers identified by the *Area Frame* in a manner that producer duplication is minimized, and producer omission is satisfactorily resolved. The resulting *Multiple Frame* is a satisfactory sampling frame for winter wheat yield forecasting.

Accurate inferences from a probability-based survey require the construction of a sampling population and, in addition, a means to select a representative sample from that population. The statistical method used to select wheat acres to be included in WOY is probability proportional to size (PPS). In statistics, this method ensures a representative sample is selected when sample elements vary in their size. The wheat acres sample population is constructed by explicitly identifying and separating *producers* with winter wheat plantings and noting their whole farm wheat plantings. Based on this producer record, PPS sample selection assigns higher probabilities to producers with a larger percentage of the state's total wheat plantings. Recall the sample design for WOY selects the *producer* first whereas, for corn, soybeans, and cotton the *field* is selected first. Then for each selected producer, a listing of wheat fields is constructed from which the representative *field* is selected and finally the representative *acre* is selected by

simple random sampling. Since a selected *acre* is too large for complete enumeration, a selected *sample* will be enumerated and expanded up to the one-acre level.

In mid-April of each year, professional statisticians in the ten WOY states train field enumerators to properly identify and prepare selected samples for data collection. Generally, enumerators will have many years of experience in WOY data collection and are well qualified to conduct their assignments. Practical field training is also available through relationships developed among individual enumerators and their supervisors. Overall, the preparation for data collection is rigorous and includes quality control processes that continue through the wheat growing season. In late April field enumerators will begin visiting and enumerating samples and will continue personal visits at monthly intervals throughout the season until final harvest.

Each sample consists of two units (or plots) to be utilized in forecasting final net yield per acre. Each unit consists of three parallel 21.6-inch sections of row. During each visit, enumerators count or measure each required plant characteristic, the required characteristics being determined according to the units' maturity level. For example, at early maturities enumerators will count plants and measure row spacing. At later maturities, head weights are measured and recorded. After each data collection period, a forecast of each unit's final net yield is constructed. Essentially forecasts of wheat yields depend on two items: 1. the current year's measurable plant characteristics and, 2. the historic relationship between plant characteristics measured in previous years, at the same level of maturity, and final net yield. Each item is needed to forecast a samples final net yield per acre. In cases where the historic relationship is missing, a five-year average of the final plant characteristic is substituted. During some early forecasts there may be no plant characteristic that NASS measures that can be related to the final net yield. Head weight is the prime example. In these cases, historic averages are used to forecast final net yield.

Data Collected

Field enumerators count and measure several items within or near the units. Data items are used to measure the size of the unit, number of heads, weight per head, and harvest loss. The following lists the data items collected and objective of these measurements.

Data items used to measure the size of each unit:
- Distance between two rows (one row middle)
- Distance between five rows (four row middles)

Data items used to forecast or estimate the number of heads:
- Number of stalks in each row
- Number of late boot heads in each row
- Number of emerged heads in each row

Data items used to forecast or estimate grain weight per head:

- Number of fertile spikelets on 10 heads
- Number of grains on 10 heads
- Weight of mature heads (before threshing) and weight of late boot heads
- Weight of grain threshed from mature heads
- Moisture content of the threshed grain

Data items used to estimate harvest loss:

- Distance between two rows (one row middle)
- Distance between five rows (four row middles)
- Grain weight of heads between Row 1 and Row 4
- Grain weight of loose kernels between Row 1 and Row 4.

Maturity Categories

At each visit, the enumerator makes maturity assessments within the units and a maturity category is established for the sample. Forecast equations are estimated using data collected during the previous five years, by month, and by maturity classification. The maturity definitions used by the enumerators are:

Maturity	Definition
1 - Pre-Flag	There is no swelling in the stalks and no flag leaf is present.
2 - Flag or early boot	A flag leaf is present, and the collar of the flag leaf has emerged above the top foliage leaf. The enclosed head is located below the collar of the top foliage leaf.
3 - Late boot or Flower	The wheat is in the late boot stage from the point where the swelling has occurred above the top foliage leaf until the head has emerged and will show a water clear liquid turning milky white.
4 - Milk	The kernels are soft, moist, and filled with a milky liquid.
5 - Soft dough	The contents of the kernels are soft and can be kneaded like dough.
6 - Hard dough	The grain is firm and can be dented with the thumbnail, but not easily crushed.
7 - Ripe	The grain is hard and breaks into fragments when crushed.

Forecasting and Estimating Number of Heads and Grain Weight per Head for Sample Fields

Wheat is no exception to the general rule of objective yield surveying, i.e., successfully forecasting fruit weight is more difficult than forecasting fruit count. Therefore, it should be no surprise that over the course of time and program development, one model has been deemed sufficient for predicting number of heads whereas two are employed to forecast final weight. The regression equations for all three models are developed at the sample level by relating counts and measurements of plant characteristics made during the growing season to actual counts, measurements, or weights made at harvest time. For example, the April stalk count of a sample may be easily related to the number of heads recorded just before farmer harvest in July. Many relationships like this are developed for forecasting and each relationship is estimated from the most recent five years of data.

The major early season independent variable used to forecast the final number of heads (used for pre-flag and flag or early boot maturities) is the observed stalk count. At this stage of development there are very few observable plant characteristics that are associated with final weight per head. Consequently, to forecast a yield, it is necessary to rely on the historical head weight (5-year average) as the forecast of end-of-season head weight.

As the crop develops toward mid-season, more plant characteristics appear that can be accurately defined, measured, and related to final yield. It is in this period of early head development (late boot or flower) that the plant enters a transition stage. The plant shifts from development of vegetative growth to grain development. At this time, it is possible to accurately forecast final head numbers. The maximum fruit load has been or is nearly set. The number of emerged and late boot heads are used to forecast the final number of heads. It is also possible to make the first forecast of head weight based on observable and measurable plant characteristics. Wheat heads have spikelets which are clearly distinguishable when the stalk reaches the boot stage. Within most of these spikelets one to three grains will form. Therefore, using the number of spikelets in a regression equation provides the first current indication of the end of season head weight.

When the wheat plant reaches the late stages of development (milk and soft dough), the physiological processes of the wheat plant are directed totally toward kernel development. Head development has also reached the point where kernels are filling and can be accurately identified and counted. The observed number of grains per head and the observed clip unit green weight per head of emerged and late boot heads are weighted together by their R-square values and used at this stage for predicting the final head weight. At this time, forecasts become more precise since the effect of unfavorable weather or environmental conditions on final biological yield is reduced considerably.

When a field reaches the hard dough or ripe stage (maturity codes 6 and 7), the sample units are harvested. Number of heads, average grain weight per head, and the moisture content of the

grain are determined for each sample. The number of heads in the sample units is expanded to heads per acre and grain weight per head is adjusted to industry standard moisture of 12 percent. These actual yield components are used to compute the final sample gross yield per acre.

Table 1 illustrates the independent variables used to forecast the final number of heads and the grain weight per head over all maturity classes.

Table 1. Variables Used to Forecast the Components of Final Yield

Maturity Category	Final Number of Heads	Final Weight per Head
	Independent Variable	Independent Variable(s)
Pre-Flag	Number of stalks	Historical Average
Flag or Early Boot	Number of stalks	Historical Average
Late Boot or Flower	Emerged heads plus heads in late boot	1. Fertile spikelets per head, and 2. Historical Average
Milk	Emerged heads plus heads in late boot	1. Grains per head, and 2. Clip Unit Green Weight per head
Soft Dough	Emerged heads plus heads in late boot	1. Grains per head, and 2. Clip Unit Green Weight per head
Hard Dough and Ripe	Actual count of emerged heads, detached heads, and heads in late boot	Actual threshed weight per head adjusted to standard moisture determined from the laboratory work.

The forecast model for final heads per sample has the following form for each maturity class:

$$y_i = a + bx_i + \varepsilon_i$$

where x_i is the number of stalks in sample *i* if maturity is 1 or 2, and a combination of emerged heads and late boot heads in higher maturity classes, y_i is the number of heads for sample *i*, and ε_i's are random departures from the relationship. The historic coefficients a and b are estimated by ordinary least squares from data collected during the previous five years. This model produces R^2's between 50 and 80 percent for maturity class 1 and 2, in the 80's and 90's for class 3, and in the 90's to high 90's for classes 4 and 5.

The forecast models for final weight per head are a weighted combination of two regressions for each maturity except maturity class 1 and 2 which rely solely on historic weights. The form for each model is:

Model 1: $$wy_i = a + bx_i + \varepsilon_i$$

Model 2: $$wy_i = c + dz_i + \epsilon_i$$

where x_i is one plant measure such as the number of fertile spikelets, or the number of grains per head, or the weight of harvested green heads, and z_i is the other; y_i is the final weight of heads for sample i, and ε_i and ϵ_i are random departures from the relationship and there is no specification for the relationship between ε_i and ϵ_i. The coefficients (i.e., a, b, c, and d) are estimated by ordinary least squares from data collected during the previous five years. Table 1 highlights the variables used for estimating Model 1 and Model 2 for each maturity. Generally speaking, Model 1 and Model 2 have poor diagnostics, the R^2 values begin at 0 percent for maturity class 3 and neither model is regularly above 50 percent even for maturity class 5. Since Model 1 and Model 2 sometimes lead to performance ambiguity, a composite is constructed for each maturity class as follows:

$$\widehat{WY}_i = \frac{R_1^2\ (\widehat{wy}_{i1}) + \ R_2^2(\widehat{wy}_{i2})}{R_1^2 + \ R_2^2}$$

where, $\widehat{WY}_i$ is the weighted forecast of grain weight per head for sample i from Model 1 and 2, and R_1^2 and R_2^2 are the multiple correlation coefficients for Model 1 and 2 respectively. In class 3, the 5-year historic grain weight per head is manually assigned an R^2 of .2.

Forecasting Yield for Sample Fields

$$\widehat{GY}_i = (\hat{y}_i) * \left(\widehat{WY}_i\right) * \frac{CF_i}{RS_i}$$

where, RS_i is the row spacing measurement across eight rows and CF_i is a conversion factor defined as [(43560)(8)(12)] / [(6)(60)(453.58)(21.6)] = 1.186 adjusts for area measured and weight changes from metric to English. The parts are as follows:

43,560 is the number of square feet per acre,
8 adjusts for measuring across 8 row spaces,
12 converts inches to feet,

6 is rows counted in the sample units,
60 converts pounds to bushels,
453.58 converts grams to pounds and
21.6 is the width of the wheat frame in inches.

State Average Forecasts and Estimates

In each month, the sample level gross yield forecasts are averaged to the State level. Since the sample is self-weighting, the simple mean of the sample forecasts is an unbiased estimate of the State gross yield. Therefore,

$$\bar{G} = \frac{1}{N_G}\sum_{i}^{N_G} G_i$$

Where $\bar{G}$ represents the mean gross yield for each state,

and the standard error of the estimate is:

$$S_{\bar{G}} = \sqrt{\sum_{i}^{\mathrm{N_G}} \frac{(\mathrm{G_i} - \bar{\mathrm{G}})^2}{\mathrm{N_G}(\mathrm{N_G} - 1)}}$$

Simple means are also calculated for most all independent variables and harvest loss measures. These calculations are used to support analysis of the State level gross, or net, yield. Notably, the State average grain weight per head is calculated using a weighted mean. The weighting variable is the sample's Heads per Square Foot.

State Average Grain Wt. per Head = ∑(Sample Field Grain Wt. per Head * Sample Field Heads per Sq. Ft) / ∑(Sample Field Heads per Sq. Ft)

Harvest Loss

State level harvest loss is forecasted as the mean of all sample level harvest losses.

$$\bar{L} = \frac{1}{N_L}\sum_{i}^{N_L} L_i$$

where L_i is the harvest loss in sample i and N_L = number of samples with gleaning data.

$$S_{\bar{L}} = \sqrt{\sum_{i}^{N_L} \frac{(L_i - \bar{L})^2}{N_L(N_L - 1)}}$$

State Level Net Yield

Net Yield is a direct extension of the State level aggregates.

$$\overline{NY} = \bar{G} - \bar{L}$$

And the standard error of the State level net yield :

$$S_{\overline{NY}} = \sqrt{S_{\bar{G}}^2 + S_{\bar{L}}^2 - \frac{2}{N_G} COV(G,L)}$$

Where

$$COV(G,L) = \sum_{1}^{N_L} \frac{(G_i - \overline{G})(L_i - \overline{L})}{N_L - 1}$$

Production for the State

Production, P, for the State is the product of estimated State-level net yield and acres to be harvested for grain

$$P = (A_{harv})\ (\overline{NY}),$$

And a standard error of:

$$S_P = \sqrt{(A_{harv}^2)\left(S_{\overline{NY}}^2\right) + (\overline{NY}^2)\left(S_{A_{harv}}^2\right) + \left(S_{\overline{NY}}^2\right)(S_{A_{harv}}^2)}$$

Gross Yield for Samples with Incomplete Data

Gross yield is forecasted or estimated from the current month's survey data. In some cases, current data are unavailable and data from a previous month may be used to compute gross yield, or no gross yield may be computed for the sample. The different cases are discussed below.

Refusals

If the farmer refuses permission to enter the field, the sample is lost for the season. In this case the yield for this sample is left missing. Consequently, the refused sample contributes no new information to the state level average yield. The sample, and the acreage represented by it, is assumed to be the state's average gross yield.

Inaccessible Samples

Occasionally, some samples are inaccessible due to scheduling or field conditions. If data from a previous visit are available, the previous forecast is carried forward. Otherwise, the sample is excluded from gross yield calculations. The sample must still be intended for harvest as grain.

Early Farmer Harvest

If a previously laid out sample is harvested by the farmer before current data can be collected, the previous month's predicted yield is brought forward.

Lost, Abandoned, Destroyed Samples

If a sample is lost, abandoned, destroyed, and so forth, no gross yield is computed for the sample. The sample contributes nothing to the sample-level yield indication.

Removing Bias from State Level Calculations

Forecasts of State level yields are inherently subject to differ from the final, administratively-determined, yield. The difference is due to weather and crop conditions yet to be encountered at the time the forecast is produced, the difference between weather and crop conditions in the current year and the historic weather and crop conditions utilized to predict current yields, and systematic non-sampling errors which contribute to forecast error. NASS makes every attempt to minimize the impact of these three sources of error by means of administratively determine final values. These final, sometimes referred to as *official* values are used as dependent variables to estimate an ordinary least squares equation with the averages calculated from objective yield samples as the independent variable. For example, each state will have an administratively-determined, *official* wheat yield for all previous years. To forecast the *official* wheat yield for the current year, NASS regresses the objectively determined State level average wheat yields from previous years to the corresponding *official* wheat yields. This process provides a historical context for weather, biases, and non-sampling errors.

The regression equation is:

$$y_t = a + b\bar{x}_t + \varepsilon_t$$

where y_t is the Official state yield in time period t, $\bar{x}_t$ is the calculated State average net yield in time period t, a and b are ordinary least squares estimates, and ε_t is a random departure from the relationship.

Computational Examples

Yield

Yield indications are derived by initially calculating the two yield components, number of heads, and weight per head. These components are forecasted by applying linear regression models to sample data. The models used by a State vary by class of wheat, geographic district, and maturity category. The parameters for these regression models are computed from the 5 most recent years of historical sample data for that State.

The following pages will demonstrate, by example, how models are used to forecast yield in the various sample maturity categories.

Maturity Category 1, pre-flag:

For samples in the pre-flag maturity category, the sample variable used to forecast number of heads is number of stalks. The variable to forecast the weight per head is the historical average weight per head.

Suppose the appropriate regression models are:

Number of heads = 180 + .2 * (total number of stalks),

and

Weight per head = .64

If the sample has 920 stalks, then the forecasted number of heads = 180 + .2 * (920) = 364.

Therefore, the forecast of gross yield per acre from a sample with an 8-row width of 6.4 would be:

Gross yield = [(number of heads)(weight per head)(conversion factor)] / (8-row width)

= [(364)(.64) (1.186)] / 6.4 = 43.17.

Maturity Category 2, flag or early boot:

For samples in the flag or early boot maturity category, the sample variable to forecast number of heads is number of stalks. The variable to forecast the weight per head is the historical average weight per head.

Suppose the appropriate regression models are:

Number of heads = 90 + .4 * (total number of stalks)

and

Weight per head = .64

If the sample unit has 650 stalks, then the number of heads = 90 + .4 * (650) = 350.

Therefore, the forecast of gross yield per acre from a sample with an 8-row width of 6.4 would be:

Gross yield = [(number of heads)(weight per head)(conversion factor)] / 8-row width

= [(350)(.64) (1.186)] / 6.4 = 41.51

Maturity Category 3, late boot or flower:

For samples in the late boot or flower maturity category, the variable to forecast number of heads is the sum of emerged heads and heads in late boot. Two models are used to forecast weight per head. The first model uses the number of fertile spikelets per head and the second model uses the historical average head weight. These are weighted together using R-square of the first model and a weight of 0.2 for the second.

Suppose the appropriate regression models are:

number of heads = 23 + .9 * (total number of emerged heads + heads in late boot)

weight per head (Model 1) = .12 + .04 * (number of fertile spikelets), with an R-square of .31

weight per head (Model 2) = .64

If the sampled unit has a total of 336 emerged heads and heads in late boot, and 15 fertile spikelets per head,

then,

number of heads = 23 + .9 * (336) = 325,

and weight per head (Model 1) = .12 + .4 * (15) = .72

The composite weight per head forecast is:

$$\text{weight of heads} = \frac{R^2 \text{ Model 1}(\text{wt per head Model 1}) + R^2 \text{ Model 2}(\text{wt per head Model 2})}{R^2 \text{ Model 1} + R^2 \text{ Model 2}}$$

so that in this example:

weight per head = [.31(.72) + .20 (.64)] / [.31 + .20] = .69

Therefore, with 8-row width of 6.4,

Gross Yield per Acre = [(number of heads)(wt per head)(conversion factor)] / 8-row width

= [(325)(.69)(1.186)] / 6.4

= 41.56

Maturity Category 4, milk:

For samples in the milk maturity category, the variable to forecast number of heads is the sum of emerged heads and heads in late boot. Two models are used to forecast weight per head. The first model uses the number of grains per head and the second model uses the clip unit green weight per head. There are weighted together using the R-squares of the models.

Suppose the appropriate regression models are:

number of heads = 6 + 1.0 * (total number of emerged heads + heads in late boot),

weight per head (Model 1) = .59 + .003 * (number of grains per head), with an R-square of .95,

and

weight per head (Model 2) = .5 + .16 * (clip unit head weight), with an R-square of .97

If the sampled unit has a total of 331 emerged heads and heads in late boot, 18 grains per head, and a clip unit green weight of .74,

then

number of heads = 6 + 1.0 * (331) = 337,

weight per head (Model 1) = .59 + .003 * (18) = .64,

and

weight per head (Model 2) = .5 + .16 * (.74) = .62

The composite weight per head forecast is

$$\text{wt per head} = \frac{R^2 \text{ model 1(wt per head Model 1)} + R^2 \text{ Model 2(wt per head Model 2)}}{R^2 \text{ Model 1} + R^2 \text{ Model 2}}$$

so that in our example

weight per head = [.95 (.64) + .97 (.62)] / [.95 + .97] = .63

Therefore, with 8-row width of 6.4,

Gross yield per acre = [(number of heads)(wt per head)(conversion factor)] / 8-row width

= [(337) (.63) (1.186)] / .64 = 39.34

Maturity Category 5, soft dough:

For samples in the soft dough maturity category, the variable to forecast number of heads is the sum of emerged heads and heads in late boot. Two models are used to forecast weight per head, one using the number of grains per head and the other using the clip unit green weight per head. These are weighted together using the R-squares of the models.

Suppose the appropriate regression models are:

number of heads = 7 + 1.0 * (number of emerged heads + head in late boot),

weight per head (Model 1) = .33 + .02 * (number of grains per head),
with an R-square of .98,

and

weight per head (Model 2) = .37 + .33 * (clip unit green wt.),
with an R-square of .99.

If the sample unit has a total of 332 emerged heads and heads in late boot, 21 grains per head, and a clip unit head weight of .93,

then

number of heads = 7 + 1.0 * (332) = 339,

weight per head (Model 1) = .33 + .02 * (21) = .75,

and

weight per head (Model 2) = .37 + .33 * (.93) = .68

The composite weight per head forecast is

$$\text{wt per hd} = \frac{R^2 \text{ model 1(wt per hd Model 1)} + R^2 \text{ Model 2(wt per hd Model 2)}}{R^2 \text{ Model 1} + R^2 \text{ Model 2}}$$

so that in the example

weight per head = [.98 (.75) + .99 (.68)] / [.98 + .99]

= .71

Therefore, with an 8-row width of 6.4,

Gross Yield Per Acre = [(number of heads)(wt per head)(conversion factor)] / 8-row width

= [(339) (.71) (1.186)] / 6.4 = 44.60

Maturity Categories 6 and 7, hard dough and ripe:

Actual number of heads and actual head weight are used to calculate gross yield per acre. The following final lab data and gleaning counts and measurements are obtained for a sample:

number of emerged heads, detached heads, and heads in late boot = 350,
moisture content of enumerator harvested grain = 12%,
number of heads threshed = 250, and
threshed weight of grain = 180

weight of gleaned grain = 20
moisture content of post-harvest gleaning grain = 14%

Calculate weight per head, gross yield per acre, harvest loss per acre, and net yield.

Wt. per Head = [(threshed wt of grain)(1.0 - moisture)] / [(number of heads threshed) (.880)]

= [(180) (1.0-.12)] / [(250) (.880)]

= .72

Assuming an 8-row width of 6.4,

Gross Yield Per Acre = [(number of heads)(wt per head)(conversion factor)] / [8-row width]

= [(350) (.72) (1.186)] / 6.4

= 46.70

Harvest loss per acre = [(wt of threshed grain)(1.0-moisture content of grain) (conversion factor)] / [(.880)(8-row width)]

= [(20) (1-.14) (1.186)] / [(.880) 6.4]

= 3.62

Net Yield = Gross Yield - Harvest Loss

= 46.70 - 3.62

= 43.08

CHAPTER 9 PREPARATION OF OFFICIAL STATISTICS

Overview

A fundamental principle behind the estimation process is that precision of the sample survey estimates is greatest at the aggregated regional and U.S. levels. The precision of a sample survey estimate is measured by the estimated sampling error. In theory, many independent sample surveys could be conducted simultaneously, each producing estimates of acreage, yield, or production. The extent to which these independent estimates would differ from each other is called the sampling error and can be estimated from each sample. For NASS surveys, the sampling error at the U.S. level for corn acres is about 1.0 percent, 2.3 percent in major States and 10-15 percent in other States.

The sample surveys are designed to produce State level estimates of acreage, expected yields, final yields, and total production. The surveys are conducted by each State, and the first level of analysis is done by each State. Each Regional Field Office does its independent appraisal of the relationships between the survey estimates and the final official statistics and forwards this information to Headquarters.

While each Regional Field Office is analyzing its survey data, statisticians in Headquarters are doing a parallel analysis of all survey data at the State, U.S., and regional levels. For the major field crops discussed in this paper, a formal Agricultural Statistics Board is convened to review regional indications and determine the official forecast or estimate. This Board is made up of 7 to 10 statisticians representing different divisions of NASS. Each Board member evaluates the regional survey indications and supporting data and determines their forecast or estimate. Each member brings their individual perspective to the review which can result in different conclusions being drawn. Through review and discussion, the Board must collectively reach a consensus and establish the National number. The Board process ensures all perspectives are examined and the national or regional forecast or estimate is the result of a thorough analysis. The summation of the individual State estimates as prepared by each State is compared to the Board number. The Headquarters statisticians will re-examine all national and State data relationships and either adjust State estimates, so they sum to the U.S. or change the previously determined U.S. number.

Domestic supply is a key factor in the marketing of any commodity and affects the day-to-day business decisions of the industry. As a result, crop production forecasts and estimates are extremely sensitive data. Premature or privileged disclosure of NASS numbers would give individuals or groups an unfair advantage in the marketplace. NASS must ensure that all official numbers are made available to everybody at the same time, making security a very big issue. All data, both individual and summary, are protected against disclosure at every step of the forecasting and estimating process. Data access must be always restricted in the Regional Field

Offices and Headquarters. As data are summarized and aggregated to regional or national levels, the security is heightened. Yield forecasts and estimates from the largest producing States are encrypted before transmission to Headquarters. As data are received in Headquarters and commodity statisticians begin the review process, offices are designated as secure offices and visitors are denied access.

The formal meeting of the ASB to establish the final numbers and prepare the report is conducted under "lock up" conditions. Lock up begins with a complete isolation of all facilities required by the Board. All doors are locked, windows and elevators are covered and sealed, phones are disconnected, and the computer network inside "lock up" is isolated from the full network. Transmitters are not permitted, and the area is monitored for electronic signals. Highly speculative data are decrypted only after the area is secure. Only after all security is in place does the Board begin final deliberations. The area remains locked up until a prescribed release time (12:00 p.m. for Crop Production) at which time the report is disseminated in electronic and paper forms.

This chapter is devoted to describing the interpretation process followed by commodity statisticians to arrive at the best number. A brief discussion of acreage estimates is followed by a detailed explanation of forecasting yields. The last two parts address end of season estimates of acreage, yield, and production followed by an overview of how balance sheets are used as a check on the final estimates.

Acreage Estimates

The summary programs provide point estimates of acreage planted, called **direct expansions**, and measures of change from a previous estimate, called **ratio estimates**.

Direct expansions measure the level of the value of the item being estimated. For area frame surveys, every segment of land selected from the area frame has a known probability of selection. The inverse of the probability of selection for each sample unit (expansion factor) multiplied times the acres found in the segment are summed across the sample to determine a direct estimate of acres planted to each crop. List samples also have known probabilities of selection and their data can be similarly expanded to provide direct expansions in a multiple frame design.

Ratio estimates are used to measure change from a previous estimate of the same item (preliminary acres for harvest) or a related item (previous year's planted acres). These types of ratios rely on matched reports from both surveys. The area frame sample is divided into five independent rotation groups with four groups carried over from one year to the next. The consecutive year's data from these four rotation groups can be matched to provide a measure of the percent change in acres planted. The list sample can be similarly structured to provide survey to survey matched samples and ratios can be computed in the multiple frame design.

The determination of the official estimates of acres planted is based on an analysis of the historical and current direct expansions and ratio estimates as they compare to the final estimates of planted acres. The analysis is based on "difference" estimates which measure the average difference between the survey indications and the final estimates. This analysis is done at both the State and U.S. levels with any differences being reconciled in Headquarters.

The June Area Survey (JAS) and the June Agricultural Survey provide the benchmark estimates of acres planted. In some years, weather related problems delay planting activities which means farmers are reporting acres they still intend to plant. When this occurs, subsamples of farms included in the JAS are re-surveyed in July to determine the acres actually planted. These updated acreage estimates are reviewed similarly to the procedures followed in June. Yield surveys provide ratio indications which are used to monitor changes in acreages.

Acres harvested and to be harvested are key variables for deriving production forecasts and estimates, respectively. Direct expansion and ratio of change estimators are also used to estimate harvested acres. In addition, the ratio of harvested to planted acres as provided by the survey can be multiplied times planted acres for another indication of harvested acres. The "difference" analysis described above is also used to determine the official harvested acreage estimates.

Yield Forecasts

Arguably, the most watched publications of NASS are the Crop Production Reports containing the early season forecasts of production for the major field crops. Early season production forecasts are key pieces in the price discovery mechanism for these billion-dollar crops. This kind of scrutiny demands a review as comprehensive as the security provisions to ensure the best forecasts and estimates.

The yield surveys produce vast amounts of data for analysis. The modeling processes described in previous chapters produce multiple indications of net yield per harvested acre. The first monthly forecasts for a crop feature some key indications that include but not limited to average field level yield regressed to official estimates, average counts regressed to official estimates, average projected yield reported by farmers in the Agricultural Yield (AY) Survey regressed to official estimates, and model-based indications. Once harvest begins, average realized farmer yields are regressed to official estimates. In addition to the point estimates, forecast errors of the regression equations and models are also computed. Adding and subtracting these forecast errors from the forecast value forms a forecast range for each indication. Usually, the ranges for these indications overlap defining the range that simultaneously satisfies all forecasts.

Merely selecting a yield from within the overlapping range is not the end of the process. Commodity statisticians must determine if all of the other pieces of available data support the "candidate" yield forecast. Some of the more important things to evaluate are:

1. Average maturity category - Enumerators determine the maturity category of each OY sample. The average maturity category helps commodity statisticians align the crop calendar with the monthly report calendar. This maturity should be consistent with weekly crop progress data. Extremely late (below average maturity) crops and extremely early (above average maturity) crops often produce data that lie in the fringes of the historical data and may result erratic forecasts due to extrapolating the forecast equations.

2. Forecasted fruit count - Even in the first survey month, plant counts are obtained for all OY samples and forecasts of the number of fruit per acre can be made every month for every crop regardless of maturity. As the fruit develop, counts of immature fruit are used to provide even more precise forecasts of fruit expected at harvest. Experience has shown that forecast equations for fruit count have very high R-square values and produce very accurate forecasts. In fact, the linear relationship is so strong these equations are robust against extrapolation.

3. Forecasted fruit weight - As easy as it is to forecast count, forecasting weight is equally difficult. In the early months, there is no measurable characteristic to use in a model and historical average fruit weights must be used. Even after the fruit set and measurements can be made, data are extremely variable, and correlations are not very high. Thus, fruit weight forecasts have much larger forecast errors than fruit count. Extreme maturities can significantly impact weight models. Fruit weight often becomes the key discussion factor in Board deliberations.

4. Averages of the raw data - The raw counts are definitionally stable across years. As noted in earlier chapters, parameter estimates for the forecast equations are recomputed each year using a “rolling” dataset. Changes in forecasts from one year to the next are a combination of changes in the current raw counts and new equations. These changes are confounded in the forecast and isolating the changes in farmer practice from the differences in the crop season from the trends in yield is difficult. The raw counts give insight into true shifts in the components of yield like planting patterns and plant populations, fruit per plant, size of ears, etc. When the number of plants per acre is higher than ever recorded before, a record fruit count forecast and, possibly, a record yield should be no surprise.

5. Interaction of fruit count and fruit weight - Statisticians can obtain insight into yield levels by looking at the interaction of the two main components, count and weight. The final yield may be the same for 2 years, but they may be a result of different components. A simple scatterplot of count against weight with points labeled as to year clearly show how the current forecasts compare to the final estimates of previous years.

6. Month to month shifts - Each of the five items discussed above can apply to a stand-alone, single month analysis. However, after the first forecast month, each can be applied in a month-to-month analysis. The second and third forecasts are measured against the previous forecast and the statistician must understand what is causing the indications to move up or down. Are the raw counts and measurements changing? Are the models forecasting the components at a different level? How are the farmer assessments of their yield prospects in the AY survey changing? What effect is final harvest data having on the indications?

This process is done independently in each State and at the combined level in Headquarters. Headquarters statisticians make the final determination, and, when necessary, will establish forecast or estimate that differs from the State(s) recommendations so the State numbers are additive to the U.S. level.

Final Estimates - Acres, Production, Stocks

Chapter 2 contains a discussion of the Agricultural Surveys and how they relate to yield surveys. The September and December Agricultural Surveys are the vehicles by which final acreage, yield, and production data are obtained. Final end-of-year estimates are prepared from these data. The September survey focuses on the small grains and is timed to be conducted as harvest is nearly complete. The December survey collects the row crops end of season data, and it is also timed to occur as harvest winds down. Respondents to these surveys report actual acres harvested and the actual yield or production realized from harvest. Grain in storage data are collected at this time and are used to estimate “carry out” stocks which are used in balance sheet reviews of the major crops.

The OY sample plots are harvested at crop maturity. A sample of plots are gleaned for harvest loss after the sample fields are harvested. These crop cuttings form a secondary final yield indication, but, more importantly, they are used to compute parameter estimates in future years. The final OY observations serve as the values of the dependent variables of the regression models.

Balance Sheets

The end-of-season estimates of acres harvested, yield, production, and stocks are reviewed in combination using a balance sheet approach. Up to this point, the approach is to consider acres and yield independent of the supply and demand relationship. The balance sheet offers a more global look at how the estimates fit into the bigger picture. Using estimates from NASS surveys, and administrative data from outside sources, commodity statisticians can construct a balance sheet to see if the estimates reconcile with these sources. Using corn as an example, a December 1 balance sheet analysis would look as follows:

Quantity carried over from previous year (September 1 on-farm and off-farm stocks for corn and soybeans, June 1 for wheat)

Plus Imports since September 1
Current Production (NASS estimate)

Equals Beginning supply as of December 1

Minus Disappearance since September 1
Exports
Processing
Feed and seed

Balance Sheet Indication of December 1 stocks

Survey Indications of December 1 stocks (on farm and off farm)

Residual

The residual component of the balance sheet is the difference between the survey indicated stocks and the balance sheet stocks. Each survey component of the balance sheet contains sampling and non-sampling errors. The disappearance items such as exports and processing are based on administrative sources with varying levels of completeness. For these reasons, it is not reasonable to expect a zero residual; however, an unreasonable residual is cause for alarm and triggers a second review of the elements in the balance sheet. The objective is to have a reasonable balance and still have the estimates within the range indicated by the surveys.