AGRICULTURE DATA SYSTEMS - A U.S./CANADA COMPARISON

Robin O. Roark
USDA/NASS, International Programs Office
So. Agriculture Bldg Room 4132; Washington, D.C. 20250-2000

Key Words: NASS and AgDiv

The Canada/U.S. Free Trade Agreement has opened the border for more agricultural trade between Canada and the United States. It will also increase the need for agricultural data and comparison of statistics from each country. The Agriculture Division (AgDiv) of Statistics Canada (STC) provides a wide array of agriculture statistics for Canada just as the National Agricultural Statistics Service (NASS) provides for the United States. However, procedures for sampling, data collection, analysis, and compiling data can be quite different. Even the structure of the agriculture industry, the structure of the two governments and of the two agencies plays a role in how data are collected, summarized, and published.

NASS, the statistical agency for the U.S. Department of Agriculture (USDA), is responsible for agriculture production and inventory statistics and some of the economic statistics. Economic Research Service, another economic agency within USDA, is responsible for compiling the statistical data into farm income and economic projections. World Agriculture Outlook Board, also one of the USDA economic agencies, uses the U.S. agriculture statistics from NASS, along with data from other countries, to estimate the world agriculture supply and demand.

The Agriculture Division of Statistics Canada is responsible for agriculture production, inventory and economic statistics and also compiles data for farm income and economic projections. Some economic data analysis work for outlook projections is done in conjunction with Agriculture Canada. Agriculture Canada is the agriculture policy ministry (department) of the Canadian Federal Government.

The primary difference in the structure of NASS and AgDiv is that AgDiv is part of a centralized statistical system known as Statistics Canada. Within STC there are several program divisions, including AgDiv, that have the responsibility of preparing statistical data related to their division. Other divisions have specific supporting functions to all STC program divisions, such as research, survey design, computer programming, dissemination, etc. The de-centralized U.S. statistical system has several statistical agencies. Each statistical agency is responsible for a particular area of data but must also support their own research, survey design, programming, dissemination, etc. NASS is the statistical agency for agriculture within USDA.

Both organizations have field offices to facilitate data collection, but the functions of these offices are different. For NASS, State Statistical Offices (SSO) are responsible for list maintenance, data collection, data entry, summarization, analysis and administrative work. The SSO's are a significant part of the NASS structure and have a great deal of input into the analysis and estimation of the data. They receive guidance and support from the main office in Washington, D.C. and concentrate primarily on agriculture related statistics. The State offices also have the freedom to be involved in state funded projects that are not part of the National program.

For AgDiv, data collection is done at the Regional Offices (RO). The ROs are part of STC and are primarily data collection and report dissemination centers. The ROs are not directly tied to AgDiv and they collect all types of statistical data. They generally do not get involved in the analysis, summarization or estimation of the data. However, the AgDiv does have agreements with each of the Provincial Governments. These agreements state that the Provincial Statisticians will review the data and estimates prior to the publication of the data. These statisticians are allowed some input into the level of the published estimates.

Both organizations use sales of agricultural products, without regard for acreage, as the defining factor for establishing an operation as a farm. However, the cut-off for the sales value is different. The definition of a farm for the U.S. is any operation that has $1,000.00 in agriculture sales or expected sales. For Canada the definition of a farm is any operation that produces agricultural product(s) for sale.

Census of Agriculture

The Agriculture Census for Canada is conducted every five years by AgDiv. The enumeration coincides with the Census of Population. Therefore, a question is asked on the population census questionnaire about farming interests. If the response is positive, then an Agriculture Census questionnaire is filled out by the respondent. The questionnaires are delivered by a STC enumerator and are mailed back. The enumerators do follow-up of the non-response. Analysis of data gives an expected under-coverage of about 1.5% to 3%, depending on the estimate.

Agriculture Census data are reviewed by AgDiv analysts at the provincial, and sub-provincial level, concentrating on the top contributors with most of the manual editing done on a macro-level. Only if severe problems are detected, or in the review of extremely large operators, are individual records reviewed by analysts. After Census data have been reviewed and published, AgDiv completes a 5 year historic review of all acreage, production, inventory, and economic estimates. Generally, AgDiv estimates, for the year of the census, are revised to match Ag. Census estimates, with minor adjustments due to differences in reference dates. The impact of the under coverage or duplication is assumed to be negligible.

The U.S. Census of Agriculture is conducted every 5 years by the Agriculture Division of the Census Bureau, part of the Department of Commerce. The U.S. Agriculture Census is a stand alone collection, not tied to the U.S. Population Census. The U.S. Agriculture Census is a mail out survey with telephone follow-up of the non-response. The Agriculture Division of the Census Bureau maintains a list of farms. The list is updated from responses to their surveys and from outside sources, such as NASS, income tax records, etc. Under-coverage from the Agriculture Census is about 13% of the farms. Under coverage from the Census is primarily with the smaller farms. The Agriculture Census uses the NASS area frame to estimate potential under coverage. However, Census published numbers are totals from the survey and are not adjusted for the under coverage. Duplication also causes some problems, especially with producer contract arrangements. Census data are reviewed by NASS statisticians at the county, State, and National level. Like AgDiv, data are reviewed on a macro-level with review of individual records limited to severe problems and extremely large operators. After Census data have been reviewed and published, NASS completes a 5 year historic review of all acreage, production, and inventory estimates. However, estimates from the Agriculture Census are not used as official NASS estimates. NASS makes adjustments to Census data to account for duplication, under coverage and differences in reference dates.

The linkage between the Canadian Census of Agriculture and the Population Census, allows for more Census estimates of the social characteristics of agriculture. However, the Agriculture Division of the U.S. Bureau of Census conducts follow-on surveys to establish estimates for most of the same statistics.

Sampling Frames

List - The farm register (list frame) for AgDiv is based primarily on names received during the Ag. Census. The Ag. Census is used as the basis for the list frame for a 5-year period. The majority of updates are based on changes found during surveys conducted during the 5 year period. The list frame, for most probability surveys, is frozen between Census years. The samples for these surveys are selected shortly after the current Census. New names are not generally added to the frame, except names of operators that are new to agriculture and take over an existing operation are allowed to replace an existing name. Some surveys, such as the fruit and vegetable survey, do use producer organization lists to update new names in between Census occasions. These samples are re-drawn every year. Coverage at the time of the Census is estimated to be about 97% for most samples, but this percentage drops at the rate of about 1-2% per year after the Census, depending on the commodity being measured.

The list frame for NASS is continually updated and samples are redrawn every year. Since updates to control data are based on information received during surveys and on information from producer organizations and government program participation lists, etc., not all names and control data are updated each year. NASS does not receive names or control data from the U.S. Agriculture Census. The NASS list frame was built many years ago from outside organization lists. Coverage runs at about 55% for the number of all farms, but varies significantly by State. Coverage is concentrated on larger farms with coverage of farm land at about 80%.

Area - The AgDiv area frame is designed to produce a weighted segment indicator to be used in conjunction with list frame surveys. When screening is done, the

primary data collected are name and address information, total acres, acres in the segment (a piece of land with identifiable boundaries used as a sampling unit in area-frame sampling), and some general information about the type of farm. Agricultural operations are then determined to be either overlap (included on the list frame) or non-overlap (not included on the list frame). The non-overlap operations are then included in subsequent multi-frame surveys. Virtually no data indicators are produced from the area frame alone.

The NASS area frame survey is designed to produce closed segment indicators, open segment indicators and weighted segment indicators. Indicators from the June Area Frame Survey are used both as independent indications and also used in conjunction with list surveys to produce multi-frame indications. Area screening includes collection of tract (the area of land located within a segment that is under a single operating arrangement) data for all agriculture operations found in the area frame sampled segments and entire farm data for all operations that are not known to be overlap with the list frame. Non-overlap operations are also included in Agriculture Survey program for the rest of the survey year. The NASS area frame plays a much more significant role in the estimation program for NASS since the list coverage is lower.

AgDiv has begun using telephone enumeration to identify area frame operators in some areas of the Prairie Provinces (Alberta, Saskatchewan, and Manitoba). Segments in this region are drawn to follow the range and township boundaries. Since only limited data are collected at the time of the screening, preliminary results have been very favorable. Due to the complexity of segment boundaries in the other regions, the segments are personally enumerated. All NASS segments are personally enumerated due to both the precise segment boundaries and the amount and type of data that must be collected. The design of the AgDiv area frame specifically excludes areas not considered to be involved in agriculture, such as rangeland and urban areas. The NASS design includes these areas but samples them at a proportionally lower rate.

Sample Design

Both organizations use probability sample designs on all major surveys. Crop estimates for AgDiv are based on a series of surveys called the Crop Panel surveys. The Crop Panel surveys begin with a Seeding Intentions and Grain Stocks Survey in early April. The panel surveys continue with the June Acreage Survey, July 31 Yield and Grain Stocks Survey, September 15 Yield Survey, November Acreage & Production Survey, and December 31 Production & Stocks Survey.

Crop estimates for NASS are based on a series of integrated surveys called the Agriculture Survey program and the monthly Agriculture Yield Surveys. The March 1 Agriculture Survey is used to estimate seeding intentions. The June 1 Agriculture Survey is used to establish the planted acres and preliminary harvested acres. The September 1 Agriculture Survey is the end-of-season indicator for production of small grains and the December 1 Agriculture Survey is the end-of-season indicator for production of other field crops and hay. The Agriculture Surveys are also used to collect quarterly on-farm grain storage data. The monthly Agriculture Yield surveys collect yield and production data on crops during the growing season. The crops included in the survey will vary from month to month depending on the growing season of each crop and the program for that crop. The first small grain yield survey is conducted in May and the first row crop/hay yield survey is in August.

The actual sample design is fairly similar for the two organizations with both designs stratified by acreage of cropland. The NASS sample is stratified by State on cropland and grain storage, with some strata for specialty crops such as tobacco. The AgDiv Crop Panel is stratified on cropland by sub-provincial regions.

The Agriculture Survey sample for NASS is also stratified to collect quarterly hog & pig inventory and farrowing data. Reference dates for hog inventory estimates are the same as the dates for the four Agriculture Surveys. The cattle & sheep estimates reference dates are January 1 for both cattle and sheep and July 1 for cattle only. Therefore, cattle and sheep data are collected via a separate survey with separate stratification and sampling. The AgDiv livestock surveys are done twice each year and include cattle, hogs, and sheep. The reference dates for the livestock surveys are January 1 and July 1.

There are also numerous probability and non-probability surveys conducted by both organizations to obtain statistics on commodities such as fruit, vegetables, poultry, prices paid and received by farmers, farm income and expenses, etc. Due to the

structure of the farm programs and marketing boards, there are more administrative data available to the AgDiv than are available to NASS. The supply managed commodities, which are milk, eggs and poultry meat, have extensive administrative data available that are used by the AgDiv in lieu of survey data. Farm expense data for AgDiv are obtained through a sample of income tax records rather than an additional survey of farmers. Both organizations use administrative data whenever available to help relieve respondent burden and lower data collection costs.

Data Analysis

NASS questionnaires that are not collected using Computer Assisted Telephone Interview (CATI) procedures are put through a complete manual review prior to data entry. Discrepancies are reviewed and corrected. Within AgDiv, the manual editing of data prior to data entry is virtually non-existent. Data that are not collected with CATI are briefly reviewed by the data entry division for clarity. However, data are not edited as being "correct" or "incorrect".

Computer editing of data, after collection, is used in both organizations. NASS's computer editing is designed to review data and flag errors with only a small number of the errors corrected by the editing program. The remaining errors are then reviewed and corrected by statisticians. The computer editing program for AgDiv is designed to make corrections or perform imputations for most of the errors. Statistician review and correction of the remaining micro-level errors is not as prevalent.

Expansion (or raising) factor adjustment is used by both organizations in most probability surveys to account for missing data and refusals. Some surveys in both organizations use automated imputation procedures. Response rates for the production surveys are very similar between the two agencies.

Estimation

The estimation program in NASS requires that most data and estimates be reviewed, analyzed, and approved by the Agriculture Statistics Board (ASB). The ASB is made up of the ASB Chairperson, the Estimates Division Director, the respective commodity statistician(s) and their Branch Chief, and 1 or more statisticians from 1 or more SSOs. For selected estimates, the Secretary of Agriculture, or a representative from the Secretaries office is briefed about the estimates prior to the release of the data. Within AgDiv, generally only the statisticians (Federal and Provincial statisticians) and their immediate supervisor review the data and estimates prior to the release.

The concept of livestock inventory estimates are nearly identical. The weight groups, age groups, and livestock definitions are virtually the same. However, the reference dates and estimation procedures are different. AgDiv produces sheep inventory estimates twice a year while NASS produces these estimates only for January 1 each year. Both organizations produce January 1 and July 1 cattle inventory estimates. For hog estimates, the reference dates are off by one month. NASS's reference dates are December 1, March 1, June 1, and September 1 with AgDiv dates being January 1, April 1, July 1, and October 1. NASS conducts a survey for each of the 4 quarterly estimates while AgDiv makes the estimates for April 1 and October 1 without the use of a survey. The inventory estimates are based on administrative data and previous survey data.

Both organizations, in spite of the differences in structure, make use of market trends and information provided by field experts. The primary estimation tool for both organizations is the balance sheet. AgDiv estimates are made so the balance sheet residual is zero where NASS will generally allow small residuals to remain.

Crop estimates for seeding intentions and for preliminary yield surveys have a different base concept between NASS and AgDiv. The seeding intentions estimates from NASS are designed to forecast what the actual planted acres will be, thus requiring the statistician to make a forecast of the seeded acres. For AgDiv, seeding intentions are designed to be a point estimate showing current seeding plans of farmers, without making a projection of what planted acres will actually be.

The same concept holds true for yield surveys. With NASS, data analysis done with the monthly yield surveys is designed to forecast the final yield and production of the each commodity. NASS uses objective yield surveys for the wheat (winter, spring & durum), corn, soybeans, cotton and potatoes. The objective yield surveys are conducted in the major producing states and usually account for over 80% of the production. The objective yield models are designed to compare current conditions with historic

conditions and compare to final yields. For AgDiv, data analysis is designed to estimate current yields, with no adjustments made for historic trends or comparisons of survey indications to final yields. The only objective yield survey for AgDiv is for potato production.

Conclusion

Most of the differences mentioned above have advantages and disadvantages when compared together. Despite the differences in structures and external controls, both organizations produce a wide array of agriculture statistics that are used to establish agriculture policy for the respective countries. However, the background and even the definition of what the data represent are often quite different. Therefore, when comparing the data from each organization, the data concepts are equally as important as the numbers themselves.

History and Procedures of Objective Yield Surveys in the United States

Eric Waldhaus, Eddie Oaks, Mike Steiner, National Agricultural Statistics Service
Eric Waldhaus, NASS, USDA, Room 4162, South Building, Washington, D.C. 20250

KEY WORDS: Crop cutting, Enumerators, Objective Yield, Yield Surveys

This paper reviews the efforts made by the National Agricultural Statistics Service (NASS) in the United States Department of Agricultural (USDA) to measure crop production by direct measurement of plant characteristics. NASS implies the current agency and all of its predecessors. These efforts are collectively known as Objective Yield Surveys (OYS). This program is similar to Crop Cutting surveys done in many parts of the world. The major difference is that OYS includes non-destructive field counts prior to harvest to facilitate yield forecasts during the growing season. Yield is defined as the weight of targeted crop, at standard moisture, produced per unit of harvested area. Sampling for NASS Objective Yield Surveys have been on a probability basis from the inception. This is, however, not the implication of the word 'objective' in the title. Objective refers to the direct collection of plant characteristic measurements, instead of subjective estimates of yield reported by an observer.

NASS is a recognized world leader in the use of objective yield technology. Objective yield surveys produce the primary indications for yield forecasts and estimates for the major feed and food grains in the United States. Additionally, NASS has made long term commitments to make this technology available internationally. Through cooperative arrangements NASS has demonstrated or helped implement objective yield programs in many countries of Asia, Africa, and Central and South America.

Three aspects of NASS' objective yield program for major field corps are considered: The history and evolution of the program, current sampling procedures, and general concepts of objective yield survey field procedures. Specifically not included is a discussion of the use of survey data in preparing yield forecasts. This major topic is covered in another paper presented at this conference.

HISTORY

Yield and production of major field crops in the United States have been forecasted and estimated by USDA since President Abraham Lincoln's administration in the 1860's. Crop condition surveys were prepared monthly by the Statistics Division, USDA as early as 1863, the year following the creation of the Department. Until 1884 pre-harvest reports were in terms of condition as compared to an 'average' crop. In 1884 the reporting concept changed. Condition began being asked as a percent of a 'normal' crop, given no adverse effects of weather, disease or pests.

Although crop area changes from year to year, some of the largest variations in crop production are caused by fluctuations in production per unit area or yield. For more than a century, yield forecasts were based solely on voluntary producer appraisals of expected yield. It was recognized early that actual changes in yield were not fully reflected in subjective grower appraisals. By 1898 traveling agents supplemented farmer-crop reporters' information with on site observations of crop conditions. By 1903 more than 100,000 agriculture related business operators, including cotton ginners, millers, elevator operators, and transportation agents were paneled to gain insight into the agricultural situation.

In 1910 a shift began in the practice of reporting crop condition to forecasting actual production during the growing season. By 1915 cotton production forecasts became available during the growing season. The transition from condition to yield forecasts required regression modeling. This was almost entirely done by visual interpretation of charts prior to the use of computers in the late 1960's.

Objective measurements for forecasting yield started with cotton in 1928. These early efforts involved statisticians driving along the perimeter of cotton fields, making boll counts at predetermined locations in fields. There appears to have been no effort made to relate the field counts to yield. Thus, it may be more appropriate to think of this early effort as 'Objective Condition' surveys. Later corn and wheat were added to this program, but this early effort in objective methods was discontinued at the start of the World War II. Research into objective measurements of wheat, corn, and cotton resumed in 1954.

The 'birth' of probability sampling for agricultural statistics and objective yield methods came in 1957 when the United States Congress funded an initiative titled "A Program for the Development of the

Agricultural Estimating Service". The project provided for an annual enumeration of a large area frame probability sample for crop area estimates. This area frame survey evolved into the current June Agricultural Survey (JAS). Target crop fields identified during the JAS provide the sampling universe for the OYS (except winter wheat).

Cotton and corn objective yield programs became operational in 1961. Wheat came on line a year later. Soybeans joined the national program in 1967 and potatoes in the early 1970's. Grain sorghum, sunflowers and rice were added in the 1980's, but due to budget constraints grain sorghum and sunflowers were dropped in 1988. The rice program was reduced then and finally discontinued in 1993.

OBJECTIVE YIELD SURVEYS OVERVIEW

NASS is organized into 45 State Statistical Offices (SSO). There is one in each state except in New England where six states are combined. There is a centralized Headquarters in Washington, D.C. Sample design and selection, planning and coordination between states, centralized data processing, and quality assurance are the major roles of Headquarters in the OYS. Headquarters prepares and distributes two major OYS manuals. The Supervising and Editing manual is focused on the tasks completed in the SSO. The Interviewers' Manual is a training and reference manual for enumerators in the field. Each SSO coordinates field work and other data collection activities independently within established guidelines.

Qualified, adequately trained field personnel, including SSO staff and field enumerators, are essential for a quality job. States send a survey statistician, designated the State Survey Statistician, to a National training workshop to learn and reinforce correct procedures. State Survey Statisticians return to their states to train field supervisors and enumerators.

Objective Yield Surveys begins with intensive training for field enumerators. Training is more intensive for OYS than many other NASS data collection operations. The need for rigorous training stems from the fact that data collection usually is accomplished in remote locations in the field where supervision is minimal and there is not the opportunity to clarify procedures. It is also recognized that data collection is often a very time sensitive process so it may be impossible to reconstruct an 'interview' when errors are discovered after the field work is complete. The cost of training is very high, but the need is critical. NASS has consistently recognized this need and continues to make a substantial resource commitment to training.

NASS field enumerators are part time employees of the National Association of State Departments of Agriculture (NASDA). NASDA contracts their services to NASS. There are approximately 600 NASDA enumerators who work on OYS. OYS enumerators are almost exclusively rural people, and most are from farm families, typically retired or part-time farmers or farm spouses. Understanding agricultural practices is a prerequisite for a successful OYS enumerator. Enumerators also have to demonstrate literacy and computational skills about equivalent to a high school graduate.

In addition to training, the State Survey Statistician is charged with monitoring survey progress, and is the resource person for enumerators. Responsibly also extends to oversight of all SSO processing of survey data, and supervising laboratory processes. The State Survey Statistician and assistants review all edit and summary output. In most states the final yield recommendations (proposed estimates) submitted to Headquarters are not prepared by the Survey Statistician, but a Commodity Specialist.

Field Quality Control is conducted by supervisory enumerators and statisticians from the State office. A random sample of each enumerator's field work is selected for personal inspection. The sample selected for quality control is unknown to the enumerator and the supervisor in advance to insure an accurate assessment of quality. The sample pattern is such that at least one quality check for each enumerator is insured, and multiple checks throughout the survey cycle are possible. Supervisors may inspect additional work of the enumerators in their charge on an 'as needed' basis.

Occasionally, deficiencies in field procedures are discovered by the quality control process. When this occurs remedial action is taken, both to correct errors in a particular sample and to re-train the errant personnel. Discovery of deliberately falsified survey results is another potential benefit of the Quality program. The authors, with about twenty years of objective yield experience each, have no personal knowledge of this ever occurring.

Objective yield surveys are timed for making crop production estimates which are released to the public in the monthly *Crop Production* report. *Crop Production*

is published during the second week of the month, between the 8th and the 12th. To complete field work, process all data, and remain timely, the OYS adheres to a very rigid schedule. Data collection starts on the 22nd of the month prior to the survey reference date, and must be completed by the first of the reference month. Laboratory work, data processing, and summary review are completed, and recommendation submitted to the NASS Agricultural Statistics Board in Headquarters by the second day before the *Crop Production* release.

Concepts and methodology used in the OYS for forecasting and estimating yields are similar for all field crops. Two components of yield -- weight of the fruit and number of fruit -- are used to forecast a yield. Various plant characteristics are used to predict these components during the growing season. Harvest losses, estimated by gleaning small plots in the sample fields after harvest, are deducted to obtain a net yield.

During the early growing season, crop maturity varies considerably by region. As the season and plant maturity progresses the plant characteristics and measurements made to forecast yield change . The enumerator determines the maturity stage of the crop in the sample field during each visit and makes the appropriate counts and measurements for the growth stage.

Observations for each sample are made on two randomly selected plots (units) in each of the selected fields. Each plot consists of a specified number of parallel rows of predetermined length, or a rectangular unit drawn to specification if crop rows are indistinguishable.

SAMPLING

OYS samples are selected from acreage reported of the target crop in the March Agricultural Survey (MAS) or the June Agricultural Surveys (JAS). Spring and durum wheat, corn, cotton, potatoes, and soybean samples are selected from the JAS. The winter wheat sample comes from the MAS.

Winter wheat samples are unique as they are selected from the March Agricultural Survey using a multiple frame (combined list and area survey) design. Also, winter wheat varies in that samples are drawn from 'fields to be harvested for grain', while other crops are sampled from fields 'planted and to be planted' on the parent survey.

The objective yield sample for each crop is allocated to the most important production states such that 80 percent or more of the nations crop is included. Allocations are made to minimize production estimate coefficient of variation (CV). Until about 1990 allocations were made to maintain minimum harvest level CV's. As estimation models have improved, an effort has been made to allocate samples to maintain a minimum CV across the growing season.

The JAS, which is the parent survey for OYS, is the major, once a year, multiple frame survey conducted by NASS. Nationally, the area frame component includes approximately 15,500 segments, each about 1 mile square, representing about 52,500 farms which are enumerated in early June to identify land use. The area of target crop planted is expanded by the associated expansion factor for the area frame sample. OYS samples are then selected proportional to the expanded acreage. Proportional sampling insures that the distribution of the OYS sample will approximate the distribution of the crop as discovered in the JAS. Sampling procedures are similar for winter wheat except MAS is the base survey.

Survey States, sample size, and sample distribution are reviewed annual, but NASS has attempted to maintain consistent State involvement and sample sizes to maintain year to year comparability. In 1993 1,670 winter wheat samples were selected in 13 States. Spring wheat samples totaled 380 in four States, and 150 durum samples in one State were selected. Corn samples equaled 2,010 spread over 10 States, while 1,360 Cotton samples were drawn in six States. Soybeans samples totaled 1,330 in eight States, and 2,080 Potatoes samples were distributed over 11 States.

FIELD PROCEDURES

Enumerators are provided aerial photograph with the area frame segment containing the selected sample field outlined in red. Operators of land in these segments were interviewed during the JAS. Within the segment there may be more than one tract (farm). The enumerator locates and interviews the operator of the tract which contains the selected target crop field for OYS.

Six reporting forms are used through the growing season to collect information from the farm operator or to record counts and measurements. The reporting forms are identified by letter initials, which reflect the chronological order of use of the forms during the growing season. The data collected on each form are

similar for all crops in the OYS program.

A convenient way to describe the field procedure for implementing the OYS is to describe each reporting form, and explain its use.

Form A - is an interview form, used to update the crop acreage intended for harvest and to identify the sample field. It shows which field (area frame) or how to select a field (list frame) that will be used for making actual field counts and measurements. The Form A is completed on the first visit to the selected farm. It is also used to gain permission from the farmer to enter the field to set out OYS sample units, and to query the farmer about pesticide usage so the enumerator can take appropriate personal safety precautions.

Pesticide usage has expanded over the years both in the crops treated and the variety of chemicals available. Consequently pesticide safety training and enumerator exposure monitoring has become an integral part of the OYS program. This is especially true for Cotton OY, where the use of organophosphorus pesticides is nearly universal.

Form H - also an interview form, is used to collect data on seed, fertilizer, and pesticide application rates and tillage practices. These data are used for further economic analysis, and are not part of the yield estimation program directly. It is completed at the same time as the Form A.

Form B - is a field observation recording form. It is used to record counts and measurements of the plants and fruits. This form also reiterates instructions for locating, constructing, and processing the sample units.

The following two sections: Locating the Sample, and Counts and Measurements are presented here because these activities are associated with completion of Form B. A separate Form B is completed each survey month until harvest time, when a final Form B is completed.

LOCATING THE UNIT:

After completing Forms A and H, the units are constructed in the sample field by the enumerator. Two units are laid out for each sample. Unit 1 and Unit 2 are located independently of each other (except in wheat where unit locations are dependent). The random number of rows and paces for locating Units 1 and 2 are computer generated and preprinted on a label on the Form B.

The point of entry into the field, or starting corner, is the first corner reached when approaching the field that allows the units to have a chance of falling anywhere within the field boundaries. The shape of the field must be considered to insure that the entire field has a chance of selection. Research has indicated that there is no statistical differences related to starting corners. Therefore, any field corner which does not exclude some part of the field is acceptable.

The following steps are followed when locating and laying out units:

Step 1: The enumerator marks the staring corner with a piece of plastic flagging ribbon so it will be clearly visible on later visits.

Step 2: The enumerator then walks along the end of the crop rows the number of rows (or paces for wheat and broadcast seeded fields) indicated for Unit 1. A piece of flagging ribbon is tied onto the first plant in Row 1. This helps locate the same row on later visits. The next row in the direction of travel will be Row 2 of Unit 1.

NOTE: The enumerator walks his or her normal paces when locating the units within the field. It is not necessary to measure the distance traveled as it is not necessary to locate a precise point in the field, only one determined by a random process.

Step 3: The enumerator then walks the required number of paces into the field between Row 1 and Row 2, starting the first pace 1.5 feet outside the plowed end of Row 1. This makes it possible for a unit to fall anywhere in the field including the very edge.

Step 4: After the last of the required paces is taken, a dowel stick is laid down so that it touches the end of the enumerator's shoe. The dowel is placed across Row 1 and Row 2, at a right angle to the rows. The unit is laid out in the direction of travel of the last pace.

Step 5: The zero end of a 50 ft. tape is anchored at the dowel stick directly beside the plants in Row 1. The sample number is written on a florist stake and inserted at the anchor point.

Florist stakes are colored lath about 6 to 8 inches long. They are highly visible markers commonly used in nursery and greenhouse operations to mark seed beds. Florist stakes deteriorate quickly so no hazard will be created if lost or abandoned in the field after the

survey.

Step 6: In row 1 a starting florist stake is placed exactly 5 feet from the anchor point. It is marked "U1-R1". This measured 'buffer zone', helps insure that the unit location is not subjectively biased in its location by the enumerator. The florist stake should be placed beside the row about 2 inches from the base of the plants. The marker is placed outside the plant row to avoid any damage to the developing crop.

Step 7: Working outside the unit, the enumerator carefully measures the unit length and places a florist stake at the designated point. Corn, cotton and potatoes have larger unit lengths which are measured with a tape. For example, the corn count area is 15 feet long. A rigid metal frame is used for marking wheat and soybeans where the unit size is smaller. The wheat unit is 21.6 inches.

Not all fields are square or rectangle and other special situations may arise when locating and laying out a unit. The Interviewers' Manual gives details on how to handle most of these situations. Some of the problems that more commonly occur include: blank areas in the field that were known or unknown during the mid-year survey; the field is not large enough to accommodate the number of rows or paces specified; row direction changes; odd shaped fields are encountered as circular fields under pivot irrigation; fields planted on contours; or crop rows that are not distinguishable due to sowing practices. These situations are covered with precise instructions.

The Form B is the recording form for counts and measurements that are made at the units. Visits to these sample units will take place monthly during the growing season except for potatoes, when only one visit is made within 3 days of harvest or when vines are dead.

Because the same sample unit must be revisited monthly it is important the enumerator precisely mark the location of the unit. Plastic flagging ribbon is used. This is highly visible, but like the florist stakes, quickly disintegrates so it may be abandoned after the survey.

COUNTS AND MEASUREMENTS

Step 1: Measure 1-row space and then 4-row spaces. Measurements are made from the plants in row 1 to row 2 and then from row 1 to row 5. These measurements are used to calculate area of the unit.

Step 2: Count the number of plants in each row in the designated unit.

Step 3: Classify the unit by maturity category. Descriptive four page handouts with color picture examples are helpful in determining maturity.

Step 4: Make the specific counts and measurements of plant characteristics required. Different counts are made depending on the maturity level category. The crop and type of counts are as follows:

Soybeans: 1) plants; 2) nodes; 3) lateral branches with blooms, dried flowers, or pods; 4) blooms, dried flowers and pods; and 5) pods with beans.

Corn: 1) plants; 2) average length of kernel rows; 3) diameter of ear; 4) stalks with ears or silked ear shoots; 5) number of ears; 6) ears with kernel formation; and 7) cob length; and 8) field weight of corn.

Cotton: 1) plants; 2) burrs, open and partially opened bolls; 3) large unopened bolls; 4) small bolls and blooms; and 5) squares.

Wheat: 1) stalks; 2) heads in late boot; 3) emerged heads on all stalks; and 4) detached heads.

Potatoes: 1) hills; 2) tubers; and 3) field weight of tubers in the unit.

After completing Unit 1 counts and measurements go back to the beginning of the Row 1, and walk to the designated row, or number of paces, for Unit 2. Continue in the original direction of travel as when locating Unit 1 if Unit 2 count exceeds the Unit 1 count. After locating the Row 1 of Unit 2, walk the required paces into the field to set up Unit 2, and make the counts and measurements required.

A Form B is done for each month until very near harvest. Close contact is made with the operator so a sample field will not be harvested before a final Form B (just before harvest) can be completed. During this last visit before the farmer harvests, a sample of mature crop is sent to the laboratory. This sample is the basis for at harvest yield estimates.

FORM C-1 and C-2 - These forms record laboratory observations, and are not seen by the field enumerator. Form C-1 records data from pre-harvest field visits, while the C-2 is generated from the last field visit made at, or just before the farmer harvest.

FORM D - is used to record the actual number of acres harvested at the end of the year and the operator estimated yield of the field.

FORM E - is a field observation form used to collect data for determining field harvest loss so a net yield estimate can be made. The field visit to collect data must be within 3 days after harvest to determine harvest loss accurately as loose grain deteriorates quickly or is lost when left in the open. Harvest losses are subtracted from gross yield to arrive at a net yield. Finding the location of this post-harvest unit is similar to the original unit location. A measured rectangle is staked out and fruit from the crop is collected, and sent to the lab. There it is counted, weighed, and moisture tested to determine the field loss.

NON SAMPLING ERROR

Controlling non-sampling error is a major concern of the OYS program as in any large scale sampling survey project. Cause for OYS non-sampling error can be divided into two major categories: faulty procedures, and faulty procedure implementation. Additionally, as OYS use sub-samples from other surveys, non-sampling error present in the parent survey is passed on or magnified. This is out of the control of the OYS personnel except to monitor the larger survey for consistency. This source of error will not be considered further herein.

Non-sampling error which are the result of faulty procedures can be dealt with in a straight forward manner. The NASS research unit continuously reviews various aspects of the OYS program to insure survey validity. Validation surveys are conducted for each crop on a rotational basis. These surveys explore many aspects of OYS, such as the independence of the starting corner as noted earlier.

The survey quality program is also useful in discovering faulty procedures. Most often procedural difficulties that are discovered in the quality control program relate to some 'special case' which was not adequately considered when preparing manuals. Instruction changes that clarified selecting starting corners that do not exclude some part of the field developed largely through this route.

Insuring that procedures are consistently and accurately followed across the country is the greatest challenge in controlling non-sampling error. The most important control for non-sampling error is training. OYS training is continuous. The training cycle starts with training for State Survey Statisticians at National workshops. Usually there are three held in a year, one for Wheat, another for Corn, Cotton and Soybeans, and the third for Potatoes. Corn, Cotton, and Soybeans are combined for training because procedures, growing seasons, and States involved largely overlap.

Training continues with workshops for field enumerators, conducted by the State Survey Statistician. Assistance from the Headquarters OYS unit is available to the SSO's in conducting local training. This can be an important resource for a new State Survey Statistician, and gives Headquarter personnel the opportunity to observe local operations.

The formal quality control program in which individual enumerators have work inspected at random is an important part of the NASS non sampling error control program. While the potential is in place to discover an enumerator who is intentionally falsifying reports or 'table topping', this is not a major concern. The real value of the quality control program is to assess the level and effectiveness of training. Another important benefit of the program is its moral boosting effect on enumerators. The normal out come of quality control is that the enumerator is 'caught doing it right'. When fed back to the enumerator in a positive way this can be excellent reinforcement for continued quality field work.

REDESIGN OF THE CANADIAN AGRICULTURE COMMODITY SURVEYS

J. Trépanier and A. Théberge, Statistics Canada
J. Trépanier, Statistics Canada, Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6

1. INTRODUCTION

From 1983 to 1992, the Canadian National Farm Survey (NFS) has produced estimates on cropland areas, livestock and farmers' expenses and income. The NFS was the second most important agricultural survey after the quinquennial Census of Agriculture. It was a multiframe survey and was conducted in all provinces of Canada except Newfoundland. The major agricultural surveys have traditionally been redesigned following each Census of Agriculture which provides information on a multitude of farms' agricultural activities. Consequently, the last Census of Agriculture that took place in 1991 gave rise to a complete redesign of the NFS based on the 1991 Census data. An evaluation of the Agriculture Statistics Program and an analysis of its probable evolution over the next few years suggested that the NFS be replaced by three separate surveys. Starting in 1993, surveys covering respectively crops, livestock and farm financial data will be conducted and supplemented by a common area sample collected by the Area Farm Survey (AFS).

Since 1988, the NFS design relied on one list frame and list sample for the Maritime provinces (Prince Edward Island, Nova Scotia and New Brunswick), Québec, Ontario and British Columbia (except the Peace River district). Two list frames, each with their own sample, were used for the Canadian Wheat Board (CWB) area, that is the Prairie provinces (Manitoba, Saskatchewan and Alberta) and the Peace River district in British Columbia (Julien and Maranda, 1990). Note that the Peace River district is agriculturally similar to the Prairie provinces. Both NFS list frames included 1986 Census farms except those on Indian reserves and institutional farms. Small farms in terms of cropland areas were also excluded in the CWB area. The list frames were complemented by an area frame in the CWB area, Ontario and Québec. In the CWB area, income and expense estimates made use of only one of the two list frames, the one with the largest farms, and the area frame.

Under the new design, the crop and livestock surveys are both multiple frame surveys with a list and area frames. They share the same frames. The new list frame includes 1991 Census farms of all provinces except Newfoundland. Only farms on Indian reserves and institutional farms are excluded. These exclusions represent only 0.3% of the total number of farms in the target population. The crop component of the NFS which dealt with cultivated areas is now integrated into an already existing group of crop surveys currently dealing with seeding intentions, yields and stocks of grain, giving a new series of six crop surveys conducted throughout the year. The crop samples are based on univariate stratification method. A master sample is selected and a predetermined subsample of it is taken for each crop survey. On the other hand, the livestock surveys will be conducted in July and January. A sample is first selected for the July survey and then subsampled for the January survey. Multivariate stratification methods are used. Even though the strata are the same, sample allocation is performed separately for the July and January surveys. Methods have been implemented to reduce the overlap between the crop and livestock survey samples in order to control respondent burden.

As in previous designs, an area frame survey is used to account for new operations and those missed in the Census. The new AFS uses a one stage design with stratified random sampling. This new method requires much of Canada's land to be segmented using geographical information systems.

2. CROP SURVEYS

Under the previous design, estimates of cropland areas were produced from NFS data. However, estimates of yields, stocks of grain and seeding intentions were provided by a series of seven crop surveys conducted over the course of the year. For more information on these surveys see Bélanger (1990). It was decided to combine the crop part of NFS and this series of seven crop surveys. The new series of crop surveys consists of six surveys: December (stocks of grain), March (seeding intentions and stocks of grain), June (crop areas), July (stocks of grain), September (yields) and November (yields).

The design of the previous series inspired the sampling design of the new crop surveys. Even if six surveys are conducted, the list frame described above was stratified only once. A sample, called the master

sample, was selected and randomly partitioned into a predetermined number of subsamples. Each crop survey's sample is a union of one or more of these subsamples. This method makes it possible to control the overlap between each survey as we will see in more detail in what follows.

2.1. Stratification

For the crop surveys, the first level of stratification is the province. To facilitate the work, the Peace River district and the rest of British Columbia are treated as two distinct provinces. Within each province, several farms with specific characteristics are selected with probability one. All other farms within the province are stratified and samples are selected within strata.

Because of their large size, some farms must be in the sample of more than one of the six individual crop surveys. These farms are grouped in non partitioned take-all strata which permit control over the inclusion of them in any of the subsamples. There are three such strata for the crop surveys. Some Census farms on the frame are part of farming enterprises with complex operating arrangements. These are called multiholding farms and special data collection procedures have been put in place for them. For this reason, and because they are farming operations with frequently changing structures, all multiholding farms are grouped in a take-all stratum. Since community pastures are also treated separately at the time of data collection, they are also grouped in a take-all stratum.

There are other farms on the frame that can be very large. These farms are often very different from the majority of the farms. In order to avoid the undesirable situation of stratum jumping due to large farms changing hands, which can lead to very high weights, these farms are also placed in a take-all stratum within each province. Special procedures ensure that they remain in their stratum. These largest crop farms are identified by what is called the Sigma-Gap Rule, and are called the Sigma-Gap farms. Total cropland area, as defined below, is the Sigma-Gap variable. The Sigma-Gap Rule can be summarized as follows. Within a given province, let X_i be the value of the Sigma-Gap variable X for farm i. Let P be the set $[X_i : X_i > 0]$, M be the median of X on P, σ be the standard error of X on P, and $X_{(1)}$, $X_{(2)},\ldots,X_{(N)}$ be the ordered values of P. Now let k be the smallest number, if it exists, where $X_{(k)} > M$ and $X_{(k)} - X_{(k-1)} > \sigma$. All farms with $X_i \geq X_{(k)}$ are identified as the Sigma-Gap farms. For the crop surveys, Sigma-Gap farms represent 0.02% of the number of farms on the list frame. All the remaining farms were stratified using the method described in the next two paragraphs.

In Canada, estimates of crops are calculated at both the provincial and subprovincial levels but published at the provincial level only. These subprovincial levels correspond to agricultural regions within provinces. The number of agricultural regions in a province varies from one to twenty. Provincial authorities are increasingly more interested in subprovincial estimates, but budget, and consequently the sample size, are often a constraint. Some of the crop surveys have a sample size of approximately 11,000 units for a population of 280,000 units. Agricultural regions could form the second level of stratification within the province for these surveys and thus provide improved subprovincial estimates. A study was conducted before the redesign to analyze this possibility. Stratification where strata are created within the agricultural regions was compared to stratification where strata are created within the province. If the total number of strata to be formed is fixed within a province, stratification within the agricultural regions showed that improvements to the estimates (reduction in coefficients of variation) at the agricultural region levels could not justify the use of the method because of the deterioration of the provincial estimates (increase in coefficients of variation). A compromise was investigated and made. Similar agricultural regions were grouped. The number of subprovincial regions was thus approximately cut in half. These new subprovincial regions became the second level of stratification, even though estimates are still to be computed at the finer regional level. Strata were then formed within each province and subprovincial region. The number of strata created in each subprovincial region varies from 4 to 10 depending on the size of the region.

The new series of crop surveys covers cropland area, seeding intentions, yields and stocks of grain. Census information does not include variables on intentions, yields and stocks. Thus the 1991 Census total cropland area was the only variable used for stratification. Despite the fact that hay is not a variable of major interest by itself, the Census total cropland area includes hay area. Canadian farmers (except in the CWB area) use hay as part of their crop rotation; what is hay acreage one year can easily become grain the year after. Thus hay acreage was kept in total cropland area except in the CWB area where it was excluded. Total cropland, as modified above, is the stratification variable. The boundaries of

the strata were calculated with Sethi's algorithm when Neyman's optimal sample allocation is used (Sethi, 1963).

2.2. Master Sample and its Subsamples

Canada, for the purpose of the crop surveys, is divided into three major areas: Maritime provinces and British Columbia (excluding Peace River district), Québec and Ontario, and CWB area. There are six crop surveys involved in this redesign. Each survey's sample size was fixed for each of the areas defined above as well as the size of the overlap, if any, between the survey samples. For instance, the March survey sample and the July survey sample have no common element, but together they also form the sample for the December survey. Thus, the size of the master sample in each major area was determined. Simultaneously, the number of subsamples needed and their sizes were fixed. Remember that the master sample is partitioned into a number of subsamples. The union of some of these subsamples forms the sample of each survey. The final size of the sample must be close to the target size. The size of each subsample was then allocated to the provinces within the major area. Allocation to the strata that are not take-all was performed using Neyman's optimal sample allocation. Total cropland area (excluding hay acreage in the CWB area) was the allocation variable. The size of the master sample is approximately 98,000 units.

Keeping in mind that good estimates are wanted at the agricultural region level, proportional representation was ensured by sorting the observations of all strata by agricultural region and selecting the units of the master sample using a circular systematic method. Units within an agricultural region were randomly sorted to ensure that samples within regions are the equivalent of simple random samples.

In a given province, the master sample was randomly divided into a certain number of subsamples of equal size. This partition was performed within each stratum. The representativity of the population by each subsample was verified by computing sample statistics for many of the variables of interest. A number of subsamples is assigned to each of the six surveys. If an overlap is desired between two surveys, the same subsamples are used. Otherwise, different subsamples are assigned to minimize respondent burden. The number of subsamples in a province varies from 5 to 18.

3. LIVESTOCK SURVEYS

The NFS livestock data used to be collected annually in July. In January, a subsample of the July sample was contacted for the January Farm Survey (JFS) and used to provide another series of livestock estimates. This redesign retains the same basic idea. See Julien and Maranda (1990) for more details on the NFS design. Starting in 1993, a July Livestock Survey and a January Livestock Survey are being conducted. Once again, these surveys use the list frame described in Section 1.

3.1. Stratification

For the livestock surveys, the province is the first level of stratification. Unlike the crop surveys, the livestock surveys do not treat the Peace River district separately from the rest of British Columbia. Within a province, the multiholding farms were again grouped into a take-all stratum as were the community pastures. The Sigma-Gap Rule was applied to the remainder of the frame for each of the following variables: beef cows, milk cows, sows, total number of cattle, total number of pigs and total number of sheep. If a farm happened to be a Sigma-Gap farm for any one of these variables, it was included in the take-all stratum created for the Sigma-Gap farms. Sigma-Gap farms represent 0.08% of the total number of farms on the list frame.

A significant number of Census farms on the list frame do not have any livestock (in terms of cattle, pigs and sheep). These can represent up to one half of the total number of farms in a province. Because many farms can have no livestock during the season when Census is done, but yet have some during winter when the January Livestock Survey is conducted, these "zero" farms were kept on the frame. Retaining the "zero" farms also ensures that farms that were misclassified at Census or have since started breeding livestock are also covered. The Census farms with no livestock were grouped in one provincial stratum. All the other farms were stratified using the multivariate procedures Fastclus and Cluster of the SAS software (SAS Institute Inc, 1985). The stratification variables were beef cows, milk cows, sows, total number of cattle, total number of pigs and total number of sheep. These were standardized to have zero mean and unit variance before using the clustering algorithms. Empirical investigation showed that standardized stratification variables achieved a better balance of coefficients of variation than unscaled variables.

The number of farms to be stratified was quite high. Due to limited computer resources, the Fastclus procedure was first performed in order to produce an initial clustering of 150 clusters. (The Cluster procedure produces a hierarchical stratification starting its number of strata with the number of observations in the population and ending with one stratum regrouping all population units. It can be very expensive when the number of units in the population is very large.) Then the Cluster procedure was applied to these 150 clusters using Ward's minimum variance method. The final number of clusters varies from 11 to 42 depending on the province.

3.2. Samples

The sample size for the July Livestock Survey is about 30,000 units whereas that of the January survey is approximately 13,000 units. The provincial sample sizes were fixed. Part of the sample size is taken up by the three special take-all strata. For both surveys, a sampling rate of 1/60 was used in the stratum of farms with no livestock. Then, independently for the January and July surveys, allocation of the remaining sample size to the other strata was performed. Both allocations were done to optimize the precision of the estimates of the same livestock variables: beef cows, milk cows, sows, total number of cattle, total number of pigs and total number of sheep.

The 1993 July sample was selected using stratified simple random sampling without replacement. For January, a stratified simple random sample was drawn from the July sample according to its own allocation. In order to reduce respondent burden, in subsequent years, these two samples will be rotated annually at a rate of 50% for the stratum of Census farms with no livestock and 20% for the other strata.

3.3. Overlap Reduction between the Crop Master Sample and the July Livestock Sample

Controlling respondent burden has always been an important issue at Statistics Canada. Each year, a number of agricultural surveys are conducted and consequently the respondent burden can be significant. The crop and livestock surveys are a major part of the agricultural program and thus it was decided to minimize the overlap between the crop master sample and the July livestock sample. A method proposed by Kish and Scott (1971) was chosen and adapted for this purpose. In each intersection of crop and livestock strata, the units in both the crop and livestock samples are identified. As many of these units as possible are removed from the livestock sample and replaced by non sampled units in the intersection. It can be mathematically described as follows:

Let $U_{c.}$ and $U_{.l}$ be respectively the crop stratum and the livestock stratum. Let also $U_{cl} = U_{c.} \cap U_{.l}$ and let

s_{cl} = units of U_{cl} in both the crop and livestock samples
n_{cl} = number of units in s_{cl}
U'_{cl} = units of U_{cl} that are in neither the crop sample nor the livestock sample
N'_{cl} = number of units in U'_{cl}

Two different actions are possible:
(1) If $N'_{cl} \geq n_{cl}$, then n_{cl} units are selected from U'_{cl} by simple random sampling. These selected units replace the units of s_{cl} in the livestock sample.
(2) If $N'_{cl} < n_{cl}$, then N'_{cl} units are selected from s_{cl} by simple random sampling. These selected units are replaced by the units of U'_{cl} in the livestock sample.
In both cases the crop sample remains unchanged.

This method was applied to every possible U_{cl} and resulted in a reduction of approximately 60% in the overlap between the crop master sample and the July livestock sample. After rotation is applied to the livestock sample, reduction of the overlap is performed using the same method with U'_{cl} = units of U_{cl} that are in neither the crop sample nor the livestock previous and current samples.

4. AREA FARM SURVEY

The purpose of the Area Farm Survey (AFS) is to complement the list frames of various agricultural surveys, in particular the common list frames of the livestock surveys and the crop surveys. Farms that are not on this list frame include those that were missed by the 1991 Census and new farms that started operating after the Census. It will also be possible to produce from the AFS, estimates of number of farms and of total farm area.

4.1. Frame

Under the previous design, the area sample, which was a component of the NFS, was selected in two stages. The first stage was the selection of Enumeration Areas (EA) which correspond to the area that was canvassed by a Census enumerator. The selected EA's were then manually partitioned into land segments from which a sample was drawn.

Technological advances have now permitted to segment the whole country automatically, and therefore to select the sampled segments in one stage. The AFS, just as the list frame, covers all provinces except Newfoundland whereas the previous design covered only Québec, Ontario, and the CWB area.

In the Prairie provinces and in part of the Peace River district, regular cells of 3 miles × 1 mile were created. These correspond to the fairly regular legal land description in use there. Elsewhere, cells of 3 km × 2 km were constructed using the Universal Transverse Mercator grid. The cells are then intersected with the EA's. EA's with no farm headquarters cover very large areas and have little agricultural activity, therefore cells that do not intersect EA's with farm headquarters were discarded automatically. Other cells were discarded after being examined using a topographical map if, for example, they were completely inside a national park. At the same time, cells were combined when, for example a large part of them was water. Each remaining cell, or group of cells if combining was done, corresponds to one of the segments that make up the frame.

4.2. Stratification

The same subprovincial regions as those used for stratification in the crop surveys form the second level of stratification after the province. Statistics for the segments on the frame were arrived at by allocating EA Census totals to the intersecting segments in proportion to the area of intersection. A composite measure of agricultural activity based on the total numbers of cattle, pigs, and sheep, total farm area, and number of farms was computed and used as input into the Fastclus procedure of SAS to form strata within the subprovincial regions. The number of strata was such that we had on average a sample of 25 segments per stratum.

4.3. Sampling and Estimation

A sample of 2,100 segments was allocated to provinces roughly in proportion to size (total number of segments in the province) while taking into account the sample sizes under the previous design and respecting a minimum size of sampled segments in each province. Sample allocation to the strata was proportional to the square root of the stratum size (total number of segments in the stratum). Each stratum's segments were then sorted randomly within Census Subdivision (usually municipalities) before systematic random sampling was employed to ensure a good coverage of the land. A rotation rate of 25% is planned for sampled segments in each subsequent year.

All farms with land in a selected segment are enumerated. These farms are unduplicated against a given survey's list frame and non listed farms are surveyed as part of the data collection activities of that given survey. The fraction of the farm that lies inside the segment is used to arrive at segment totals.

5. CONCLUSION

The first crop and livestock surveys using the new list samples were respectively the 1992 December Farm Survey and the 1993 January Livestock Survey. Furthermore, the Area Farm Survey was first conducted in April and May 1993. Up to now, few results on the evaluation of this redesign are available, but will be produced shortly.

REFERENCES

Bélanger, Y. (1990), "Redesign of the Crop Surveys at Statistics Canada", Statistics Canada, Methodology Branch Working Paper, No. BSMD-90-014E/F.

Julien, C. and Maranda, F. (1990), "Sample Design of the 1988 National Farm Survey", *Survey Methodology*, 16, 117-129.

Kish, L. and Scott, A. (1971), "Retaining Units after Changing Strata and Probabilities", *Journal of the American Statistical Association*, 66, 461-470.

SAS Institute Inc. (1985), *SAS User's Guide: Statistics*, Version 5, Cary, North Carolina: SAS Institute Inc.

Sethi, V. K. (1963), "A Note on Optimum Stratification of Populations for Estimating the Population Means". *Australian Journal of Statistics*, 5, 20-33.

STRATEGIES FOR EVALUATING THE DATA QUALITY OF THE CANADIAN CENSUS OF AGRICULTURE

Stuart Pursey, Statistics Canada
Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A OT6

KEY WORDS: Census of Agriculture, Data Quality Evaluation

This paper discusses approaches to evaluating the data quality of the Canadian Census of Agriculture (CEAG). Section 1 describes the role that the 1991 data quality evaluation has played in encouraging development of future data quality evaluations. Section 2 discusses approaches to future data quality evaluations: considering the meaning of data quality, the purposes of data quality evaluations, and the framework of a data quality evaluation. A model of an information system provides a useful framework for a data quality evaluation. Section 3 describes one such model that has been developed in the context of an agriculture information system.

1 The data quality evaluation of the 1991 CEAG

This section discusses the data quality evaluation of the 1991 Census of Agriculture and its role in influencing the need to examine the purposes, objectives, and methodologies of future data quality evaluations.

From 1988, when the first ideas about the proposed form of the 1991 data quality evaluation took shape, methodologists and subject matter specialists expressed a sense of unease about the role of the data quality evaluation within the Census of Agriculture. There seemed to be an understanding that many traditional data quality evaluation activities had provided useful benchmark data -- but had provided little information that was useful in making practical improvements in census data quality. By 1990 a decision had been taken to concentrate on two objectives for the 1991 data quality evaluation:

- to identify issues that must be considered to improve the data quality of the 1996 Census of Agriculture and
- to consider the 1991 evaluation as useful means of planning future data quality evaluations.

The traditional evaluation activities were not ignored but less emphasis was placed upon them. These activities included recording responses rates, measuring the frequency and impact of edit changes, measuring the impact of data capture errors, measuring the frequency and impact of imputation, and measuring the differences between census estimates and other estimates from comparable data sources.

Two data quality activities proved especially useful in understanding the data quality of the 1991 Census of Agriculture. As well, they influenced the path followed in exploring and understanding data quality from the 1991 Census of Agriculture.

The first activity concerned the experiences of the Census Representatives (CRs). CRs represent our first contact in the field with the data responses of farm operators during data collection. Their experiences with the data quality were derived from focus group sessions. The analysis provided invaluable ideas about the impact of CEAG data collection activities on data quality of the 1991 CEAG.

The second activity concerned the traditional agriculture economic analysis of the final census estimates (Data Validation) before release to the public. The discussions and reports of the many staff from this exercise also proved invaluable in understanding the impact of many CEAG activities on data quality of the census estimates.

1.1 Ideas from the 1991 data quality evaluation

In the 1991 Census of Agriculture data quality developed from two types of effort. Activities before and during Data Collection **generated** a level of microdata quality. Activities at Head Office **repaired** data of poor quality derived at an earlier stage in census processing. They identified and isolated potential data problems and successfully instituted remedial action where required.

There is some concern about the relative importance that repair efforts played during the 1991 CEAG. We would rather "do it right the first time". Thus improving the data quality at early stages of the census pays enormous dividends at later stages of the census. Some of the most important earlier stages of the process include:

- the quality of the design and layout of the CEAG Questionnaire;
- the effectiveness of the CEAG public relations efforts to motivate respondents' to complete the CEAG questionnaire accurately and fully;
- the field procedures during data collection that help the CRs find agriculture holdings, obtain a completed CEAG Questionnaire, and perform the field questionnaire edits; and
- the training and motivation that the CEAG Project Team provides to the CRs.

2 Looking to future data quality evaluations

This section provides some tentative ideas on the form of future data quality evaluations. These are the major issues that must be considered: meaning of data quality; objectives and purposes of a data quality evaluation; relative importance placed in evaluating elements of the CEAG (relevance of the Census of Agriculture, conceptual framework of measured variables, accuracy of data, dissemination of data to users, and analysis and interpretation of data); and identifying who should evaluate what elements of the CEAG.

Often a statistical program is divided into two parts: the **product** is the set of goods and services provided by the CEAG to data users and the **process** is the set of procedures used by the CEAG Project Team to generate the product. The data user is called the **customer** and the CEAG Project Team is called the data **supplier**; see Colledge and March (1993) for a more detailed discussion.

2.1 The meaning of data quality

The first task is to derive an understanding of the meaning of data quality for the Census of Agriculture. The concept of data quality is nebulous. Certainly it must be understood and defined in terms of data users' needs. Colledge and March (1993) discuss the application of quality management to a national statistical agency -- and in this context quality has been interpreted to mean fitness for use from the viewpoint of the customer rather than the supplier.

2.2 Objectives and purposes

The second task is to understand the objectives and purposes of a data quality evaluation. The purposes of a data quality evaluation are closely twinned with the objectives. In the short term we want to provide data users with data quality information on the most recent statistical product of the CEAG. In the long term we want to provide data suppliers with the information they need to improve future CEAGs. Thus two objectives may be:

- to derive information, with analysis, that provides data users with an awareness and understanding of the data quality of the **product** of a CEAG, and
- to provide the CEAG Project Team with information and analysis to evaluate the quality of the CEAG **process**.

2.3 Elements of the statistical program

The statistical program we develop is based on our understanding of both the agriculture sector and the data user needs. The third task is to allocate and balance evaluation efforts among the elements of the statistical program.

Traditionally data quality evaluations have emphasized the accuracy of data. But there are many more elements:

- the quality of our understanding of user needs,
- the relevance of the set of agriculture variables that we measure for data users,
- the appropriateness of the concepts and definitions behind the agriculture variables,
- the effectiveness of the dissemination of the statistical product to users,
- the timeliness of data availability to users, and
- the effectiveness of the analytical and interpretive material provided to users.

As well, it may be appropriate to evaluate not only the effectiveness of the process but also its efficiency.

2.4 Methodological approaches

The specification of methodological approaches presupposes our understanding of the meaning of data quality and the objectives and purposes of the data quality evaluation. Nevertheless some ideas can be developed from the ideas expressed above and from current characteristics of the Census of Agriculture.

The fourth task is to translate the more abstract notions of data quality meanings and data quality objectives and purposes into specific action. This implies a set of methodological "things to do" that measure data quality.

First, we must understand user needs well enough to use them as a base for comparison to the statistical program. Thus we must develop a way to measure user needs.

Second, periodically we would want to examine the appropriateness and effectiveness of the conceptual basis of the set of measured agriculture variables. This broad issue is intertwined with the Canadian Agriculture Statistics System and thus should not be examined only within the Census of Agriculture.

Third, more directly (and traditionally) we would want to examine the accuracy of the measured variables. We might organize this in terms of the main uses of the CEAG. These are listed below, with the main quality requirements.

The provision of a **database of "Census of Agriculture" estimates** requires accurate macro-estimates by province, small area, and user-defined domains.

The provision of a **micro-database for agriculture socio-economic research and analysis** requires adequate population coverage and accurate micro-data.

The provision of **small area estimates** requires accurate macro-estimates by "small area".

The development of **list frames for the design of agriculture surveys** requires adequate population coverage and accurate micro-data from design variables.

The provision of **benchmarks for Agriculture Division series** requires accurate macro-estimates.

Examining the list above we see the importance of measuring census coverage, the accuracy of micro-data, and the accuracy of macro estimates.

The Statistics Canada **Policy on Informing Users of Data Quality and Methodology** provides a set of guidelines to follow in working one's way through the methodological components of a data quality evaluation. The guidelines include an examination of concepts, definitions, population coverage, sampling and non-sampling errors, response rates, the impact of edit and imputation, the comparability of data over time, and the comparability with data from other sources.

2.5 Distribution of evaluation tasks

The value and impact of a data quality evaluation depends upon the credibilty and independence of the evaluators. The fifth task is to allocate evaluation tasks to appropriate areas. The Census of Agriculture Project Team members, subject matter staff, methodologists, systems analysts, and auditors are several examples of the types of individuals within Statistics Canada that should be involved. But for certain tasks, perhaps an examination of the relevance of the CEAG, outside independent evaluators may provide the most effective evaluative insights.

3 An inquiry and data system -- its use in data quality evaluation

The use of a generally accepted model of an information system is appealing since it provides a clear structure for a data quality evaluation. This section explains a particular information system that has been developed in the context of agriculture.

Figure 1 shows a diagram of this information system. It was developed by James T. Bonnen (1975) and first appeared as part of his presidential address at the American Agriculture Economics Association Annual Meeting in 1975.

The ultimate purpose of a statistical program is to improve our understanding of reality. Yet reality is hard to understand and therefore we simplify by developing theoretical concepts to represent it.

Theoretical concepts in agriculture include, for example, the **production** of agriculture products, the **flows** of agriculture products through the economy, the **use** of capital, the social **characteristics** of people involved in agriculture, and the economic **condition** of the sector and its people.

Theoretical concepts are usually abstract and so we develop an operational structure for them, explicitly identifying a set of agriculture variables to be measured and defining what we mean by them.

We might understand the economic conditions of the sector and its people by choosing the measurement variables net farm income, off-farm income, changes in the value of inventories, and the value of agricultural capital. These variables should be explicitly defined with reference dates and within an accounting framework.

Next we measure these variables, thus producing agriculture data. Then we output data to users through a system of data dissemination. There it is analyzed and interpreted, turning data into information for decision makers.

The left hand side of the diagram refers to the data system that is developed by the statisticians (data suppliers). The right hand side refers to the inquiry system developed by the analysts (data users). The inquiry system and data system must meet on a common conceptual ground. Together, the inquiry and data system comprise the information system.

Given the implementation of a particular data system -- the Census of Agriculture -- we wish to evaluate its quality. By following the structure of the data system we develop a framework for a data quality evaluation. That is, we measure and understand the closeness of the statistical program (its theoretical concepts, operating structure, measured variables, and data dissemination) to the needs of data users as represented by the inquiry system.

4 Concluding remark

This paper has outlined the approach we are taking to developing a framework for the 1996 CEAG data quality evaluation. It includes considering the meaning of data quality, understanding objectives and purposes, balancing evaluation efforts, developing methodological approaches, assigning evaluation efforts to appropriate organizational areas, and exploiting the usefulness of a model for an agriculture information system.

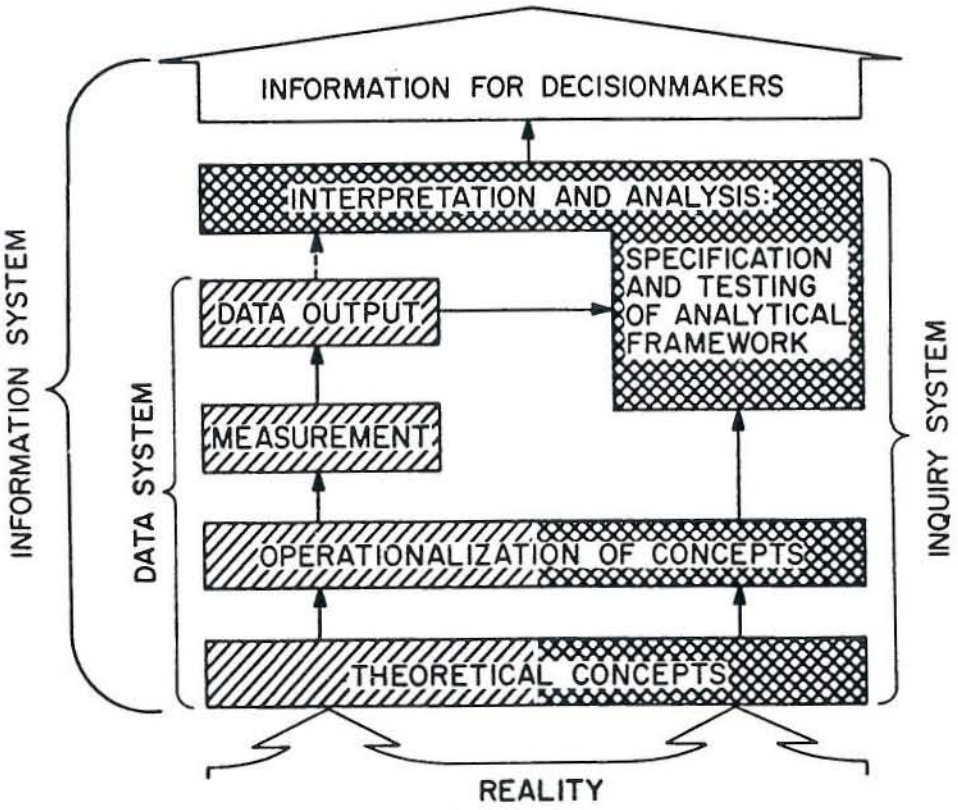

Figure 1: An agriculture information system
(reproduced with the permission of the American Journal of Agriculture Economics)

References

BONNEN, JAMES T. (1975) "Improving Information on Agriculture and Rural Life"; American Journal of Agriculture Economics; 57:753-763.

COLLEDGE, MICHAEL and MARY MARCH (1993) "Quality Management: Development of a Framework for a Statistical Agency"; Journal of Business and Economic Statistics; Vol 11, No. 2: 157-165.

THE SAMPLING DESIGN OF FINNISH AGRICULTURAL SURVEYS

Paavo Väisänen
Statistics Finland, P.O.Box 504, FIN-00101 Helsinki

Key words: Agricultural surveys, stratified sampling

1. Introduction

The Finnish National Board of Agriculture and Statistics Finland, the national statistical institute, conduct agricultural surveys by means of the same sample. The Board of Agriculture uses the sample to collect data on cereal crops, livestock production and the labour input of farmers, while Statistics Finland uses it to collect data on agricultural income. Sample size is 16,000 farms. Additionally, part of the sample is used for a survey in June, which is conducted to produce estimates of the sowing areas of cereals.

The data on production and field area are collected from the farms by interviews and mail questionnaires, or they are obtained from the Farm Register, which also serves as the sampling frame. The data for income statistics are collected from tax return forms, sparing the farmer the burden of responding to a questionnaire. The surveys are conducted once a year. They were based on separate samples up until 1985. From then on they share a sample, bringing savings in sampling costs and enabling compilation of input-output tables, plans for which are under preparation.

The distributions of the variables under study are skew, for which allowance is made through stratification. The stratification variables are rural district, area under cultivation and production sector. There are great variations in crops and income between different years. The climatic conditions in the south and in the north of Finland differ considerably, influencing agricultural production in that the best conditions for growing cereal crops exist in the south. In the northern, eastern and central areas of the country, cattle farming is the most important production sector. Pig farming is also regionally concentrated, and reindeer farming is carried on in the northernmost areas of the country. By combining the production sector and the variable describing the region or the location of the farm, the sample can be allocated in a manner that permits the income, output and input variables to be estimated from the same sample. The sample is not the optimum for any one variable but represents a compromise between the variables, producing a reasonable result under all conditions. Variations between years due to sampling can be reduced by using the rotation design, in which part of the sample remains unchanged from one year to the next. In the rotation design, about one-third of the sample is changed each year, the same farm thus remaining in the sample for a period of three years. Large farms of 100 hectares or more stay in the sample on a continuous basis, though a proportion of them, too, is changed as from 1992.

The farms are allocated to the strata using the Neyman method, with variances calculated for agricultural income, a commensurable variable suitable for all farms. Incomes vary a great deal from year to year, which influences the sample sizes of the strata. Problems arise when the allocated sample size does not correspond to the exiting panel's sample size. The register would also allow the field area to be used as the allocation variable, as was done earlier, but with the production sector selected as a stratification variable, production in the strata related to livestock farming does not depend on the field area.

The parameters estimated are crop totals, the mean and median of incomes and working hours, and the standard deviations of these. As auxiliary information, the income totals of the population are included into the estimation by means of ratio estimators.

2. Stratification

The sample is drawn using stratified simple random sampling without replacement (STRWOR). The Farm Register is used as the sampling frame. It comprises all farms and is updated annually. The register of a given year is usually completed by March the next year, which means that the sample is drawn from an up-to-date register. The information on farms in the Farm Register includes such items as field area, land use, forest area, production sector, certain administrative divisions, address, and a number of items concerning the owner and the farmer. The Farm Register is maintained by the Board of Agriculture. Rural district was selected as the regional stratification variable.

Finland is divided into 18 rural districts. Production sector was selected as the stratification variable describing the production of farms with a field area under 100 hectares. As the register's 23 production sectors constitute too detailed a classification for stratification purposes, certain sectors of livestock production were combined to form one sector. For example, the sectors of piglet production, fattening pig production and other pig farming were combined to form the sector of pig farming. Similar procedures were applied in the case of cattle and poultry farming as well. The fruit and vegetable production sectors were combined. Forestry, farm tourism and certain other heterogeneous production sectors of minor importance were combined to form one sector. The sector of cereal production was divided into five substrata by field area. Farms of a hundred hectares or more formed separate strata, to which the division into rural districts was not applied but which were combined to form three major areas: the south, the west and the rest of Finland. By combining the rural districts and the production sectors a total of 178 strata were formed. Some strata turned out to be too small in practice. In this year's sample, adjoining rural districts were merged in production sectors containing only a few farms, bringing the number of strata down to 153.

Inclusion into the target population was determined on the basis of the field area. In the cereal production sector, the target population comprised all farms with three hectares or more of arable land. The other production sectors had no such limitations. The population consisted of approx. 125,000 farms.

3. Allocation of the sample

The sample cannot be allocated only on the basis of the distributions of variables. First of all, the proportion of farms of 100 hectares or more is large in the cereal production sector. Therefore, such farms were assigned to a separate stratum and were all included in the sample. As from 1992, the allocation was changed so that 60 per cent of large farms in the south of Finland, 80 per cent of those in the west and 100 per cent of those in the north and east are included in the sample. Neyman allocation was applied to the other strata, with taxable agricultural income, excluding income from forest transactions, used as the allocation variable.

The allocation calculations were performed using the following equation (Cochran 1977):

$$n_h = \frac{N_h s_h^*}{\sum_{h=1}^{L} N_h s_h^*} n_t \qquad (1)$$

where h = 1, 2, ..., L and s_h^* = the dispersion of income in the stratum as calculated from the sample of the preceding year, n_t = the size of the new panel = 16,000 - $(n_{t-1} + n_{t-2})$.

As allocation based on the dispersion of income is less than satisfactory from the point of view of individual agricultural quantifiers, such as the field area, the figures in the allocation calculations were adjusted with the help of agricultural census statistics.

In addition, changes in sample size from 1991 to 1992 were taken into account. In the 1991 sample, province and farm size were the stratification variables. As the revised structure of stratification and the new allocation variable affected sample size, a strict Neyman allocation design was not used in order to avoid too great changes. The uncertainty of allocation was compounded by the fact that the income data related to the year 1988 because the preparation of taxation data by the National Board of Taxes was delayed in 1989, 1990 and 1991. The rotation of the sample by one-third did not succeed in the revised stratification design: in some strata, more than one-third of the farms were removed, to be replaced by only a few new ones; in some others, all farms that had participated three times were removed, to be replaced by more farms than had been removed. In co-ordinating Neyman allocation and the rotation design, one aim was to ensure that rotation would continue as planned. This meant that the sample size of a stratum was limited to three-quarters of the size of the population at most.

The co-ordination of Neyman allocation and the rotation design creates problems every year: as incomes vary from year to year, the number of farms to be removed and the number of farms to be selected do not agree at the level of the strata. The panels selected in 1990 and 1991 did not agree with Neyman allocation in 1992. In three years' time, however, the sample should have changed so as to agree, more or less, with Neyman allocation. This means that 1994 will see a sample whose allocation by income agrees with the Neyman method.

4. Estimation

The business statistics of the Finnish farm economy provide information on agricultural income. In estimating the data according to stratified sampling, the following formula is used:

$$t_{ySTR} = \sum_{h=1}^{L} N_h \, y_h \qquad (2)$$

where $y_h = \Sigma \, y_{hi} / n_h$.

The total of the agricultural income of the farms are estimated by means of a separate ratio estimator:

$$t_{yR} = \sum_{h=1}^{L} \frac{t_{yh}}{t_{xh}} T_{xh} \qquad (3)$$

where t_{yh} and t_{xh} are the estimated totals in the stratum h and T_{xh} is the total income of the farms in the stratum as calculated from the taxation register.

In the production statistics, the data on field use are estimated by means of a separate ratio estimator (3) in which the area under cultivation, x, is used as an auxiliary variable and where T_{xh} is the total of x in the stratum as calculated from Farm Register data.

A corresponding estimator is used in estimating livestock production. The number of farm animals, the data on which are obtained from the livestock file of the Farm Register, is used as an auxiliary variable. The data on crops production are estimated using the estimator of the total (2) for stratified sampling.

The standard errors and the variation coefficients (CV) of estimates are calculated according to the STRWOR design. Standard errors are published only for some variables (National Board of Agriculture (1992) and Statistics Finland (1993)).

The estimation process makes no use of rotation design data. No statistics are produced on parameters describing change, such as differences between two years; calculation of changes is up to the individual user of the statistics.

5. The sub-sample used for the survey on areas under cultivation

A compilation of preliminary statistics on field area use is released every year, providing information on the areas under different cereals. Data collection takes place in late May or early June. The sample size of 1,500 farms ensures fast results. (This year's results were published on Friday, June the 18th.) The statistics are used for forecasting annual crops.

The sample is selected from the sample of the previous year's agricultural surveys. The sample design is based on two-phase stratified sampling. Region and field area are the stratification variables, each of which has three classes, yielding a total of nine strata. The sample is allocated, by the Neyman method, according to the dispersions of the totals of wheat, barley, rye and rape, with rape replaced by hay and ensilage for areas in northern Finland. The allocation design was revised this year: after allocation of the sample by the Neyman method according to the different cereals, the mean of the sample sizes is calculated, which then is used as sample size.

In the first phase, the sampling probability π_k' agrees with the stratified sampling design and is constant in each stratum h:

$$\pi_k' = n_h / N_h$$

In the second phase, the sampling probability π_k'' agrees with the stratified sampling design and is constant in the new stratum of the second phase

$$\pi_k'' = n_l'' / n_l$$

where n_l is the number of farms belonging to the first phase sample in the new stratum l, and

$$n'' = \Sigma \, n_l''$$

is the sample size of the second phase.

The sampling probability π_k of unit k is

$$\pi_k = \pi_k' \pi_k'' = (n_h / N_h)(n_l'' / n_l)$$

(Särndal et al., 1992).

The total is estimated using the estimator
The areas under cereals are estimated using a

$$t_{y\pi} = \sum_{k=1}^{n''} \frac{y_k}{\pi_k} \qquad (4)$$

combined ratio estimator in which t_x', the total estimated from the sample of the first phase, is used as auxiliary information

$$t_{Ry} = (t_{y\pi}''/\, t_{x\pi}'')\, t_x' \qquad (5)$$

We have not calculated the design-based standard error for this estimator.

6. Evaluation of the sampling design

The distribution of the variables describing agricultural production and income is skew. Production depends on the farm's field area and geographical location. Finland has approx. 500 farms with 100 hectares or more under cultivation. Stratification can be used to focus the sample to such subgroups of the population which show large variations in the variable. In multipurpose surveys, allocation requires a variable which depends on all the variables under study. Variation by income level does not necessarily correspond to variation by crops or by hours worked, which is a function of many different factors. The surveys contain a large number of variables, and standard errors have only been calculated for the most important ones. The coefficients of variation are used to describe the sampling errors of the estimates. Design effect (deff), i.e. the ratio of the variances of the stratified sampling and simple random sampling without replacement

$$deff(t_y) = \frac{V_{STRWOR}(t_y)}{V_{SRSWOR}(t_y)} \qquad (6)$$

was used to express how well the stratification and allocation function for the variables under study.

The statistics describing agricultural income in 1992 will not be completed until the completion of taxation in 1994. The most recent statistics available on agricultural income relate to the year 1990, when the sample was stratified and allocated in a different manner from 1992. The table below shows that the relatively large sample size results in small coefficients of variation at the national level. Considered by region, the coefficients of variation are influenced most by the number of farms in the sample representing the region and by the regional distribution of the variable.

Table 1
Variation coefficients of agricultural incomes by province in 1990

Province	Dairy %	Crops %	Incomes %
Uusimaa	10.3	4.0	3.1
Turku and Pori	8.2	3.1	3.4
Häme	6.5	8.2	4.0
Kymi	5.3	5.1	3.5
Mikkeli	6.7	15.9	5.8
Northern Karelia	10.3	23.7	4.9
Kuopio	7.9	25.9	5.5
Central Finland	10.6	16.7	6.0
Vaasa	9.0	9.6	4.3
Oulu	6.6	21.4	4.7
Lapland	7.0	44.8	4.8
Åland Islands	18.2	10.9	7.4
Total	**2.8**	**2.5**	**1.5**

(Statistics Finland, 1993)

In Tables 2 and 3 the totals of field areas and crops in the survey of 1992 were calculated using the estimator (2).

Table 2
Harvested area: total, CV, deff

Cultivated plant	Est. tot 1000 ha	CV(t_y) %	deff(t_y) %
Winter wheat	12.5	6.4	0.6
Spring wheat	78.1	3.4	0.7
Rye	11.6	6.5	0.6
Barley	45.6	1.7	1.0
Oats	349.7	1.8	1.3
Cereals, mixed	9.5	11.3	2.4
Peas	16.0	4.9	1.1
Sugar beet	33.1	4.4	0.8
Hay	219.8	1.7	1.8
Green fodder	37.5	4.7	1.8
Turnip rape	79.3	2.7	1.5
Potatoes	34.6	8.7	1.8

Calculated by SUDAAN 6.30

The deff figures for the estimated totals of harvested areas and crops were calculated using SUDAAN software (Shah et al., 1991). For winter wheat and rye the deff statistic of harvested area was 0.6 - 0.7, which shows that stratification gave gain in the estimation of these cereals. Design effects close to one (0.8-1.3) were obtained for barley, oats, peas and sugar beet. Here the sample design is as effective as SRSWOR. Deffs of 1.8

or over were obtained for mixed cereals, hay, green fodder and potatoes. The results were much the same for crops. The deff figures were little lower for crops than for harvested areas as can be seen from Table 3.

Wheat and rye were cultivated on in the southern part of the country, on farms with cereal production as the production sector, which was divided substrata according to cultivated land. Farms of 100 hectares or more formed a separate strata with three regions, to which special allocation was applied (see sec. 2). Barley and oats are cultivated in the whole country and they are produced for both bread and fodder grain. Thus a farm's production sector may be other than cereal production, which is why farms are selected to the sample from the cattle or pig farming stratum, for instance. This gives high variances in the strata for barley and oats. The same is true of hay, green fodder and mixed cereals. The most important areas for potato cultivation are in the southern and western parts of the country. Potato production penetrates the strata and half of all sampled farms had values for this variable.

Table 3
Yield on cereal crops: total, CV, deff

Cultivated plant	Estm. tot. 1 M kg	CV(t_y) %	deff(t_y) %
Winter wheat	33.5	6.3	0.5
Spring wheat	171.9	3.2	0.6
Rye	26.8	6.7	0.6
Barley	1188.9	1.7	1.0
Oats	954.8	1.7	1.2
Cereals mixed	23.7	14.3	2.1
Peas	25.5	5.4	0.8
Sugar beat	799.2	4.8	0.7
Hay	569.4	2.5	1.6
Green fodder	338.3	7.8	1.7
Turnip rape	136.7	2.6	1.5
Potatoes	573.4	10.0	1.7

Calculated by SUDAAN 6.30

Areas and crops vary within the strata in regard to potato cultivation. As the Farm Register does not contain a suitable variable for stratification or allocation, SRSWOR would have been a better design for potatoes than the STRWOR used. Neyman allocation by incomes may be the reason for the high deff values for turnip rape, sugar beet, and other crops with low frequencies in the population.

The calculations of the values of livestock variables showed a high deff figure (2.3) for horses and foals.

Table 4
Livestock: number of farms (N), estimated total, CV and design effect

	N	Total (1000)	CV	Deff
Horses	550	23	7.9	2.3
Foals	168	4	14.4	2.2
Cows	3 909	468	1.3	1.9
Cattle	5 313	1 391	1.2	1.2
Sheep	400	58	16.2	2.0
Pigs	1 545	1 391	2.5	0.6
Hens	1 273	5 474	8.3	0.4

Calculated by SUDAAN 6.30

Poultry farming was concentrated on the west coast. CV for the estimated total of hens was high, 8.3, whereas deff statistic was small, 0.4. A result like this is due to the small sample sizes and homogeneous farms in the strata.

The stratification was too dense for variables with small frequencies, such as foals and sheep which had high CV figures (foals 14% and sheep 16%). The strata were collapsed in the estimation, but within the collapsed strata the deff figures became high. Production sectors were combined within rural districts: for instance, pig farming, poultry and mixed livestock farming were combined into the single stratum in northern Finland, which increased variation.

Table 5
Livestock: The highest and lowest values of CV and deff in the rural districts

	CV %	Deff %
Horses	18.8 - 50.0	0.9 - 7.1
Cattle	4.3 - 13.8	0.5 - 2.5
Sheep	19.5 - 89.9	0.6 - 4.6
Pigs	4.5 - 78.6	0.2 - 5.3
Hens	10.6 - 90.9	0.2 - 4.2

Calculated by SUDAAN 6.30

The number of observations influences the coefficients of variation, but not necessarily the deff values. Frequencies were small in some rural districts, which have caused the high CV values. The estimated totals of livestock Table 5 have also been published according to rural districts (National Board of Agriculture, 1992).

When comparing the estimates of totals with the published ones (National Board of Agriculture, 1992), small differences can be observed. These are due to the fact that in these calculations the latest version of the Farm Register was used in weighting and that the ratio estimator was not used.

The sample had too many strata, and for a number of variables there were not enough observations in some strata. In the sample for the year 1993 the stratification was reduced to 153 strata.

References

Cochran W.G. (1977). *Sampling Techniques.* 3rd Edition. John Wiley & Sons. U.S.A.

National Board of Agriculture (1992). *Monthly Review of Agricultural Statistics*, Numbers 6, 7 and 11/1992. Helsinki: Government Printing Centre

Särndal C.E., B. Swensson and J. Wretman (1992). *Model Assisted Survey Sampling*. Springer-Verlag

Shah B.V., Barnwell B.G.,Hunt P.N, LaVange L.M. (1991). *SUDAAN User's Manual, Release 5.50.* Research Triangle Park, NC,27709

Statistics Finland (1993) *The Business and Income Statistics of the Farm Economy 1990.* Agriculture and Forestry, 1993:1. Helsinki: Government Printing Centre (in Finnish and Swedish).