United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Statistical Methods Branch

SMB Staff Report
Number SMB 06-01

May 2006

THE YIELD FORECASTING PROGRAM OF NASS

The Statistical Methods Branch

THE YIELD FORECASTING PROGRAM OF NASS, by the Statistical Methods Branch, Estimates Division, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C., May 2006. NASS Staff Report No. SMB 06-01.

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 Statistical Methods 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.

ACKNOWLEDGMENTS

This paper is a collaboration of the Statistical Methods Branch staff; Bill Arends, Tom Birkett, Herb Eldridge, Gary Keough, Mark Schleusener, and Dave Aune. Special thanks goes to Fred Vogel, Director, Estimates Division, for contributing Chapter 10. Finally, thanks to our predecessors in Methods Branch and to the commodity specialists who have shaped this ever evolving program.

TABLE OF CONTENTS
THE YIELD FORECASTING PROGRAM OF NASS

CHAPTER 1 - OVERVIEW 1

CHAPTER 2 - SAMPLE DESIGNS 4

CHAPTER 3 - AGRICULTURAL YIELD SURVEYS 10

CHAPTER 4 - GENERAL YIELD FORECASTING PROCEDURES 18

CHAPTER 5 - CORN OBJECTIVE YIELD METHODS 27

CHAPTER 6 - SOYBEAN OBJECTIVE YIELD METHODS 45

CHAPTER 7 - COTTON OBJECTIVE YIELD METHODS 61

CHAPTER 8 - WHEAT OBJECTIVE YIELD METHODS 76

CHAPTER 9 - POTATO OBJECTIVE YIELD METHODS 98

CHAPTER 10 - PREPARATION OF OFFICIAL STATISTICS 103

REFERENCES

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. Vogel and Bange [1] provide a detailed discussion of the WAOB process and the interaction between the two Agencies.

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); ©) 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. Acres with multiple harvests 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 2 weeks in June. Expected corn and soybean yields are obtained monthly, August through November, from two different types of yield surveys. 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 an objective measurement survey. The survey of growers, the Agricultural Yield Survey (AYS), 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 survey, known as the Objective Yield (OY) Survey, covers wheat, corn, soybeans, cotton, and potatoes. The OY surveys consist of a sample of fields in which counts and measurements are made to plants in random plots laid out in each field.

Data from the yield surveys reflect conditions as of the first of the month, as data are collected during the last week of the previous month and the first 2 or 3 days of the survey month. Crop production forecasts are based on conditions as of the survey reference date and projected assuming normal conditions for the remainder of the season. For OY modeling, the concept of “normal” is the data relationships contained in the historical datasets used to estimate the parameter values of the forecast equations. From a laymen’s perspective, the assumption of "normal conditions" is that temperatures and precipitation will be at historical averages for the

remainder of the season. It is assumed that the first killing frost will occur on the historical average date. The crop maturity and conditions at the reference date are evaluated against the time remaining until the expected frost--if one third of the crop will not reach maturity until the frost date has passed, it is assumed that some frost damage will result. For AYS, "normal" is the collective judgement and experience of the respondents.

The primary goal is to provide the most accurate production forecast possible given the available survey data. If there is a significant change in conditions between the survey period and the report release date such as a killing freeze, serious heat wave, beneficial rains, etc., the analysts must still forecast within the range of data indicated by the survey. While the official estimate may represent a departure from the survey averages, it will still reflect the current crop conditions within the ranges provided by the data. When NASS states as policy that it is forecasting based on conditions as of the first of the month, it is saying that it will establish yields within the range of the survey results.

When forecasting crop yields, NASS does not attempt to predict future weather conditions. Long-range weather forecasts are not used in any forecast models. To the extent that conditions depart from normal, the forecasts will follow.

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. Objective measurement surveys are conducted for corn, cotton, soybeans, wheat, and potatoes only and are referred to as Objective Yield Surveys.

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 (mail and telephone). The list frame does not provide complete coverage of all farms and is not stable since farming arrangements are constantly changing. Both frames are maintained by the State, allowing some flexibility to customize for local situations.

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 Objective Yield survey samples are selected from respondents to the March Crops/Stocks Survey or June Agricultural Surveys (JAS). Samples for corn, cotton, soybeans, durum, and spring wheat are selected from JAS area tracts having the commodity of interest. Winter wheat samples are selected from March Crops/Stocks Survey respondents with winter wheat planted for harvest as grain. For potatoes, samples are selected from JAS list operators reporting fall potato acreage planted.

Samples are selected as soon as possible following the final summary of the March Crops/Stocks Survey or JAS. For geographic representation of the samples, the records are first sorted by state, district, county, segment, tract, crop and field. 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. These 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 Wisconsin. This frame was constructed in 2001 with seven land-use strata covering all 55,011 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 sample sizes and expansion factors for 2005 are shown. The expansion factors are inverses of the sampling fractions. Note the allocation of the samples favors the more intensely cultivated areas with 180 of 219 segments falling in the first three strata. Strata with little or no agriculture are lightly sampled.

WISCONSIN - 2005

Stratum	Square Miles	Segment Size	Segments in Frame	Sample Size	Exp Factor	Stratum Definition
11	11,836	1.00	11,819	70	169	>75% Cultivated
12	5,664	1.00	5,665	30	189	51-75% Cultivated
20	18,203	1.00	18,195	80	227	15-50% Cultivated
31	1,128	0.25	4,500	5	900	Agri-Urban
32	139	0.10	1,399	2	700	Commercial
40	17,719	2.00	8,857	30	295	<15% Cultivated
50	322	pps	34	2	17	Non-Agricultural
Total	**55,011**		**50,463**	**219**		

The list frame is also stratified for crop and on-farm stocks surveys. Acreage survey designs use total cropland as the main stratification variable and stocks uses storage capacity. Speciality strata may be included to deal with commodities that are difficult to measure using an “all purpose” stratification. These items may be handled separately or combined into a dual purpose survey. Each State uses this basic design with strata definitions customized for their State. Again, different sampling rates are used to achieve the most efficient sample with larger operations sampled more heavily. In the strata containing the largest operations, all operations are selected.

An example of a combined stratification describing the 2005 Illinois Crops/Stocks Survey is shown below. Note that this design gives capacity a higher priority than cropland and one specialty stratum (73) is included. A priority ordering of the strata is established and each list record passes through this hierarchy and is classified into the first (highest) stratum for which it qualifies. This ensures strata are mutually exclusive and independent.

ILLINOIS - 2005
CROPS/STOCKS

Strata	Boundaries	Population	Sample Size	Interval
97	Capacity 500K+	50	50	1.0
95	Cropland 7,000+	21	21	1.0
79	Cropland 2,500-7,499	831	159	5.2
78	Capacity 50K-499,999	6,516	485	13.4
73	Sorghum 1+	813	307	2.6
72	Cropland 600-2,499	6,847	492	13.9
66	Capacity 10K-49,999	9,108	387	23.5
65	Cropland 100-599	12,232	464	26.4
62	Capacity 4K-9,999	1,101	24	45.9
	Total	**37519,**	**2,389**	

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 as beginning in June with

September, December, and March completing the cycle. Each survey employs multiple frame methodology.

All surveys have list samples of about 50,000 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 survey features complete enumeration of an area sample of about 10,800 segments. The June area frame sample allocation favors spring planted crops. 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 March and September samples include only incompleteness (nonoverlap) tracts from the June area sample, usually around 5,700 tracts, and provide multiple frame indications.

Survey content differs each quarter to meet the varying requirements of the estimating program. The June, December, 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) Durum Wheat OY (area) Other Spring Wheat OY (area) Potato Objective Yield (list)
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	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

Two grower reported surveys, called the Agricultural Yield Surveys (AYS), cover most of the field crops estimating program. The survey covers most crop yield data needs for each State. The AYS 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, and barley. The second AYS 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, hay and tobacco.

The subsampling design for the AYS restricts the selection to sampled list strata. This excludes the largest (preselect) list stratum and the nonoverlap tracts from sampling The assumption is made that mean yields from the excluded subpopulation are the same as those included in the subpopulation for AYS. The AYS uses a multivariate probability proportionate to size (MPPS) sample design, with list frame control 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 the AYS. The overall sample size varies, depending upon the month, from a maximum of 27,000 in August, to a minimum of 5000 in June. In the AYS, targeting is especially important for the commodities that are considered rare or for specialty crops.

Objective Yield Surveys

Objective measurement surveys (OY), are conducted for corn, cotton, soybeans, winter wheat, other spring wheat, durum wheat, and potatoes. These surveys are very costly and are conducted only in the top producing States. The States in the OY program usually produce in excess of 75 percent of the U.S. total. For each commodity except potatoes, a series of monthly net yield forecasts culminates in a final net yield at maturity. Only the final net yield is measured for

potatoes.

As noted in the previous table, all OY samples except potatoes and winter wheat are drawn from an area frame parent survey. June area data are collected and recorded at the field 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 right down to the exact field. Fields 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.

Potato and winter wheat acres are collected at the farm level on the Crops/Stocks 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 also a self-weighting sample. Farms and fields within farms may be selected more than once.

CHAPTER 3 AGRICULTURAL YIELD SURVEYS

The Agricultural Yield Survey (AYS) 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 AYS 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 AYS uses a multivariate probability proportionate to size (MPPS) sample design, with list frame control data used to determine a unit's selection probability. 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 as the name implies, there are multiple commodities or control 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 having the largest proportion of total and sample size .

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

In August the AYS sample includes operations from both SG and RC samples, a composite weighting methodology was developed. Using such an approach allows maximum use of the information obtained from AYS responses. That is, information about SG crops that was obtained from AYS RC only sample records can have that SG information used in AYS SG survey indications. Similarly, information about RC crops that was obtained from AYS SG only sample records can have that RC information used in AYS 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. A single survey instrument is prepared and respondents are asked all questions regardless of the sampling base

Data Collection

The reference date of every AYS is the first of the month. The States are instructed to collect data as close to the reference date as possible. In practice, the data collection period begins on the 25th of the previous month and ends about the 3rd of the survey month. This amounts to about 7 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. Electronic data reporting (EDR) via the internet will begin with the 2006 crop year.

The complete questionnaire for AYS includes acres for harvest and yield for each crop of interest. Nearly all the AYS crops will be asked in the initial month of the survey - May for SG and August for RC. There are a few crops which are not asked during the initial month and for these crops acres to be harvested and yield are asked during the month in which they begin. Harvested acres, once reported, are not re-asked but carried forward for ensuing months the crop yield is asked. If an operator is inaccessible during the initial month of a crop, harvested acres is asked once contact is made. Acres reported from the base survey - March Crops/Stocks Survey for SG and June Agricultural Survey for RC - are carried on the sample master for each AYS respondent. For most crops, acres to be harvested are carried forward to the AYS survey. In some cases where crops are planted later in the year, planted acres are carried forward instead of harvested acres, and in a few crops, both harvested and planted acres carried forward from the base survey.

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 CATI versions also include a question on whether each crop has been harvested and the reported yield is final. Once harvested, the yield for those crops are not asked on subsequent interviews, but are brought forward and used in subsequent months summaries. The AYS has an additional ability to measure harvested acreage changes during the crop year when extreme weather conditions exist. This distressed acres sub-survey provides an additional ratio indication of current acres versus previously reported acres for harvest. The sub-survey can target specific crops in States that have experienced extreme weather conditions.

States are expected to achieve a minimum response rate of 80 percent. States are expected to conduct a follow-up of mail and telephone nonresponse sufficient to achieve this level. States must also monitor response by crop to determine the amount of follow-up necessary to achieve 50 usable reports for major crops.

Analysis

All AYS data are processed through a central edit program as the first step in data review. This edit performs all within-record data (microdata) checks. Data from paper versions are manually reviewed, keyentered, and merged with CATI data. Data collected through EDR data is merged with CATI data prior to editing. The machine edit checks that reported data are within absolute limits, compares acres reported on the AYS 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 made in the main edit are performed by the CATI software which allows enumerators to probe for additional information and correct errors during the interview when suspect values are recorded. The CATI data are still processed through the central edit to merge them with the data from non-CATI sources, to prepare the dataset for the summary program, and to ensure consistency.

The next step is an across-record (macrodata) review of the raw data via the Interactive Data Analyses System (IDAS). 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.

IDAS displays extreme differences between surveys. In May and August, AYS 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. IDAS will also display current acres for harvest versus previous reported acres for harvest when acres are reasked due to distressed conditions.

The IDAS output can be displayed in graphical and tabular form. The graphs provide a frequency distribution of the reported data. It also provides tabulations of actual record level data.

Summarization

The AYS summary program is really two summaries combined into one print output. The first

part, the probability summary, applies a combination of the appropriate mpps sample weight and a nonresponse adjustment weight, to produce an indication with associated measurable statistical error. The second part, a non-probability indication, pools the useable reports from all respondents for each crop within an Agricultural Statistics District (ASD). The reported yields, weighted by harvested acres are generated at the ASD level. The nonprobability state yield is calculated by weighting each ASD yield by the total acres in the ASD for that specific crop. This produces a state level yield indication, however there is no measurable level of precision associated with this indications. Both the probability and nonprobability indications are sorted by ASD prior to output for comparative purposes.

The probability summary computes three types of indications:
1. Average expected yield.
2. The ratio of yields reported on consecutive AYS surveys.
3. The ratio of any two acreage items from either the AYS, the parent survey, or both.

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 non-response weight is calculated using the crops/stock 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 useable respondents. The k*th* stratum non-response weight would be calculated as:

$$w_{nr}(k) = N_k / n_k$$

where N_k = *Total sample size for statum* (k)

n_k = *Total number of useable responses 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 non-response weight:

$$w_{i\,k} = w_{mpps_i} * w_{nr_k} .$$

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

$$p_i = (y_i)\ (a_i) .$$

The d*th* ASD production is then the sum of the product of each i*th* useable respondent's production and weight within the that district:

$$P_d = \sum_{j \text{ (where } j=d)} \sum_i (p_{ij})(w_i)(k_i) .$$

where k_i = 1 if the i*th* sample response is useable
0 else.

Likewise, total acreage (A) for the d*th* ASD would be calculated as:

$$A_d = \sum_{j \text{ (where } j=d)} \sum_i (a_{ij})(w_i)(k_i) .$$

Yield (Y) for the d*th* ASD would then be the ratio of sum of all production over all acreage for that district:

$$Y_d = \frac{\sum_{j \text{ (where } j=d)} \sum_i (p_{ij})(w_i)(k_i)}{\sum_{j \text{ (where } j=d)} \sum_i (a_{ij})(w_i)(k_i)} .$$

For state level indications one simply sums across all *k* useable records and districts in the state for that particular crop.

The ratio of yields reported on consecutive AYS surveys quantifies the change in the collective judgement of the respondents. The actual computations are made by converting the current and previous month's 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. A kth useable value The equation for dth ASD appears below:

$$Yield\ R_d = \frac{\sum_{j \text{ (where } j=d)} \sum_i (y_{ij})(a_{ij})(w_i)(k_i)}{\sum_{j \text{ (where } j=d)} \sum_i (y'_{ij})(a_{ij})(w_i)(k_i)} .$$

where y'_{ij} = previous month's expected yield for sample I in district j
k_i = 1 if both current and previous month i*th* response useable
0 else.

Again, the state level indication for the ratio of yields across surveys would simply be the

summation across all *k* useable records for that particular crop across all *d* districts in that state.

The ratio of any two acreage items from the AYS and the parent survey offers several key indicators. Any ratio of AYS reported acres to the acres reported on the parent survey provides a link between the AYS and the parent survey. Every AYS 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. Ratios of harvested acres are computed for most crops, planted acres are used for some crops, and a few crops have both. Acreage ratios serve a couple of purposes. First, they provide an assessment of the presence or absence of reporting errors in the data used to determine the AYS subsampling population. In years without unusual conditions, these ratios would be expected to be near 1.0, verifying data quality in both surveys. The second use is as an indication of changes to acres occurring due to extreme conditions. In years with delayed planting, a planted acres to planted acres ratio provides a measure of unfulfilled intentions. In a drought year, harvested to harvested ratios provide insight into increasing abandonment. Under these conditions, reporting errors become confounded with true change in the acreage level.

The ratio of harvested to planted acres within the AYS data set is the cleanest measure of current year abandonment. Percent of acres abandoned is fairly constant from year to year unless an extreme condition exists.

The general formula for the ratio between surveys for the d*th* district is shown below. Note that the current AYS weights apply in all instances.

$$\text{Acreage } R_d = \frac{\sum_{j \text{ (where } j=d)} \sum_i (a_{ij})(w_i)(k_i)}{\sum_{j \text{ (where } j=d)} \sum_i (a'_{ij})(w_i)(k_i)}$$

where,

a'_{ij} = acres reported on the parent survey for sample I in district j

k_i = 1 if both current and previous month i*th* response useable
0 else.

The formula for the harvested to planted acreage ratio for the d*th* district is:

$$H/P_d = \frac{\sum_{j \text{ (where } j=d)} \sum_i (hv_{ij})(w_i)(k_i)}{\sum_{j \text{ (where } j=d)} \sum_i (pl_{ij})(w_i)(k_i)}$$

where hv_{ij} = acres for harvest as grain in sample I in district j
pl_{ij} = acres planted in sample I in district j

k_i = 1 if both current and previous month i*th* response useable
0 else.

Although all calculations in the probability summary are correctly performed within design stratum, the printed output presents the results by ASD. This facilitates the interpretive process by making it easier to compare AYS results to other data sources reported by ASD. Temperature, precipitation, and crop progress data provide additional evidence to support the AYS 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.

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 nonprobability formulas are identical to the probability with design weights (w_i) eliminated and ASD weights substituted for stratum.

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

$$Y_d = \frac{\sum_{j\ (where\ j=d)} \sum_i (y_{ij})(a_{ij})(k_i)}{\sum_{j\ (where\ j=d)} \sum_i (a_{ij})(k_i)} .$$

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

$$Yield\ R_d = \frac{\sum_{j(where\ j=d)} \sum_i (a_{ij})(y_{ij})(k_i)}{\sum_{j(where\ j=d)} \sum_i (a_{ij})(y'_{ij})(k_i)}$$

The ratio of acreage items between the AYS and the parent survey and the harvested to planted ratio are derived as:

$$\text{Acreage } R_d = \frac{\sum_{j \text{ (where } j=d)} \sum_i (a_{ij})(k_i)}{\sum_{j \text{ (where } j=d)} \sum_i (a'_{ij})(k_i)}$$

$$H/P_d = \frac{\sum_{j \text{ (where } j=d)} \sum_i (hv_{ij})(k_i)}{\sum_{j \text{ (where } j=d)} \sum_i (pl_{ij})(k_i)} \ .$$

Let E_i denote any ASD level non-probability estimate shown above.

The State level estimate is:

$$E = \frac{\sum_i w_i E_d}{\sum_i w_d}$$

where,

w_d = external ASD acreage weight .for the d*th* district.

The summary output displays the yields and ratios by ASD and State. A special note concerning bias in the AYS data must be made here. The yields obtained are the judgement of the respondent. 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 GENERAL YIELD FORECASTING PROCEDURES

Yield Forecasting, Estimation, and Indications

To begin the discussion of NASS crop production forecasting and estimating, it is important to understand how NASS uses the terms forecast, estimation, and indications. The differences between official forecasts and final estimates is in the timing of the release. Forecasts are made before the entire crop has been harvested whereas estimates are made after the crop has been harvested. Indications are the result of applying a statistical estimator to the survey data and the resulting point estimates are interpreted by commodity statisticians to make forecasts and estimates.

The major goal of the OY program is to produce indications of expected yield and final harvest yield with actual plant counts and measurements. OY indications calculated from actual plant counts and measurements eliminate some of the biases found in the farmer reported yields.

Both regional and State level indications are produced from the OY data. Therefore, questionnaires and procedures for an OY crop are the same across the multi-state region. Regional level indications are derived by weighting the data for all OY States by harvested acres. They are used for analysis by the ASB in the same manner as the States review their indications. In some States, indications are also produced for ASD which are groups of counties with similar agricultural characteristics.

The OY surveys produce indications for harvested acres, yield, and production. Objective measurements (counts of plants, ears, pods, etc.) are made on small plots of land. At maturity, the small plots are harvested and yield is calculated based on the actual production taken from these small plots less an allowance for harvesting loss.

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 done on a field by field basis. The main focus is on verifying the subsampling frame by checking the acreages reported on the parent survey and recording any changes. Changes may be due to recording or reporting errors in the parent survey, failure to fulfill planting intentions, or switching to other utilizations. Any other data that must be obtained from the operator, for example planting date, are collected at this time. The final question asks for permission to enter the sample field and make counts and measurements. Ratio indications

comparing the initial interview acres to the parent survey are computed to determine if acreage revisions are in order.

Various counts for each plot are obtained each month until the crop is mature or harvested. Plant and fruit counts, fruit measurements, and maturity determinations are recorded each month. Early season data are fed into regression equations used to forecast gross yield and the components of yield, number of fruit, and weight per fruit. These forecasts are made using two approaches. The first approach applies forecast equations to sample (field) level observations to compute a yield forecast for each sample field, and averages these forecasts to the State level. This indication is called **OY B**. The second approach computes the State average of the raw counts and measurements and applies forecast equations derived from State level data. This is called the **OY X** indication. It is important to remember the difference between OY B and OY X since both use the same set of data. At maturity, the final visit obtains crop cutting data used to directly calculate final gross yield. The counts and measurements from all visits are added to the historical database used to derive future forecast equations.

Regional laboratories record measurements on fruit sent in from field enumerators. Lab samples are submitted for every sample hand harvested by enumerators. Lab measurements include weighing the fruit (ear, pods, bolls, or heads), weighing the grain after threshing, and determining moisture content. These data are obtained in a controlled environment using more accurate scales and moisture meters. The data are used to true up weights obtained in the field, calculate threshing fraction, and adjust to standard moisture. Lab measurements for wheat record spikelet and grain counts from "green" heads early in the season. These counts are used to forecast grain weight per head.

Models

Models are used extensively in forecasting and estimating crop yields and production. NASS uses models of similar structure for all OY crops and months of the growing season. It is important to remember that the term, model, refers to any mathematical equation used to represent the relationship between two or more variables and not to a class of estimators.

The general formula for forecasting or estimating the yield of any crop can be stated as:

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

where

Y = net yield per area,
F = average number of fruit per area,
W = average net fruit weight in standard units and moisture content, and
L = average harvest loss per area.

This general formula describes net yield *Y* as a function of three components. This formula can be applied at different levels of area, i.e., sample, district, State, or region.

The above model is then adjusted using a component to remove bias from indications. This formula can be stated as:

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

where

B = a bias adjustment. *B* can be as simple as a straight average of the bias from the previous 10 years, a more sophisticated statistical measure, or more subjective measure of say, the average bias from similar crop years. This bias also varies by crop, State, and region.

NASS conducted 11 corn validation studies from 1954 through 1983. A majority of these studies, conducted to examine relationships between objective survey estimates and actual yield of corn, showed an unexplained difference of between 2.0 and 4.8 percent. However, differences between the objective survey estimates and the official final estimated yields for a region of 10 major States generally were between 6 and 12 percent. The principal recommendation from these studies was that the official estimate be adopted from final average yield for the 10 State region, adjusted for non-sampling errors, as its final estimated yield of corn for grain for that region.

NASS uses two different statistical modeling approaches for forecasting and estimating yields and the components. One approach is to use sample level models for calculating the components of the general yield formula. The other approach is to use models at the State and regional level for calculating the components of the yield formula. General model structure is discussed further in the ***Objective Yield Indications*** section. Crop specific models are discussed in detail in the individual crop chapters.

Objective Yield Indications

The OY sample level data are used to produce several different yield indications. These indications are:

1) the Field Level Forecast,
2) the Farmer Reported Field Yield for Sample Field,
3) the Field Level Indication Regressed to Final Official Estimate of Yield,
4) the Farmer Reported Yield Indication Regressed to Final Official Estimate of Yield, and
5) the State Average Counts Regressed to Final Official Estimate of Yield.

The general structure of each indication is consistent across crops and discussed in this section.

Crop specific items, such as model forms and data items used in each indication, are discussed in the appropriate crop section.

Field Level Statistics

The row counts and measurements collected in each field are the basis for the field level indication. The indicated State average net yield is defined as follows:

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

where

OYB= State average net yield,
G =State average gross yield, and
L = State average harvest loss

State Average Gross Yield

For most State average gross yields, *G*, the straight average of sample level gross yields is appropriate because the sample design allows each acre an equal chance of being selected. However, for some States, the sample design allows each acre within district an equal chance of being selected, but acres may be sampled at rates different in different Districts. Districts may be based on cropping practices, such as irrigated and non-irrigated acreage, or geographical. These districts are different from Agricultural Statistics Districts discussed elsewhere in this paper. In these cases, district average gross yields, G_d, are calculated using the straight average of the sample level gross yields. District averages are then weighted together using JAS area frame crop acreage indications to produce the State average gross yield:

$$\bar{G} = \sum_d \bar{G}_d \, a_d$$

where

d indexes the districts and
a is the acreage weight derived:

$$a_d = \frac{P_d}{\sum_d P_d}$$

where

P is the JAS area frame planted acreage indication for the crop of interest.

The district average gross yields are the straight average of sample level gross yields:

$$\bar{G}_d = \sum_{i}^{n_d} \frac{g_{di}}{n_d}$$

where

I indexes the samples,
g_{di} = the gross yield for sample *I* in district d, and
n_d = the number of samples within district *d*.

Sample level gross yields, g_{di}, are a product of the number of fruit and average fruit weight components:

$$g_{di} = f_{di} w_{di}$$

where

f = the sample level number of fruit per area and
w = the average net fruit weight in standard units and moisture content.

During the growing season, *f* and *w* are forecasts. Sample level models are used to calculate *f* while *w* is calculated using either historical averages or simple linear regression models, depending on the sample's maturity stage. Models for *f* and *w* differ for each crop and will be presented in detail in crop specific chapters.

In early maturity stages, historical average fruit weights are calculated by averaging the district final average fruit weights for the 5 previous years:

$$w_d = \sum_{t=1}^{5} \frac{w_{dt}}{5}$$

where

t indexes the previous 5 years.

After the crop is at or nearly mature, enumerators harvest and weigh fruit from the samples. In these cases, the actual number of fruit harvested and the average weight per fruit are calculated for each sample. In earlier months, samples will fall in different maturity categories. Separate forecast equations are derived for each maturity category within month within State. A sample's maturity category determines which equation is employed. The computed average yields will be forecasts from multiple maturities and, in later months, include yields calculated from final harvest data.

Since the models used to determine the gross yield indication are applied separately to fruit count and fruit weight, each can be isolated and analyzed individually which broadens the scope of the

review. The commodity statistician can obtain a better understanding of how the yield figure is coming together and how this year's data relate to previous years'. Is a near record yield forecast a result of very high fruit count and an average yield or vice versa? Analysts can also examine how each model is influenced by extreme conditions. Computationally, a State or district average forecast is the simple average of the sample forecasts in the State or district.

The above discussion focuses on modeling at the sample level and averaging to the aggregate. A second approach has been developed, using the same data, that averages the raw counts and measurements and uses models developed at the aggregate level. For example, an early season average number of corn ears per acre, an average ear length, and a calculated interaction term provide candidate independent variables for a State level forecast model. The resulting indication is discussed further at the end of this chapter.

Harvest Loss

District average harvest loss ($\bar{L}_d$) is a straight average of sample level harvest loss from one-fourth of the samples:

$$\bar{L}_d = \sum_{i \in n_{dl}}^{n_{dl}} \frac{l_{di}}{n_{dl}}$$

where

I indexes the samples
l_{di} = the loss for sample *I* in district *d*, and
n_{dl} = the number of samples in district *d* with harvest loss data.

Data collected for harvest loss consist of gleaning samples of fruit left in field after harvest. Prior to harvest, the harvest loss component is either the historical 5-year average harvest loss :

$$\bar{L}_d = \sum_{t=1}^{5} \frac{L_{dt}}{5}$$

or the based on the net yield as a percent of the gross yield:

$$\bar{L}_d = \bar{G}_d \frac{\sum_{t=1}^{5} \left(1 - \frac{\bar{L}_{dt}}{G_{dt}} \right)}{5}.$$

State average harvest loss $\bar{L}$ is calculated by weighting the $\bar{L}_d$ by a_d, the same procedure used to calculate *G*.

Regional indications are calculated using State level yield indications weighted by the estimated harvested acres.

Regressed to Board Indications

The interpretation process is dependent on the historical data relationship of the survey indications and the final estimate. The treatment of biases that may exist using average differences assume bias is a constant. These biases can also be addressed using simple regression models where the final yield is the dependent variable. Different independent variables can be employed to develop models. The field level (Y) and farmer reported (F) yield indications can be regressed directly to the final estimate to get unbiased forecast equations. For the OY X indication, average raw counts can be regressed to the final yield estimate to arrive at another forecast of net yield. Separate models are developed for each State and region for every month of the growing season. The regression model can be expressed (omitting subscripts for month and State) as:

$$\hat{Y} = a + bX$$

where

$\hat{Y}$ is the current forecast,
X is the current value for the independent variable, and
a and b are regression coefficients.

Model parameters are estimated using up to 15 years of data. Outlier detection is done using the studentized deleted residual (see Neter, Wasserman, and Kutner [2], page 406). Observations with a studentized deleted residual value outside plus or minus 3 are not used for estimating model parameters.

Acreage Indications

Acreage adjustment ratios account for changes in acreage from the time of the base survey until harvest. Besides data from the base survey, acreage adjustment ratios use data from the initial interviews and from the monthly field counts. Acreage adjustments are not the main purpose of OY Surveys but are a by-product. Thus, these adjustments 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. For potatoes (list sample), the R1 ratio is calculated on the basis of "all land operated" rather than the tract.

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 active samples less any current "lost" samples, and the denominator is the total number of active samples. An active sample is where harvest has occurred or is expected to occur.

   Table 1 below identifies the ratios used in adjusting acreage by crop throughout the season.

Table 1: Acreage Adjustment Ratios, by Crop

Crop	First Month	Subsequent Months
Winter wheat	R1	---
Spring Wheat	R1	----
Corn	R1	Abandonment
Cotton	R1 R2	Abandonment
Soybeans	R1	Abandonment
Potatoes	R1	----

Strengths and Weaknesses of Each Model

The strengths of the field level models are that there is a separate model for each component (plants, pods per plants for soybeans) at each level of maturity. This allows for a high level of complexity in modeling the data. Also, the frequency with which each model is used depends on which maturity models are used more. This approach is self adjusting for early and late seasons. The weakness is that sample level data is highly variable, both for measurements and final sample level values, and these sample level component level models have a large error associated with them (that is, they are not very accurate for any one particular forecast).

The strength of the average counts approach is that by averaging thousands of observations together, the central limit theorem comes into play and the variability of the mean is greatly

reduced, both on the independent and dependent side. The disadvantage is that these models are simple one variable models with only 15 observations, and consequently are not at all complex. Also, early and late seasons must show up in the average counts since the model does not explicitly address early and late.

CHAPTER 5 CORN OBJECTIVE YIELD METHODS

This chapter presents the procedures and formulae used to calculate corn yield indications. The scope of the Corn 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.

Sample Design

Corn Objective Yield surveys are conducted in the ten major corn producing States: Illinois, Indiana, Iowa, Kansas, Minnesota, Missouri, Nebraska, Ohio, South Dakota, and Wisconsin. There are approximately 2,090 samples allocated to the States. Forecasts of acreage, yield, and production are made monthly from the August 1 Crop Report through the November 1 Crop Report with final estimates published in January.

Sample fields for Corn Objective Yield are selected from farms reporting corn planted or to be planted in the area frame of the JAS. The sample fields are selected with probability proportional to size, and the net effect is a self-weighting sample of areas of all corn for grain in each State. In Nebraska and Kansas, separate samples are selected from irrigated acres and nonirrigated acres with each being handled as if they were separate States. Data are collected from each sample at monthly intervals starting in late July and continuing through December or until the sample has been final harvested. Each month during the Objective Yield Survey, data collected from the sample fields are used to produce indications of planted acres (August only), acres for harvest, and yield.

A sample consists of two independently located units (or plots), each of which consists of two parallel 15 foot sections of row. 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. If ears have formed on the stalks, a sample of ears are measured for length and circumference. Just before farmer harvest, both units are hand harvested by the enumerator, weighed, and four ears are sent to a NASS lab where shelling fraction and moisture content are measured. Ears mailed to the lab are placed in plastic bags and sealed to preserve the moisture content from the time of field weighing. Final gross yield is computed from these data. The yield is measured as bushels of corn per acre at 15.5 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 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.

Forecasting and Estimating Number of Ears for Sample Fields

The sample number of ears per acre forecasts and estimates are influenced by two sub-components, the number of ears and the plot size of the sample in square feet. 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
- Average row space is the sum of the two 4-row space measurements divided by 8.

For immature forecasting categories, models used to forecast the final number of ears in the sample plots use data collected such as: number of stalks, stalks with ears, ears and ear shoots, and ears with kernels. The models vary depending on which variable has the best predictive value for each maturity.

Maturities categories 1-4 (no ear shoots, pre-blister, blister, and milk stages)

Two models are used to forecast the number of ears in each sample. Model 1, for maturity categories 1 through 4 (no ear shoots, pre-blister, blister, and milk stages), uses the current month's stalk count as the independent variable and the final number of ears as the dependent variable. This model has very high R-squares for each maturity category. Model 1 is:

$$\hat{Y}1_i = a + bX_i$$

where X_i is the number of stalks in both units.

Model 2, for maturity categories 2 through 4 (pre-blister through milk stage), uses a regression model of the ratio of the current month's count of stalks with ears or ear shoots to total stalks to predict the ratio of the current month's count of ears and ear shoots to the final number of ears. This predicted ratio is divided into the current month's count of ears and ear shoots to produce the Model 2 forecast of number of ears in the sample. Model 2 is not used for maturity category 1 (no ear shoots) since samples in this category have no ears or silked ear shoots. Model 2 has very high R-squares for pre-blister samples, but R-squares for maturity categories 3 and 4 (blister and milk stages) are not as high. Model 2 is:

$$\hat{Y}2_i = \frac{h_i}{a + b\,(t_i/S_i)}$$

where

h_i = the number of ears and of silked ear shoots in sample I
t_i = the number of stalks with ears and ear shoots in sample I
s_i = the number of stalks in sample I
a and b are regression coefficients developed from the relationship of the ratio of stalks with ears or silked ear shoots to total stalks with the proportion of ears and silked ear shoots to final ears.

This is a type of survival model in that it forecasts the number of ears that will develop and survive from the observed fruiting positions.

The forecasts from the two models for a given month and maturity category are weighted together to obtain a single forecast for the final number of ears for each sample. The weights are calculated from the R-squares of the models. Thus, the model which has the higher R-square will have more effect on the combined model.

$$\hat{Y} = w\hat{Y}1_i + (1 - w)\,\hat{Y}2_i$$

where

$$w = \frac{R_1^2}{(R_1^2 + R_2^2)}$$

and

R_1^2 and R_2^2 are the R^2 values from the forecast equations for Models 1 and 2, respectively.

Maturity categories 5-7 (dough through mature stages)

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

The sample average grain weight per ear is always converted to a standard 15.5 percent moisture. Models used to forecast or estimate the sample average grain weight per ear differ by maturity. All unharvested samples in maturities 1 and 2 (no ear shoots and pre-blister stage) use a 5-year historical average ear weight. Maturities 3-6 (blister through dent stage) employ models derived using kernel row length and ear diameter. Enumerator harvested samples use the average field weight per ear, shelled grain weight per ear, and the moisture content to estimate the grain weight per ear at 15.5 percent moisture.

Maturity categories 1 and 2 (no ear shoots and pre-blister stage)

Because of the immature stage of ear development, samples in maturity categories 1 and 2 (no ear shoots and pre-blister stage) use one of two 5-year historical average grain weight per ear (pounds at 15.5 percent moisture). Two averages are computed for each State (in Nebraska, which has two irrigation districts, two averages are computed for each district). For the August 1 survey, this average is computed from all samples with final lab grain weights during the last 5-years. For the September 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 September 1 or later surveys from the last 5-years. This average is based on few samples (usually 10-30) and is rarely used.

Maturity categories 3-6 (blister through dent stage)

Samples in maturity categories 3 through 6 (blister through dent stage) that have not been enumerator harvested use regression models to forecast grain weight per ear. The current model

uses the average length of the kernel row from the first five ears beyond a row of one unit. A second model is being developed that uses a derived volume as the independent variable. This volume variable is an interaction term computed from the average kernel row length and the average diameter. When an adequate dataset is built, this model could be used in conjunction with the current model or even replace it. A third model being evaluated uses the average maturity code 6 (dent stage) ear weight from the first five ears beyond a row of one unit as the independent variable. This model can only be used when the sample has matured to the dent stage and is not enumerator harvested.

The general form of any model is:

$$\hat{wt}_i = a + b\ X_i$$

where

X_i is the average kernel row length in sample I or the computed volume measurement of the ears in sample I.

Parameter estimates are calculated for each maturity category in each month for each State (or district).

Maturity category 7 (mature)

Mature samples and enumerator harvested samples use the average field weight per ear adjusted to the standard definition using measurements from ears mailed to a lab. A sample is enumerator harvested when there are three or more ears beyond a unit that are mature, the farmer intends to harvest the sample within 3 days, or it is the final visit before the survey cutoff date. Note that a sample need not be mature to be enumerator harvested. Many samples in dent stage and some in dough stage are harvested by enumerators.

The average field weight per ear is an average of the combined ear weight (cob and kernels) from all ears harvested in sample i:

$$\text{Field Wt} = \frac{\text{Total wt of ears}}{\text{count of ears}}$$

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:

$$\text{wt/ear} = \text{Field wt} * \left[\frac{w_s}{w_4 - b}\right] * \left[\frac{1 - m_i/100}{.845}\right]$$

where, for sample I,

w_s = the weight of all grain shelled from four ears
w_4 = the weight of four ears (including cob), plastic bags and rubber bands (as mailed)
b = the weight of plastic bags and rubber bands
m_i = the moisture content of the shelled grain, and
.845 = (100 - 15.5/100), the 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.

Forecasting Yield for Sample Fields

The gross yield for sample I is calculated by:

$$G_i = \frac{F_i * W_i}{56}$$

where
F = ears per acre
W = average grain weight per ear in pounds at 15.5 percent moisture
56 = converts bushel per acre

Both components, F and W, may be a forecast ($\hat{Y}_i$ and wt_i) or actual crop cutting data.

State Average Forecasts and Estimates

The State average gross yield is the average of the computed gross yields for all the sample fields. No weighting is required because the sample fields have been selected with probabilities proportional to size.

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 of sample i.

The standard error of the estimate is:

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

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 = $\sum$ (Sample Field Stalks per Acre) / N_G

State Average Ears per Acre = $\sum$ (Sample Field Ears per Acre) / N_G

State Average Harvest Loss = $\sum$ (Sample Field Harvest Loss) / N_G

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 =
$\sum$ (sample field grain weight * sample field ears per acre) / $\sum$ (sample field ears per acre)

For average ears per acre and average grain weight per ear, forecast values are used for those samples not yet harvested.

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 difference 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 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.

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:

Data items used to estimate or forecast number of ears, weight per ear, and harvest loss, by maturity.

Maturity	Number of Ears	Weight per Ear	Harvest Loss
1 No ears or ear shoots	Stalks	5-year average	5-year average[1]
2 Pre-blister	Stalks Stalks with ears or ear shoots Ears and ear shoots	5-year average	5-year average[1]

Data items used to estimate or forecast number of ears, weight per ear, and harvest loss, by maturity.

Maturity	Number of Ears	Weight per Ear	Harvest Loss
3 Blister	Stalks Stalks with ears or ear shoots Ears and ear shoots	Kernel row length	5-year average[1]
4 Milk	Stalks Stalks with ears or ear shoots Ears and ear shoots	Kernel row length	5-year average[1]
5 Dough	Ears with kernels	Kernel row length	5-year average[1]
6 Dent	Ears with kernels	Kernel row length Grain weight per ear	5-year average[1]
7 Mature	Ears with kernels	Grain weight per ear	5-year average[1] or harvest loss if available
Final	Ears with kernels	Grain weight per ear	Harvest loss

[1] 5-year average of (net yield/gross yield), not harvest loss.

Forecasting Directly to the 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, using the same data, to forecasting State yield can be applied by doing the averaging first and the modeling last. For each of the count variables, (stalks and ears), an average per acre at the State level can be calculated. Average weight per fruit can also be calculated, weighting the average weight per ear in each sample by the number of ears per acre 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 ears per acre, and final weight per ear. The distinction is State and regional averages are used as independent variables in regression models that predict State and regional level final yields, ears per acre, and weight per ear. In these models, 1-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. They forecast State and regional level indications, not sample level indications as described in the previous sections.

Independent and dependent variables used by each State and Corn OY Region, by month.

August		
State	Dependent Variables	Independent variables
OY Region, Illinois, Indiana, Iowa, Nebraska	Official yield	(Stalks with ears+ears with kernels)*(average kernel row length)
	Final Number of Ears	Stalks per acre
	Final Grain Weight (lbs.)	Average kernel row length
	Final Harvest Loss	5-year average
Minnesota, Ohio, Wisconsin	Official Yield	Stalks per acre
	Final Number of Ears	Stalks per acre
	Final Grain Weight (lbs.)	Stalks per acre
	Final Harvest Loss	5-year average
September		
OY Region, Illinois, Indiana, Iowa, Minnesota, Nebraska, Ohio, Wisconsin	Official Yield	(Average number of ears)*(average kernel row length)
	Final Number of Ears	Average number of ears per acre
	Final Grain Weight (lbs.)	Average kernel row length
	Final Harvest Loss	5-year average
October		

OY Region, Illinois, Indiana, Iowa, Minnesota, Nebraska, Ohio, Wisconsin	Official Yield	Indicated net yield
	Final Number of Ears	Average number of ears per acre
	Final Grain Weight (lbs.)	Average kernel row length
	Final Harvest Loss	5-year average
November		
OY Region, Illinois, Indiana, Iowa, Minnesota, Nebraska, Ohio, Wisconsin	Official Yield	Indicated net yield
	Final Number of Ears	Average number of ears per acre
	Final Grain Weight (lbs.)	Average kernel row length
	Final Harvest Loss	Indicated harvest Loss

Final net yield

The final net yield indication is based on the following formula:

FINAL NET YIELD = FINAL GROSS YIELD - FINAL HARVEST LOSS

The final gross yield indication is calculated using data collected from sample fields shortly before farmer harvest. This final enumeration of the sample field is also known as crop cutting. The enumerator harvests all the ears of corn in the sample units and weighs them. A subsample of ears is sent to a lab to determine moisture content and shelling fraction. These data are used to estimate ears per acre and grain weight per ear for the sample field. Ears per acre and grain weight per ear can be combined to calculate gross yield per acre, as shown previously. A straight average of the sample field gross yields is an indication of the State average gross yield.

Harvest Loss (gleanings)

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) * ∑ (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, State average harvest loss is calculated using data from the current year:

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

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

$$\text{and standard error} \quad S_L = \sqrt{\sum_{1}^{N_L} \frac{(L_i - \bar{L})^2}{N_L(N_L - 1)}}$$

where

L_i = harvest loss in sample i

N_L = number of samples with Form E data.

$$L_i = \frac{(w_e + 2w_g)\ (1 - (m_i/100)\ (43{,}560)}{(\text{average row space})\ (60)\ (453.6)\ (56)\ (.845)}$$

where

w_e = weight of ears between Row 1 and Row 3
w_g = weight of grain between Row 1 and Row 2
average row space = the sum of the two 4-row space measurements divided by 8.
m_i = the moisture content of the shelled grain for the harvest loss sample
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.

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:

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

where

$\bar{G}$ and $\bar{L}$ = were defined previously

The standard error of the estimate is:

$$S_Y = \sqrt{S_G^2 + S_L^2 - \frac{2}{N_G}COV\ (G,L)}$$

where

S_L was defined previously, and

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

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

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:

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

with standard error:

$$S_P = \sqrt{(A^2_{HARV})(S^2_Y) + (Y^2)(S^2 A_{HARV}) + (S^2_Y)(S^2 A_{HARV})}$$

Strengths and Weaknesses of Each Model

The strengths of the sample level models are that there is a separate model for each component (ears, grain weight per ear) at each level of maturity. This allows for a high level of complexity in modeling the data. The weakness is that sample level data is highly variable, both for the measurements and the final sample level values, and these sample level component level models have a large error associated with them (i.e., they are not very accurate for any one particular forecast).

The strength of the OY X approach is that by averaging thousands of observations together, the central limit theorem comes into play and the variability of the mean is greatly reduced, both on the independent and dependent side. The disadvantage is that these models are simple one

variable models with only 15 observations, and consequently are not at all complex.

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
Stalks with ears or ear shoots	76
Ears and ear shoots	89
8-row space width (ft)	20.3
Average kernel length (in.)	5.8
W, the weighting variable for the number of ears Model 1	0.52

Suppose regression models for samples in blister stage are:

Ears Model 1 = 7.2 + 0.87 (stalk count)
Ears Model 2 = [(Ears and Ear shoots)] /
[1.2 + 0.09 * ((stalks with ears or ear shoots) / stalk count)]
Ears combined model = w * Ears Model 1 + (1-w) * Ears Model 2

where $$w = \frac{R_1^2}{(R_1^2 + R_2^2)}$$

and R_1^2 and R_2^2 are the R^2 values from the forecast equations for models 1 and 2, respectively.

Grain Wt = 0.23 + 0.02 (kernel row length)

Then

Ears Model 1	=	7.2 + 0.87 (81) = 77.67 ears
Ears Model 2	=	89 / [1.2 + 0.09 * (76 / 81)] = 69.29
Ears combined model	=	(0.52)(77.67) + (0.48)(69.29) = 73.65 Ears
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)

Then

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)

Then

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 the procedures and formulae used to calculate soybean yield indications. The scope of the Soybean 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.

Sample Design

Soybean Objective Yield surveys are conducted in eight major soybean producing States; Arkansas, Illinois, Indiana, Iowa, Minnesota, Missouri, Nebraska, and Ohio. There are over 1,300 samples allocated to the States. Forecasts of acreage, yield, and production are made monthly from the August 1 Crop Report through the November 1 Crop Report with final estimates published in January.

Sample fields for Soybean Objective Yield are selected from farms reporting soybeans for harvest in the area frame of the JAS. The sample fields are selected with probability proportional to area, and the net effect of the sample design is a self-weighting sample of areas of all planted soybeans in each State. 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 (August only), acres for harvest, and yield.

A sample consists of two independently located units (or plots), each of which consists of two parallel 3.5 foot sections of row partitioned into a 3-foot section and a 6-inch section. 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 harvest, the beans from the sample plots are weighed to determine the final yield from that sample. Plant counts are made in the full unit while detailed fruit counts are limited to a small 6-inch section at the end of each row, which usually consists of 1 - 4 plants. All 3.5 feet of each row is picked and weighed at harvest to establish gross yield. The yield is measured as bushels of beans per acre at 12.5 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 and number of pods. 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 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-inch 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

To forecast each sample's yield per acre, regression models are developed by forecasting category for each survey month. In the field, the enumerators classify each unit into one of four maturity categories. The field categories are:

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

Originally, six categories were used. Experience over several years indicated categories 1 (no pods) and 6 (mature) were not needed.

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

0. No plants are present in either row of the 6-inch section.

1. Field maturity 2, no pods with beans are present and the ratio of total fruit to main stem nodes is less than 0.20.

2. Field maturity 2, no pods with beans are present in the 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 are present in the 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.

Sample Level Yield Forecasts

The models constructed for each forecasting category forecast the number of plants per 18 square feet and the number of pods with beans per plant, for each unit. The third component of yield -- weight of beans per pod with beans (hereafter referred to as simply "bean weight per pod") -- is forecasted using an historical average. The most recent 5 years of data are used to derive the regression models and the historical average bean weight per pod.

The three components are multiplied to give a unit-level forecast of gross yield in bushels per acre, where a bushel is defined as 60 pounds of beans adjusted to 12.5 percent moisture.

If regression models are unstable from year to year (usually caused by a very small sample size for a forecasting category), or missing for certain forecasting categories and survey months, models are substituted from an adjacent forecasting category in the same month or from another month for the same category. If the 5-year historical average bean weight per pod is unstable due to an unusual year, then the unusual year may not be included in the historical average.

A separate forecasting category is determined for each unit and forecasts of the number of plants per 18 square feet and pods with beans per plant are made for each unit. This is necessary because the development of soybeans can be quite different for the two units in a sample. The forecasting categories assigned by the summary are given in Table 1.

Analysis of Raw Data

When the parameter estimates for the sample level models are created, certain observations are excluded as outliers. Often historical datasets contain extreme and unusual counts. These are viewed as data aberrations which falsely influence the parameter estimates. Statistically, these are defined to be observations that have an rstudent value greater than 3 or less than -3. For a discussion of the rstudent statistic, see Belsley, Kuh, and Welsch [4]. An rstudent greater than the absolute value of three basically means that, if the observation were used in a forecast equation derived without that observation, the difference between the observation and the prediction would be greater than 3 prediction standard errors.

Forecasting Number of Plants per 18 Square Feet

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

$$Y = b_o + b_1X$$

The independent variable, X, is the total plant count in the 3-foot and 6-inch sections of a unit. These counts are obtained in a preharvest visit and expanded to 18 square feet. The dependent variable, Y, is the final plant count in the same area, also expanded to 18 square feet. The model parameters are estimated from the last 5 years' data.

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.

Forecasting Number of Pods with Beans per Plant

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

V1 = Plants per 18 square feet (the same variable used to forecast the 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:

$Y = b_o+b_1V1+b_2V2+b_3V3+b_4V4+b_5V5$

where three or four of the coefficients (b_1,b_2,b_3,b_4,b_5) are zero. Again, the model coefficients are estimated from the last 5 years' data. Y is the final (at-harvest) number of pods with beans per plant in the 6-inch section.

The table on the following page shows which variables are used for each forecasting category.

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.

If a negative number of pods is forecasted, the forecast is set to zero.

Forecasting Bean Weight per Pod

A 5-year average bean weight per pod in grams at 12.5 percent moisture is used for all forecasting categories, except 10, which uses the actual mature bean weight per pod for each sample.

All States use separate historical averages for "wide" row units (row width 1.5 feet or more or blank) and narrow row units (row width less than 1.5 feet or broadcast).

Forecasting Gross Yield

Gross yield for a unit is forecasted by multiplying the forecasts of the number of plants per 18 square feet, the number of pods with beans per plant, and the historical average bean weight per pod, and converting this forecast to bushels per acre. Unit-level gross yields are averaged to the sample level; the resulting sample-level yields are averaged to obtain a forecast of State-level gross yield. As the season progresses, more and more of the unit-level yields are based on at-harvest data rather than forecasts. State-level gross yield is then an average of forecasted and at-harvest estimates.

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	Aug	Plants nodes	Plants laterals	Laterals fruit	Plants laterals	Laterals fruit	Laterals fruit				pods with beans
	Sep					fruit	fruit pods w/beans	pods w/beans	pods w/beans	pods w/beans	
	Oct							pods w/beans	pods w/beans	pods w/beans	
	Nov										
Weight per pod		5 year average									lab data

fruit = blooms + dried flowers + pods

Forecasting Directly to the 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, using the same data, to forecasting State yield can be applied by doing the averaging first and the modeling last. For each of the count variables, (plants, nodes, laterals, fruit, and pods), an average per acre at the State level can also be calculated. Average weight per fruit can also be calculated, weighting the average weight per pod in each sample by the number of pods 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 Board yield, final pods per 18 square feet, and final weight per pod. The distinction is State and regional averages are used as independent variables in regression models that predict State and regional level final yields, pods per 18 square feet, and weight per pod. 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. They are State and regional level models, not sample level models as described in the previous sections.

The following table shows the independent and dependent variables for the indications:

Dependent variable	Independent variable
Official Final Yield	August - Average number of lateral branches per acre September - average number of pods per acre October - December - average net yield per acre
Final Number of pods per acre	August - Average number of lateral branches per acre September - December - average number of pods per acre

Final weight per pod	August average number of lateral branches per acre September - average number of pods per acre October - December - average weight per pod
Final Harvest Loss	August - November - average of previous 5 years harvest loss December - Harvest Loss for current year

The Farmer Reported Yield Indication Regressed to Official Yield

The farmer reported yield obtained using the Post-Harvest Interview is averaged to the State level. This Post-Harvest Interview indication is also regressed to the final official yield to obtain an additional indication. In effect, this is a model for bias.

Gross Yield Estimate at Maturity

When the unit reaches maturity (forecasting category 10), gross yield is estimated by the product:

Number of pods with beans per 18 sq. ft.	X	Bean Weight per Pod	X	Conversion Factor	

As with forecasted gross yield, unit-level estimates of gross yield at maturity are averaged to obtain sample-level at-maturity gross yield. If one unit is not yet mature, its forecasted yield is averaged with the mature unit's estimated yield to obtain a sample-level yield indication.

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

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

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

(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 with beans 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:

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

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 of 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 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.

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:

$$L = \frac{(\text{weight of loose and threshed beans})(1-(\text{moisture content}/100))(43560)}{(0.875)(453.6)(60)(3)(\text{4-row space, Unit 1} + \text{4-row space, Unit 2})/(2)}$$

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

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

$$\text{and standard error} \quad S_L = \sqrt{\sum_{1}^{N_L} \frac{(L_i - \bar{L})^2}{N_L} (N_L - 1)}$$

where

L_i = harvest loss in sample i
N_L = number of samples with gleaning data.

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:

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

where

$\bar{G}$ $\bar{L}$ were defined previously

The standard error of the estimate is:

$$S_Y = \sqrt{S_G^2 + S_L^2 - \frac{2}{N_G} COV\ (G,L)}$$

where

S_L was defined previously, and

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

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

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 P for the State is the product of estimated State-level net yield and acres to be harvested for beans:

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

with standard error:

$$S_P = \sqrt{(A^2_{HARV})(S^2_Y) + (Y^2)(S^2A_{HARV}) + (S^2_Y)(S^2A_{HARV})}$$

Strengths and Weaknesses of Each Model

The strengths of the sample level models are that there is a separate model for each component (plants, pods per plants) at each level of maturity. This allows for a high level of complexity in modeling the data. The weakness is that sample level data is highly variable, both for the measurements and the final sample level values, and these sample level component level models have a large error associated with them, (i.e., they are not very accurate for any one particular forecast).

The strength of the OY X approach is that by averaging thousands of observations together, the central limit theorem comes into play and the variability of the mean is greatly reduced, both on the independent and dependent side. The disadvantage is that these models are simple one variable models with only 15 observations, and consequently are not at all complex.

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:

$$X = \frac{(\text{plants in the 3-foot and 6-inch sections})(18)}{(3.5)(\text{4-row space width})/(2)}$$

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: (41+11)(18)

$$X = \frac{}{(3.5)(12.8)/(2)} = 41.8$$

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: $\hat{P}_1 = 1.2+(0.92)(41.8)=39.656$

Unit 2: $\hat{P}_2 = 1.2+(0.92)(40.3)=38.276$

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:

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

the forecast of pods with beans per plant for each unit is:

Unit 1: $\hat{Y}_1 = 42.2 - (0.6)(41.8) + (4.8)(5/11) = 19.30$

Unit 2: $\hat{Y}_2 = 42.2 - (0.6)(40.3) + (4.8)(2/9) = 19.09$

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 general formula for computing yield per acre based on each unit's data is:

$$\text{Yield} = \frac{(\hat{P})\ (\hat{Y})\ (0.437)\ (43{,}560)}{(18)\ (453.6)\ (60)}$$

where

$\hat{P}$ = predicted number of plants per 18 square feet
$\hat{Y}$ = predicted pods with beans per plant
43,560 = square feet per acre (convert to acre basis)
18 = standard size of unit
453.6 = grams per pound (converts to pounds)
60 = pounds of beans per bushels

Then the gross yield estimates for the two units are:

Unit 1: $\frac{(39.656)(19.30)(0.437)(43560)}{(18)(453.6)(60)} = 29.74$ bu/acre

Unit 2: $\frac{(38.276)(19.09)(0.437)(43560)}{(18)(453.6)(60)} = 28.39$ bu/acre

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 surveys are conducted in major cotton producing States; Arkansas, California, Georgia, Louisiana, Mississippi, North Carolina and Texas. There are over 1,300 samples allocated to these States. Forecasts of yield and production are made monthly from the August 1 Crop Report through the January 1 Crop Report with final estimates published in May.

Sample fields for Cotton Objective Yield are selected from farms reporting cotton planted in the area frame sample of the JAS. The sample fields are selected with probability proportional to size, and the net effect is a self-weighting sample of areas of all cotton in each State. Texas is further divided into two geographic districts, each of which is sampled separately. 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 (August 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 since the initial Form-B was completed.	

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:

$$Y = B0 + B1*X1 + B2*X2 + B3*X3$$

where:

Y = forecasted number of large bolls in ith unit
X1 = observed number of burrs, open bolls, partially open bolls, and large unopened bolls (40-foot equivalent) in ith unit
X2 = observed number of small bolls and blooms (40-foot equivalent) in ith unit
X3 = observed number of squares (40-foot equivalent) in ith unit
B0-B3 = 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

One model is used to forecast boll weight for all maturity categories in a district in a State. Early in the year (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. The following model is used during the season, when between 20 and 85 percent of the projected number has been picked and weighed:

$$BW = W * (A + BX)$$

where:

A and B are regression coefficients
BW = forecast boll weight
W = observed boll weight to date
X = ratio of bolls picked and weighed to large bolls forecasted

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 .5	4 ratio .5 to 2.0	5 ratio 2.0+	6 harvested or 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

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

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. For each of the count variables (plants, squares, small bolls and blooms, large unopen bolls, and open bolls), an average per acre at the State level can also be calculated. 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:

$$GY = (2.401 * LSR * LB * BW) / RS$$

where

GY	=	Gross Yield (in lbs. of lint per acre)
LSR	=	Lint/Seed Ratio (preceding 3-year average)
LB	=	number of large bolls at harvest (on a 40-foot basis)
BW	=	average boll weight (in grams at 5 percent moisture, gin equivalent)
RS	=	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.

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 of 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:

$$HL = (2.401 * WT * LSR) / RS$$

where

HL	=	Harvest Loss (lbs. of lint per acre)
WT	=	weight of cotton left in units 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)
LSR	=	Lint/Seed ratio
RS	=	row space measurement
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

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

$$\text{and standard error} \quad S_L = \sqrt{\sum_{1}^{N_L} \frac{(L_i - \bar{L})^2}{N_L(N_L - 1)}}$$

where

L_i = harvest loss in sample i
N_L = number of samples with Form E data.

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

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

where

$\bar{G}$ $\bar{L}$ were defined previously.

The standard error of the estimate is:

$$S_Y = \sqrt{S_G^2 + S_L^2 - \frac{2}{N_G} COV\ (G,L)}$$

where

S_L was defined previously, and

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

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

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

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

with standard error:

$$S_P = \sqrt{(A^2_{HARV})(S^2_Y) + (Y^2)(S^2A_{HARV}) + (S^2_Y)(S^2A_{HARV})}$$

Strengths and Weaknesses of Each Model

The strengths of the sample level models are that there is a separate model for each component (large bolls, weight per boll) at each level of maturity. This allows for a high level of complexity in modeling the data. The weakness is that sample level data are highly variable, both for the measurements and the final sample level values, and these sample level component level models have a large error associated with them, (i.e., they are not very accurate for any one particular forecast).

The strength of the State level approach is that by averaging thousands of observations together, the central limit theorem comes into play and the variability of the mean is greatly reduced, both on the independent and dependent side. The disadvantage is that these models are simple one variable models with only 15 observations, and consequently are not at all complex.

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

LB = burrs + open bolls + partially open bolls + large unopened bolls within unit
p = number of plants

LB/p = [(113+20) + (130+50) + 48 + 121] / 87 = 5.54

So, the maturity category is 5 (ratio > 2.00).

Forecast Number of Bolls

Multiple Regression Model

NB®) = # bolls = B1 + B2*X1 + B3*X2 + B4*X3

where,

X1 = burrs and large bolls (40-ft. equiv.)
X2 = small bolls and blooms (40-ft equiv.)
X3 = square (40-ft equiv.)

let,

B1 = 14
B2 = .933
B3 = .300
B4 = .110

These are regression coefficients derived from previous 5 years of sample level data.

Since burrs and open bolls and 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),

X1 = (40/46) * (all large bolls)

= (40/46) * ((113+20) + (130+50) + 48 + 121 + (33+14) + (27+11))

= (40/46) * 567

= 493.043

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

X2 = (40/6) * (4+6)

= 66.67

Since squares are counted in six feet of row (both 3-foot units),

X3 = (40/6) * (2+1)

= 20.000

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

NB®) = 14 + (.933 * 493.043) + (.300 * 66.67) + (.110 * 20.000)

= 496.210

Forecast Boll Weight

BW = W * (A + B * X)

where:

W = accumulated weight of seed cotton picked (adjusted for moisture content) divided by the accumulated number of open bolls picked

X = accumulated number of open bolls picked divided by the forecast number of large bolls (that is., this is the proportion of forecast large bolls picked by the enumerator)

A and B are regression coefficients.

For this example, let:

A = .882
B = .131

To determine BW for the current month:

Drying ratio = Dry weight / Wet Weight

= 52 /56 = .9286

Current month's weight picked = 650 * .9286

= 603.590

Current month's weight (at 5% moisture) = 603.590 * 1.0526

= 635 grams

where 1.0526 is the conversion factor to 5% moisture (gin equivalent).

So,

W = (257 + 635) / (50 + 130) = 4.956

X = 180 / 487 = .370

and,

BW = W * (A + B * X)

= 4.956 * (.882 + .131 * .370)

= 4.611 grams per boll

Forecast Gross Yield per Acre

Using the formula described previously, the estimated gross yield for this sample is:

GY = (2.401 * LSR * NB * BW) / RS

= (2.401 * .368 * 496.210 * 4.611) / 3.225

= 626.86 pounds of lint per acre

The average estimated gross yield across all samples in a district less an estimate of harvest loss produces the district estimate of net yield.

CHAPTER 8 WHEAT OBJECTIVE YIELD METHODS

This chapter presents the procedures and formulae used to calculate wheat yield indications. The scope of the Wheat 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.

Sample Design

Wheat Objective Yield surveys are conducted for three major classes of wheat: winter, durum, and other spring. Each is treated as a separate survey, however, they have identical methodologies. Winter Wheat Objective Yield surveys are conducted in the 10 major winter wheat producing States: Colorado, Illinois, Kansas, Missouri, Montana, Nebraska, Ohio, Oklahoma, Texas, and Washington. Other spring wheat is measured in Minnesota, Montana, and North Dakota. The Durum Survey is done in North Dakota only. There are approximately 1,410 samples allocated to the winter wheat States, 320 to the spring wheat States, and 120 for durum. Forecasts of winter wheat acreage, yield, and production are made monthly from the May 1 Crop Report through the September 1 Crop Report with final estimates published in late September. The other spring and durum programs begin with the July 1 Crop Report and end with the late September Small Grains Annual Summary.

Sample fields for Winter Wheat Objective Yield are selected from farms reporting winter wheat planted for harvest as grain on the March Crops/Stocks Survey. Other spring and Durum fields are drawn from farms reporting wheat planted or to be planted on the June Area Survey. The sample fields are selected with probability proportional to size, and the net effect is a self-weighting sample of areas of all wheat for grain in each State. In Texas, separate samples are selected from two different geographic regions with each being handled as if they were separate States. Data are collected from each sample at monthly intervals 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 acres for harvest and yield.

A sample consists of two independently located units (or plots), each of which consists of three parallel 21.6 inch sections of row. Field enumerators use a random number of paces along the edge of the field and a random number of paces into the field to locate each unit. A steel frame with tines exactly 21.6 inches apart is slipped into the rows to delineate the units. At each visit, enumerators count all stalks and heads. If heads have emerged from the stalks, the enumerator clips heads from outside the units and sends them to a NASS lab where spikelets and grains are counted. Just before farmer harvest, both units are hand harvested by the enumerator and sent to the lab where threshing fraction and moisture content are measured. A final gross yield is computed from these data. The yield is measured as bushels of wheat per acre at 12 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 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 derived using data collected during the previous five years for each maturity in each month. The maturity definitions used by the enumerators are:

Maturity	Code	Definition
Pre-Flag	1	There is no swelling in the stalks and no flag leaf is present.
Flag or early boot	2	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.

Late boot or Flower	3	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.
Milk	4	The kernels are soft, moist, and filled with a milky liquid.
Soft dough	5	The contents of the kernels are soft and can be kneaded like dough.
Hard dough	6	The grain is firm and can be dented with the thumbnail, but not easily crushed.
Ripe	7	The grain is hard and breaks into fragments when crushed.

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

The forecasting procedures use one model for predicting the final number of heads and one or two models for predicting final head weight. The regression equations for these 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 all States, the most recent 5 years of historical data are used to develop the forecast models. For example, the count of the number of observed heads, emerged or in a late boot stage, is the independent variable for predicting the number of heads expected at harvest time for samples in the late boot, flower, or soft dough maturity categories.

The forecasts of number of heads and head weight are made using current month counts and measurements. Harvest loss in bushels per acre is based on a straight 10-year historical average early in the season and by an average of current gleanings after harvest begins.

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. The historical average head weight is weighted (with a weight of .2) together with this model indication (with a weight of the R-square) to create a forecasted 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 (Model 1) and the observed clip unit green weight per head of emerged and late boot heads (Model 2) 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. Net yield, however, can still be affected by factors which influence the harvest loss.

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 a standard moisture of 12 percent. These actual yield components are used to compute the final sample gross yield per acre.

Independent variables used in the forecasting models of yield components at the various stages of maturity are shown in the following table:

Field or Lab Variables Used for Forecasting Final Yield Components in Sample Fields

Maturity Category	Final Number of Heads		Final Weight of Heads	
	Model	Independent Variable	Model	Independent Variables
Pre-Flag	1	Number of stalks	1	Historical Average
Flag or Early Boot	1	Number of stalks	1	Historical Average
Late Boot or Flower	1	Emerged heads + heads in late boot	1 2	Fertile spikelets per head Historical Average
Milk	1	Emerged heads + heads in late boot	1 2	Grains per head Clip Unit Green Weight per head
Soft Dough	1	Emerged heads + heads in late boot	1 2	Grains per head 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 models have the following form:

$$Y_i = a + b\,X_i$$

where,

Y_i = number of heads or weight per head,
a = the number of heads or weight per head when X equals zero,
b = the change in number of heads or weight/head for each unit increase in x, and
X_i = the independent variable from current field counts or laboratory measurements: number of stalks, number of emerged plus late boot heads, number of fertile spikelets/head, grains/head, or weight/head.

The formulae for arriving at forecasted head number (Y_h), forecasted head weight (Y_w),

forecasted gross yield/acre (GY), final head number (Y_{fh}), final head weight (Y_{fw}), final gross yield/acre (GY), harvest loss (HL), net yield (NY), and standard error of the net yield (SE(NY)) are given below. The forecast equations and R^2 are computed from the five most recent survey years. Early season forecasts for number and weight of heads will be made using current survey data as the independent variable in the forecast equations. When the crop is mature, actual plant counts and measurements from the current year are used to calculate the sample yield.

Forecast number of heads (Y_h)

$$Y_h = a + b\,x$$

where

x = number of stalks, or
x = number of emerged and late boot heads

Forecast threshed grain weight/head (Y_w)

$$Y_w = \frac{R_1^2(Y_{w1}) + R_2^2(Y_{w2})}{R_1^2 + R_2^2}$$

where,

Y_w = Combined weight per head from forecast Models 1 and 2 weighted by R^2 values.
Y_{w1} = Forecast weight per head from Model l. 1/
Y_{w2} = Forecast weight per head from Model 2. 1/
R_1^2 = Multiple correlation coefficient for Model 1.
R_2^2 = Multiple correlation coefficient for Model 2.

1/ A 5-year historical average is used with an R^2 value of 0.2 is used for the following maturity categories: pre-flag, flag or early boot, and also for Model 2 for late boot or flower maturities.

Forecasting Yield for Sample Fields

Forecasted Gross-Yield (GY)

$$GY = (Y_h) * (Y_w) * \frac{(\text{conversion factor})}{(8 - \text{row width})}$$

where,

Y_h is the forecast number of heads,
Y_w is the forecast grain weight per head, and

The conversion factor = [(43560)(8)(12)] / [(6)(60)(453.58)(21.6)] = 1.186

where,

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.

Final Gross Yield (GY):

$$GY = (Y_{fh}) * (Y_{fw}) * \frac{(\text{conversion factor})}{(8 - \text{row width})}$$

where,

Final number of heads per sample Y_{fh} is the sum of emerged heads, detached heads and heads in late boot when the sample reaches the hard dough or ripe maturity categories.

Final weight per head Y_{fw}

$$Y_{fw} = \frac{(\text{threshed grain wt.}) * (1.0 - \text{grain moisture content})}{(\text{number of heads threshed}) * (.880)}$$

Calculations of State Average Yield and Yield Components

To forecast a State yield per acre, a series of regression equations is used to forecast the two components of yield for each sample. The two components are number of heads and weight of grain per head. These components are combined to give a forecast of bushels per acre for each sample. A bushel of wheat is defined to be 60 pounds of wheat at 12 percent moisture. Since fields are selected with probabilities proportional to acreage, the average of these individual sample yields provides a self-weighted forecast of yield per acre for the State. The forecast equations used for a sample depend on the maturity classification of the sample units.

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 of sample i.

The standard error of the estimate is:

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

Simple means are also appropriate for Heads per Square Foot and Harvest Loss. No weighting is required when calculating State level averages for these items.

State Average Heads per Square Foot = $\sum$ (Sample Field Heads per Square Foot) / N_G

State Average Harvest Loss = $\sum$ (Sample Field Harvest Loss) / N_L

where

N_L is the number of sample field gleanings.

The State average grain weight per head is calculated using a weighted mean. The weighting variable is the sample field Heads per Square Foot.

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

Net Yield (NY) = Gross Yield (GY) - Harvest Loss (HL)

Standard Error of the Net Yield (SE(NY)):

$$SE(NY)=\sqrt{VAR(GY)+VAR(HL)-2*COV(GY,HL)}$$

where

$$VAR(GY)=\sum_{i=1}^{N_b}\frac{(GY_i-GY)^2}{N_b-1} \qquad VAR(HL)=\sum_{i=1}^{N_e}\frac{(HL_i-HL)^2}{N_e-1}$$

and

$$COV(GY,HL)=\sum_{i=1}^{N_b}\frac{(GY_i-GY)(HL_i-HL)}{N_b-1}$$

Note: In the above derivations GY_i is the i^{th} sample level gross yield, (either forecasted or final) GY is the State (or district) average gross yield, HL_i is the i^{th} sample level harvest loss, HL is the State (or district) average harvest loss, N_b is the number of usable field count samples and N_e is the number if usable harvest loss samples. If fewer than five usable harvest loss samples are available, the summary considers

$$Var(HL)=Cov(GY,HL)=0$$

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.

Final Net Yield

The indicated final net yield uses the following formula:

FINAL NET YIELD = FINAL GROSS YIELD - FINAL HARVEST LOSS

The final gross yield indication is calculated using data collected from sample fields shortly before farmer harvest. This final enumeration of the sample field is also known as crop cutting. The enumerator harvests all the heads of wheat in the sample units and sends them to a lab for weight and moisture content determination. These data are used to estimate heads per square foot and grain weight per head for the sample field. Heads per square foot and grain weight per head can be combined to calculate gross yield per acre, as shown previously. A straight average of the sample field gross yields is an indication of the state average gross yield.

Harvest Loss (HL):

HL = [(threshed grain wt)*(1.0-grain moist. content)*(conversion factor)] / [(.880)* (8-row width)]

where,

conversion factor = 1.186, and is defined above, and
.880 = 1 - .120 converts to standard 12.0 percent moisture

Early in the season, the computed gross yields are converted to net yields by deducting the previous 10-year average harvest loss. When at least five post-harvest gleanings have been collected and summarized, the average of the current year harvesting loss is calculated. The State average net yield then becomes the average of the self-weighting sample gross yields over a State minus the average of the post-harvest gleanings.

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

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

$$\text{and standard error} \quad S_L = \sqrt{\sum_{1}^{N_L} \frac{(L_i - \bar{L})^2}{N_L(N_L - 1)}}$$

where

L_i = harvest loss in sample i

N_L = number of samples with gleaning data.

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

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

where,

$\bar{G}$ and $\bar{L}$ = were defined previously

The standard error of the estimate is:

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

where,

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

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

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

When less than five gleanings are completed, an 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:

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

with standard error:

$$S_P = \sqrt{(A^2_{HARV})(S^2_Y) + (Y^2)(S^2A_{HARV}) + (S^2_Y)(S^2A_{HARV})}$$

Forecasting Directly to the 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, using the same data, to forecasting State yield can be applied by doing the averaging first and the modeling last. For each of the count variables, (stalks and heads), an average per acre at the state level can be calculated. Average weight per fruit can also be calculated, weighting the average weight per head in each sample by the number of heads per square foot 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 heads per square foot, and final weight per head. The distinction is State and regional averages are used as independent variables in regression models that predict State and regional level final yields, heads per acre, and weight per head. 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 data set is used for these models. The models are simple one variable regression models. They forecast State and regional level indications, not sample level indications as described in

the previous sections.

The selection of independent variables is based on reliability and availability. The goal is to choose an independent variable that will forecast as accurately as possible. However, since many field counts are only made in specific maturity stages, not every variable is available every month. Model selection varies from State to State, and month to month. The following table lists the models used to forecast directly to the State and regional levels.

Winter Wheat Yield Models used in 2001

Forecast yields for each of these indications are computed by regressing the indication against the Official yield over the past fifteen years except for the Mean Yield Limited which uses a reduced number of years. The regression equation is:

Y = a + bx where Y = the Official state yield or the 7 state yield for the OY region yield and x = the indicated yield

Yield Models*				
May	**June**	**July**	**Aug**	**Final**
Mean Yield *KS, OK TX, REGION*	Mean Yield	Mean Yield	Mean Yield	Mean Yield
Mean Yield Limited *KS, OK TX, REGION*	Mean Yield Limited	Mean Yield Limited	Mean Yield Limited	Mean Yield Limited
		Farmer Reported Yield *IL, KS, MO, OK, TX*	Farmer Reported Yield *- all states except MT*	Farmer Reported Yield

* All states and region except where noted.

Mean Yield = each sample's yield is modeled, then the mean of all samples' yields is calculated for the state or region.

Mean Yield Limited = Mean Yield with a lesser number of years used in the model.

Special component Yield Models used in 2001		
May	**June**	**July**
Emerged & Late Boot Heads x Green Weight per Head *TX*	Emerged & Late Boot Heads x Fertile Spikelets *CO, IL, KS, OH, TX, WA, REGION*	Emerged & Late Boot Heads x Fertile Spikelets *MT*
Stalks *KS, OK*	Emerged & Late Boot Heads x Grains per Head *MO, OK, TX*	Emerged & Late Boot Heads x Green Weight per Head *CO, NE, OH, WA*
	Stalks *MT*	

Special Component = the means of the modeled heads and weight/head are calculated for the state or region, then multiplied to get a state or regional indication. In early months for KS, OK and MT, stalks are used alone.

Winter Wheat Component Models used in 2001

Final Heads per Square Foot	Final Weight per Head	Final Harvest Loss
Stalks per Sq. Ft. *May - KS, OK* *June - MT*	***Green Weight per Head*** *May - TX* *July - CO, NE, OH, WA*	***5 Year Average*** *May - KS, OK, TS* *June - all states, REGION*
Emerged and Late Boot Heads per Sq. Ft. *May - TX* *June - CO, IL, KS, MO, NE, OH, OK, TX, WA, REGION (excluding MT)* *July - CO, IL, KS, MO, MT, NE, OH, OK, TX, WA, REGION*	***Stalks per Sq. Ft.*** *June - MT*	***Current Loss*** *July - all states, REGION*
	Fertile Spikelets per Head *June - CO, IL, KS, NE, OH, WA, REGION(excluding MT)*	
	Grains per Head *June - MO, OK, TX*	

Spring and Durum Wheat Yield Models used in 2001

Forecast yields for each of these indications are computed by regressing the indication against the Official yield over the past fifteen years except for the Mean Yield Limited which uses a reduced number of years. The regression equation is:

Y = a + bx where y = the Official state yield or the 3 state yield for the Spring Wheat OY region yield and x = the indicated yield

Yield Models used in 2001		
Aug	**Sep**	**Final**
Mean Yield	Mean Yield	Mean Yield
Mean Yield Limited	Mean Yield Limited	Mean Yield Limited
	Farmer Reported Yield	Farmer Reported Yield

Mean Yield = each sample's yield is modeled, then the mean of all samples' yields is calculated for the state or region.

Mean Yield Limited = Mean Yield with a lesser number of years used in the model.

Strengths and Weaknesses of Each Model

The strengths of the sample level models are that there is a separate model for each component (heads, grain weight per head) at each level of maturity. This allows for a high level of complexity in modeling the data. The weakness is that sample level data is highly variable, both for the measurements and the final sample level values, and these sample level component level models have a large error associated with them (i.e., they are not very accurate for any one particular forecast).

The strength of forecasting directly to the State level is that by averaging thousands of observations together, the central limit theorem comes into play and the variability of the mean is greatly reduced, both on the independent and dependent side. The disadvantage is that these models are simple one variable models with only 15 observations, and consequently are not at all complex.

Yield Forecast 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 # of emerged heads + heads in late boot)

weight per head (Model 1) = .12 + .04 * (# 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:

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

so that in this example:

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

Therefore, with 8-row width of 6.4,

Gross Yield per Acre = [(# 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 # of emerged heads + heads in late boot),

weight per head (Model 1) = .59 + .003 * (# 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 = [(# 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 * (# of emerged heads + head in late boot),

weight per head (Model 1) = .33 + .02 * (# 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 = [(# 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 & 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:

# of emerged heads, detached heads, and heads in late boot = 350,

moisture content of enumerator harvested grain = 12%,
# 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)] / [(# of heads threshed) (.880)]

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

= .72

Assuming an 8-row width of 6.4,

Gross Yield Per Acre = [(# 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 POTATO OBJECTIVE YIELD METHODS

This chapter presents the procedures and formulae used to calculate potato yield indications. The scope of the Potato 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.

Sample Design

Potato Objective Yield surveys are conducted in the seven major potato producing states: Idaho, Maine, Minnesota, North Dakota, Oregon, Washington, and Wisconsin. There are approximately 1,400 samples allocated to the States. Estimates of acreage, yield, and production are made for the November 1 and December 1 Crop Report with final estimates published in January.

Sample farms for the Potato Objective Yield Survey are selected from farms reporting potatoes planted in the list portion of the JAS. The sample fields are selected with probability proportional to size, and the net effect is a self-weighting sample of areas of all potatoes in each State. In Idaho, Minnesota, and Oregon, samples are selected for geographic districts with each being handled separately. Similarly, North Dakota has samples for irrigated and non-irrigated acres. Data are collected from each sample just before farmer harvest.

A sample consists of two independently located units (or plots), each of which consists of one 20-foot section of row. 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. The number of hills are counted in each unit and three hills are hand harvested by the enumerator and all tubers 1 ½ inch in diameter or greater are weighed. Final gross yield is computed from these data. The yield is measured as hundred weight (cwt.) of potatoes per acre. Harvest loss is measured in separate units located near the monthly yield plots.

Data Collected

Field enumerators count and measure items within the units. Data items are used to measure the size of the unit, number of tubers, weight per tuber, 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 hills:
- Number of hills in each unit,

Data items used to forecast or estimate weight per hill:
- Weight of tubers from 3 hills

Data items used to estimate harvest loss:

Distance between two rows (one row middle)
Distance between five rows (four row middles)
Weight of tubers left in gleaning unit

Yield

Unlike other field crops in the Objective Yield program, yields are not forecasted early in the season using regression models. Observations on the crop are only made just prior to harvest or when the vines are dead and no further growth is possible. The Potato Objective Yield Survey, therefore, is a crop cutting survey. The yield is computed at harvest time for each sample in the survey. The components of net yield are:

1. Gross yield = (hills per acre * weight per hill)
2. Harvest loss

A gross yield is determined by unit for each sample by multiplying the number of hills per acre by weight per hill. These unit yields are then averaged to obtain the measure of gross yield for the sample. Gross yield is computed in this manner to account for variations which may exist in the components. These variations include uneven row spacing, or unusual hill populations, as well as other factors which affect tuber growth and development within a field. Harvest loss is determined from gleanings taken from every fourth sample.

Hill counts, made within the 20-foot count area of a unit, are used to compute hill population for each unit in a sample. The formula used to convert hill counts and row width measurements to hills per acre by unit is:

Hills per Acre = (Hills in unit*43,560) / (Avg. row space*20 ft row)

where,

43,560 is the number of square feet in an acre.

An average weight in pounds per hill for each unit is derived from a subsample of hills within the units. For each unit, the weight is the average weight of tubers dug from Hills 8, 9, and 10. For all States except Idaho and Maine, the weighing is done by the field enumerators. Weighing for the excepted States is performed in labs.

Average weights are calculated by unit using the following formula:

Wt. per Hill in lbs. = (Wt. of tubers from 3 hills in grams)/(3 hills*453.6)

where,

453.6 is the number of grams per pound.

The sample level indications of gross yield are computed by multiplying the number of hills per acre by the average weight per hill for each unit and then taking the average of the unit values.

Gross Yield = $(H_1W_1 + H_2W_2)/2$

where H_1 and H_2 are the number of hills per acre for Unit 1 and Unit 2, respectively, and W_1 and W_2 are the weights per hill for Unit 1 and Unit 2.

Harvest Loss

Harvest loss is the negative component of the net yield equation. This component represents the weight of potatoes left in the field after harvesting. Two units are gleaned in every fourth sample field to obtain the information required for estimating harvest loss. Gleaning takes place almost immediately after harvest and must be done within 3 days of digging because potatoes deteriorate rapidly in the open air and in many areas producers disk or plow the field right behind the digger. All whole potatoes 1 ½ inch or larger and all pieces are gleaned. Smaller potatoes are not considered as part of production since they seldom leave the field. If these small potatoes (less than 1 ½ inch in diameter) do get trucked from the field, they are culled out and never reach commercial channels.

Harvest loss is calculated using the formula:

Loss = (43,560 * wt. gleaned)/(2 units *3 ft * 6 ft * 453.6)

where 3 feet by 6 feet are the dimensions of each gleaning plot and 453.6 is the number of grams per pound.

The difference of gross yield and harvest loss is the net yield indication. The indications of gross yield and harvest loss at the district level are the average of the sample level data. These district level indications are weighted to the State level using the current OY indication for harvested acres as weights. Districts represent geographical areas in Idaho, Minnesota, and Oregon. In North Dakota, the two districts represent irrigated and non-irrigated acres of potatoes. These district totals and averages are weighted together by external weights.

District Level

The formula used to calculate the indications at the district and state levels are:

Net Yield = Gross Yield - Harvest Loss

$SE_{Net} = [Var(gross)+Var(loss)-2Cov(gross, loss)]^{½}$

where the variances, Var(gross) and Var(loss), are computed using the formula for simple random sampling and the covariance (Cov) is determined from samples which were used for both the gross yield and harvest loss components.

State Level

State Net Yield = $(W_1Y_1 + W_2Y_2)/(W_1 + W_2)$

$SE_{SNY} = [(W_1^2SE_{Y1}^2 + W_2^2SE_{Y2}^2)/(W_1^2 + W_2^2)]^{½}$

where Y_1 and Y_2 are the net yield indications for Districts 1 and 2 and W_1 and W_2 are the current indications of harvested acres for the corresponding districts.

Examples

The following example illustrates the computation of gross yield and loss at the sample level. Since the sample was selected with probability proportional to size, by district, the sample is self weighting at the district level. Thus, the indication for each component can be calculated by taking a straight average of the indications for each usable sample.

Suppose the following information was obtained for sample 24:		
	UNIT 1	UNIT 2
Number of hills	18	21
4 - row spaces (ft)	12.5	12.7
Field weight (grams) Hills 8, 9, and 10	2675	2280
Weight of gleaned tubers (grams)	590	350

Computing the values for Unit 1,

Hills per acre=[18*43,560] / [(12.5/4)*20] = 12,545

Weight per hill = 2,675 / (3*453.6) = 1.9658 pounds

Continuing for Unit 2,

Hills per acre = [21*43,560] / [(12.7/4)*20] = 14,406

Weight per hill = 2,280 / (3*453.6) = 1.6755 pounds

Therefore,

Gross Yld=[(12,545*1.9658)+(14,406*1.6755)]/[2*100] = 243.99 cwt. per acre

and,

Loss = [43,560*(590+350)]/[2*3*6*453.6*100] = 25.07 cwt. per acre

For the district level, assume that average gross yield is 310.56 cwt. per acre and harvest loss was 19.07 cwt. per acre. The net yield would be computed as:

Net Yield=310.56-19.07 =291.49 cwt. per acre,

which rounds to 291 hundredweight per acre.

CHAPTER 10 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 State Field Office (FO) does its independent appraisal of the relationships between the survey estimates and the final official statistics and forwards this information to Headquarters.

While each FO 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 affect 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 must be tended or locked up at all times in the FO 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 (8:30 a. 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 Frame Survey and the June Multiple Frame Survey provide the benchmark estimates of acres planted. In some years, weather related problems delay planting activities which means farmers are actually reporting acres they still intend to plant. When this occurs, subsamples of farms included in the June Survey 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 to be harvested and actually 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 three key indications: average field level yield regressed to official estimates, average counts regressed to official estimates, and average yield reported by farmers in the Agricultural Yield Survey regressed to official estimates. Once harvest begins, average farmer reported yield regressed to official estimates is added to the set of indications. In addition to the point estimates, forecast errors of the regression equations are also computed. Adding and subtracting these forecast errors from the forecast value forms a forecast range for each indication. Usually, the ranges for the three indications overlap defining the range that

simultaneously satisfies all forecasts. Remember, the regressed to official estimates have accounted for all critical factors in yield estimating, such as standard units, harvest loss, and bias.

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 Agricultural Yield 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 addresses the row crops and it too is 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 Objective Yield 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 useful 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 nonsampling 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.

References

1. Vogel, Fred and Gerald Bange, “Understanding USDA Crop Statistics, “Unpublished manuscript, National Agricultural Statistics Service, April 1997.

2. Neter, John William Wasserman, and Michael H. Kutner, Applied Linear Regression Models, 1983, Richard D. Irvin, Inc., Homewood Illinois.

3. Warren, Fred, “corn Yield Validation Studies, 1953-1983", Statistical Reporting Services Staff Report, June 1985, SRB-85-07.

4. Belsley, D.A., E. Kuh, and R. E. Welsh, Regression Diagnositics: Identifying Influential Data and Sources of Collinearity, 1980, John Wiley & Sons, New York, New York.