Proceeding of Statistics Canada Symposium 90
Measurement and Improvement of Data Quality
October 1990

THE QUALITY OF DATA IN CROP YIELD SURVEYS

Ron Fecso[1]

ABSTRACT

Crop yield surveys conducted by many countries involve several stages of subsampling which are determined by the enumerator and often repeat visits to the ultimate sampling unit. These procedures and the often difficult measurement tasks associated with the field work provide ample opportunities for data quality problems. Studies of typical problems uncovered in surveys done in the U.S. will be reviewed. The management strategy for improving the data will be reviewed along with results of recent quality measurement studies.

1. INTRODUCTION

Agriculture is the largest economic activity in many countries and crop yield is often the most variable year-to-year component of agricultural production. Thus, crop yield and associated surveys have been of interest for some time. For example, Francis Bacon looked at relationships between weather and agricultural fluctuations in the 17th century. Through the middle of this century research on crop yield measurement flourished. The methods developed over 30 years ago to measure crop yield, although often modified, remain similar to those used today in many countries.

This paper focuses on one method of crop yield measurement, called objective measurement or crop cutting, as practiced by the National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture. The main purpose of the paper is to review and begin documentation of methodological studies which are not widely known. This documentation is important to us because we did not have much research activity in these surveys for the past several years. As we are now building a new research group to study yield measurement, and most of the staff is new, a baseline documentation of existing work will help them focus on important areas to study. We hope that this work will also encourage dialogue with other countries concerning their experiences with yield measurement.

1.1 Quality Focus

Statistical agencies are joining industry and service organizations in their interest in quality philosophies. NASS began a limited total quality management (TQM) effort using objective yield survey data as a demonstration project about five years ago. The demonstration was well received by management, but another data series was picked to be the first formal application because it was a bigger management concern. Hopefully, those TQM developments will be revived along with the yield research program. Thus, some of those TQM developments will be incorporated in this discussion. Basically, the approach focuses on how to provide adequate information to determine where to look to improve data quality from a complex survey design with a complex estimator knowing that resources for research are limited.

---

[1] Ron Fecso, Research and Applications Division, National Agricultural Statistics Service, U.S.D.A., Washington, DC 20250-2000

2. CROP SURVEYS

This section will provide a brief background on crop yield surveys. Yield data is important since is affects commodity trading prices on exchanges, allocation of transportation services (rail cars, barges, etc.), agricultural credit, and a variety of commercial and government programs. There are two basic methods of collecting yield data: reported data and objectively collected (field enumerated) data.

Yield data reported by growers and other individuals knowledgeable about their local agricultural conditions dates back to 1862 in the U.S. These surveys are usually done by mail and ask respondents to judge observed condition of the crop or expected yield. Further detail on these surveys can be found in Fecso (1990). This discussion will concentrate on the objective yield (OY) surveys.

2.1 Objective Yield Surveys

Collection of data from growing sites forms the basis of OY surveys. NASS conducts these surveys for many major crops (corn, soybeans, cotton, wheat, rice, and potatoes) and specialty crops (fruit and nut tree and grapes). A broad generalization of the survey design, which varies by crop, will be given here. For further detail, see Francisco, Fuller and Fecso (1987).

The basic survey is an annual area frame survey. A stratified sample of some 15,000 area segments of about one square mile in agricultural strata is used to locate fields of the crops of interest. Acreage is estimated from this survey and the fields are subsampled for use in OY measurement. The subsampling scheme results is a self-weighting sample. Usually, there are under 2,000 fields selected per crop, restricted to the largest producing states (usually 10 states which account for over 80 percent of the crop's production).

In each field selected, two plots, called units, are selected using a random rows and/or paces method. Unit size depends upon the crop. Tightly spaced plants like wheat have units of about one square yard while crops like corn use 15 foot long rows. Each enumerator is assigned about 15 samples which may be divided among several OY crops in their area. The small assignment which is necessary to assure timeliness of survey results makes measurement of enumerator effects difficult.

Measurements taken at the units in the field vary by crop and month of the survey. For example, the end of season yield for corn is determined by the following use of collected data:

Estimate of gross yield per acre of corn for the field

$$= \frac{\text{(43,560 square feet in an acre)}}{\text{(Length of all rows in the sample [4 x 15 = 60)] (Width across 8 rows /8)}}$$

$$\times \; [(\text{Number of Ears in Unit 1}) + (\text{Number of ears in unit 2})]$$

$$\times \; \frac{\text{Weight of Ears husked in the field}}{\text{Number of ears husked in the field}}$$

$$\times \; \frac{\text{Weight of Grain Shelled}}{\text{Weight of Ears Shelled}}$$

$$\times \; \frac{\text{(Grain weight tested for moisture content) (1-[Moisture \% } \div \text{ 100])}}{\text{(Ear Weight in Bag - Bag Weight) (Shelling fraction)}}$$

$$\times \; \text{(Adjustment of weight to bushels)}$$

The detailed end of season measurements are made to facilitate forecasting. In earlier months, the plants have not developed the characteristics measured at maturity. The plant characteristics which are present at the time, the most highly correlated with final measurements, and reasonably easy to collect are measured on the units and used to forecast the components of the product above. For example, in early months, the count of stalks is used to predict future ear numbers. More detail on the models can be found in Reiser, Fecso, and Chua (1989) and Reiser, Fecso and Taylor (1987).

The above samples provide data to measure the gross or biological yield. Economic yield, the statistic of interest, requires an adjustment for loss due to harvesting. In a half sample of fields, two additional units are located after harvest. The grain left in the unit is gleaned and expanded to a per acre basis.

3. METHODOLOGICAL STUDIES

There is a long history of methodological study of objective yield measurement, e.g., Mahalanobis (1946) and Zarkovich (1966). The intent of this section is to outline some of the findings at NASS which are not likely to be known outside of NASS. The presentation will be organized by the components of the yield estimating equation. For each component, a brief description of problems or methods developed will be given. This is not a complete profile of all errors, although this is a long range goal, but a view of the complexity of controlling data quality in these surveys.

3.1 Unit Placement

The two units are located in a field using a random method designed to give all plants nearly an equal chance of selection. NASS units are determined by a randomly chosen row and then a random number of paces into the field along the row. (When rows are not visible a random number of paces along the edge of the field is substituted.) The random number of rows and paces is determined by the acreage of the field. An algorithm determines the random numbers when the fields are selected. Selection is limited to the quarter of the field which is most accessible. This restriction has not shown biasing effects when studied. Plants along the edge and in tractor turn rows are a concern. The algorithm was changed several times in the 50's and 60's to ensure a reasonably uniform distribution of units near the edges and in turn rows.

In tree crops, another level of randomization is necessary. Enumerators select random numbers to determine a path up the limbs to an end branch for counts. Strict procedures for marking the tree and random number selection are needed to avoid preselection of easier to reach limbs.

Pacing along a row provides the enumerator with the opportunity to create bias. Seeing the condition of the field ahead can result in shorter or longer steps which reach a visually appealing place. A buffer space is used to help reduce the possible bias. When the last pace is completed, the enumerator measures or places a pole down at the tip of their toe. The pole is usually five feet long. (For some crops, a tape measure is used to mark the five feet buffer.) The end of this buffer marks the start of the unit to be sampled. A controlled experiment in wheat indicated that counts using a buffer are lower then counts not using a buffer. Differences, on average, have been observed between the counts in unit one and unit two, another indication that subtle biases in selection can occur using pacing.

3.2 Acreage Measurement

The yield per acre in a field is determined by adjusting unit yield to a per acre basis. Unit size varies by crop. Units in fields where rows are evident have lengths of fixed size, but row widths vary. In fields where rows are not obvious, a fixed-size rectangular unit is created.

The area expansion to a per acre bias is subject to errors from both the length and width of the unit. Remeasurement of plots made using a tape to measure fifteen feet past the five foot buffer found several plots 20 feet long. Plots using fixed metal frames of about three feet in length may have end point inclusion or exclusion of plants which is effectively a change in plot length.

Row widths are variable. To control the variability, enumerators make two measurements, across one and four row spaces, for units used to determine gross yield. The measurements are across one and five row spaces in additional units for harvesting loss measurements. The ratio of the four to the one row measurement had an interesting distribution. There was a large spike at four with a normal, small variance, error around four. There was also a spike at five. (The five to one post-harvest data behaved conversely.) It was speculated that instructions to measure between five rows (four spaces) may have been confused or pre- and post-harvest instructions were mixed. A year later this phenomena was mentioned in training classes. With no other changes in training, the extra spike disappeared.

3.3 Counts in the Unit

Different agronomic characteristics are measured or observed within the unit depending upon the crop and month. Fatigue and other task difficulty contribute to error in these data. Counts tend to diminish on average as work proceeds within fields, indicating fatigue problems associated with repeated counting of large numbers of items. Other tasks are difficult. Identification of maturity stage can be difficult at the borderline of a maturity category. An error in maturity stage detemination contributes to measurement error in the forecasts. Counting nodes and lateral branches on soybean plants is difficult. Measuring length of corn ears or kernel rows, finding all the cotton bolls, and determining kernel formation are other difficult measurements.

3.4 Weight Determination

Weight per fruiting body is needed. To obtain this statistic, a subsample of fruiting bodies from the units in the field is sent to a state office for further measurements. The subsampling can cause errors. For example, the average weight of corn ears sent to the state office has been found to be larger than the average weight of all ears in the unit. Counts of wheat heads in the field sometimes vary considerably from those counted in the office. Finally, lab equipment must be controlled to obtain correct weights, moisture content, and grain to other-matter ratio.

3.5 Harvest Loss

Many of the problems mentioned above may occur during the harvest loss measurement. Other errors include losses from birds between harvest and the sample visit and finding all the grain on the ground, especially when a tire has run over the area.

4. ONGOING QUALITY EFFORTS

The assortment of typical data problems just presented leads to the question of what to do about the problems. NASS uses several activities to help control and reduce data error, including training, quality control methods and procedures, and methodological studies and improvements.

4.1 Training

NASS uses a train the trainer approach which follows its organizational structure. A national training session is conducted to train the statisticians from the state offices who will be in charge of the survey for their state. Often, the training is several months prior to the start of the survey. These statisticians then hold a training session in their state. They repeat the material to train the enumerators and supervisory enumerators.

There are several areas of concern with the training program. First, the consistency of enumerator training between states can be a problem in a train the trainer system. Recent budget cuts have added to this problem. The initial yield training session has been combined with training for another survey. Also some of the more expensive training supplies such as color inserts for training manuals have been eliminated.

On the bright side, there is some growing interest in using more formal evaluation methods in our training program. Concepts quizzes are used more often and evaluations of the results are stimulating discussion about optimal training procedures. For example, it appears that little is gained from allowing more than three hours of pre-training home study by enumerators. Also, quiz score improvement is most related to over two years experience. Finally, a general survey of enumerator's opinions about job needs and concerns has been done. Interesting insights are expected from the analysis which is now beginning.

4.2 Quality Control Measurements and Procedures

Quality control methods for survey operations which are typical of those used in many survey organizations are also used in NASS OY surveys. Since the first operational OY surveys of corn and cotton in 1961, NASS has produced many of the usual quality measures. These measures include refusal rates, inaccessible rates, samples per enumerator, cost per sample, sampling error, and edit failure rates.

The main real-time quality control procedures are supervisory enumerator rechecks of work and edits of completed forms. Supervisory enumerators return to a sample done in the first day's work of each enumerator and one randomly selected sample of later work to recheck counts. This process is mostly a continuation of training as the enumerator often accompanies the supervisor. Too little data is collected for statistical adjustment of the full survey data. Also, some counts change naturally as the plant matures between time of the visits.

Data is key entered and verified as the survey progresses. Edits of the data are done in batch mode periodically throughout the survey. Edits currently are simple range and consistency checks. An area with potential to improve data quality is the development of smarter edits. For example, looking for outlier and leverage points in regressions of variables such as weight and number of fruiting bodies. Range checks can vary based on categorical information, for example, narrow row soybean yield has a higher mean than wide row plantings. Data capture with portable data recorders was studied seven year ago and is again a research project. Data transmission problems which halted past work are mostly resolved, but cost may still be a limitation given the small sample size per enumerator. Computerization of weight measurement in the state office has been successfully initiated, thus reducing transcription errors.

4.3 Methodological Studies and Improvements

NASS uses research studies designed to test specified hypothesis about errors or to test new procedures as well as general validation studies to guide the process of survey change for improvement. Examples of the results of special studies have been presented. Validation studies have a long history at NASS. The design of these studies, prior to 1987, typically consisted of one or two states with 16 to 32 fields selected. The units in the field would be replicated up to 32 times per field to measure procedural error sources. Although results have varied by state, year and crop, a meta analysis of these studies by Warren(1985) indicates procedural bias of 6 to 9 percent for corn. Although not formally documented, similar results are found in soybeans. These biases remain despite thirty years of study.

The recurring bias level led us to evaluate our validation and research methods. With yield being the product of many measurements, all of which had several potential sources of error, we questioned the ability of control procedure to achieve the state of "do it right the first time." Fortunately, we had an alternative. A double sampling approach to adjust for bias was developed. (Fecso,1986) Briefly, the new design collects actual yield for a subsample of fields by weighing the harvested grain. This is a rare example of a survey in which a true value can be obtained. Bias is estimated as the difference between the survey procedures and the harvested weight for the paired observations.

This new validation approach has been used for soybeans with the following results:

YEAR	Estimated Bias (bushels)	Standard Error of Estimate
1987	2.2	.9
1989	1.9	.8
1989	3.2	.9

The design met its main criterion, to detect biases of over 5% (about 1.6 bushels). The estimated biases for these three years have been in the 6 to 9 percent range. Large shifts in the bias will also be detectable, an important feature when survey conditions are changing over time. Preliminary data for 1990 indicates that a shift may have occurred.

Although this double sampling estimator puts the survey estimate on a fully statistical basis (rather than assuming some bias level or inferring from two states to all states), the variance of the estimate is larger than we'd like. Further work on allocation of resources can help as some aspects of the survey appear to be oversampled compared to others.

We may want to pursue further development of process control methods for survey operations as we restaff for yield research. Process control techniques used in industry can be modified for survey use. Simple time series of measures like the four-row to one-row measurement or the unit one to unit two difference can provide clues to changes in the essential survey conditions. More complex statistical techniques such as LISREL (Reiser, Fecso and Chua, 1989) could also be useful.

5. CONCLUSION

The intent of this paper is to provide a feel for the number and complexity of tasks associated with the visually and mechanically measured values needed to measure yield. Interested readers are encouraged to contact the author for more detail and especially to open dialogue about their experiences with crop yield measurements.

Total quality management principles can be applied to our renewed yield research efforts. We need to determine our most important quality problems so we may better balance the need for increased data and accuracy with the pressure of budget and personnel limitations and the desire for more timely data. The shifting emphasis on these elements over time creates changes to the survey which often create bias in the estimate, sometimes unexpectedly. Thus, we look to techniques such as the validation methodology and process measurements to point out accuracy changes and to help us detect important error sources. If successful we will have an efficient way to direct our limited resources to develop needed methodological improvements in accuracy, ease of conduct or cost reduction. Besides these "quality control" ideas, there is anticipation that exploring the panel, spatial and multivariate nature of the data will bring further efficiencies.

6. REFERENCES

Fecso, Ron (1986), "Sample Survey Quality: Issues and Examples from an Agricultural Survey,"Proceeding of the Section on Survey Research Methods, American Statistical Association.

Fecso, Ron (1990), "A Review of Errors of Direct Observation in Crop Yield Surveys," in publication review.

Francisco, C. A., W. A. Fuller and R. Fecso (1987), "Statistical Properties of Crop Production Estimators," Survey Methodology, Vol. 13, No. 1, 1987, pp. 45-62.

Mahalanobis, P. C. (1946), "Recent Experiments in Statistical Sampling in the Indian Statistical Institute." Journal of Royal Statistical Society, 109(4), pp. 325-370.

Reiser, M., R. Fecso, and M. Chua (1989), "Some Panel Aspects of the Objective Yield Survey," Proceedings of the Section on Survey Research Methods, American Statistical Association.

Reiser, M., R. Fecso, and K. Taylor (1987), "A Nested Error Model for the Objective Yield Survey," Proceedings of the Section on Survey Research Methods, American Statistical Association.

Warren, Fred B. (1985), "Corn Yield Validation Studies, 1953-83," Washington, D.C., USDA, SRS, Staff Report No. YRB-85-07.

Zarkovich, S. S. (1966), *Quality of Statistical Data*, Rome, Food and Agricultural Organization of the United Nations.