# ASSESSING NONSAMPLING ERRORS IN SURVEY DATA THROUGH RANDOM INTERCEPT MODELS

Dale Atkinson, Research Division, National Agricultural Statistics Service, U.S. Department of Agriculture
3251 Old Lee Hwy, Room 305, Fairfax, Va., 22030

**Key Words: Sample Survey; Nonsampling Error; Reporting Variability; Random Intercept Models.**

## INTRODUCTION

Survey-to-survey differences in reporting by identical sampled units can result for various reasons. Changes in reported inventories reflect real change confounded with nonsampling error. Despite our best efforts to avoid them, nonsampling errors affect the data at various stages of the survey process, resulting from actions before, during and after the interview.

Errors at interview time are attributable to the enumerator, the respondent or an interaction of the two. One potential source of survey-to-survey reporting differences is a change in respondent from one survey to the next. Another is differential interviewer effect on the response. This paper addresses these potential sources of error, compares their relative impact, and focuses on whether a systematic (and hopefully correctable) component of reporting variability can be identified.

In this research, random intercept modeling was used to quantify the percentage of variability explained by respondent change and the differential effects of enumerator assignment. Under certain assumptions, the modeling provides estimates of standard survey quality measures, such as reliability and indices of inconsistency. When enumerator assignment is incorporated in the model, the approach can also be used to identify individual assignments which contribute inordinately to the survey to survey variability; thereby providing a tool to target potential enumeration problems where additional concept training might be needed. In essence, the modeling supplies considerable information about the data collection process without additional data collection requirements.

The objectives of this study were 1) to identify systematic components in survey to survey variability that could indicate nonsampling error, 2) to quantify the potential nonsampling error underlying these components, and 3) to target areas where corrective measures may be taken to reduce the problem. The paper discusses the models used, results of applying the approach to recent data collected by the National Agricultural Statistics Service (NASS), and the potential for incorporating this survey quality monitoring tool as an integral part of the NASS survey management process.

## BACKGROUND

The National Agricultural Statistics Service (NASS) conducts an ongoing series of quarterly sample surveys from which it estimates inventory and production of various agricultural commodities. The quarterly surveys conducted in June, September, December and March have a multiple frame design consisting of both list and area frame samples. In June, sampled area segments averaging about one square mile in size are completely enumerated to record all agricultural activity within their boundaries. The enumeration units are individual land operating arrangements (tracts) within the sampled segments. Data from these tracts provide both full area frame estimates and the area component of multiple frame indications, which include area data only for area tracts with no chance of list selection. These are referred to as non-overlap (NOL) tracts.

The NOL tracts identified in June are subsampled to account for list incompleteness in multiple frame indications from the follow-on quarters of September, December and March, when no full area enumeration occurs. Since for many commodities NASS' list frames are fairly complete, there are relatively few NOL tracts. For example, the numbers of NOL tracts in June 1995 varied by State from about 10 to about 300. However, in spite of their small numbers, the NOL tracts account for a large amount of the variability in multiple frame indications. Consequently, they are sampled at very high rates in the follow-on quarters, resulting in repeated measures for most of them.

Various analyses of the NASS quarterly survey data over the past few years have indicated a substantial amount of quarter-to-quarter reporting variability for individual sample units in survey items that should be fairly stable within a survey year. "True value" reinterview surveys have been conducted by NASS to address similar reporting concerns in the past. However, these create additional respondent burden, which is especially troublesome with the already heavily burdened NOL operations. If it works, a preferable way to assess data quality and identify areas needing improvement would be to glean as much survey quality information as possible

from the regular survey contact. This type of approach was explored in this study.

## METHOD

In many studies of nonsampling errors, either an administrative source or a reinterview survey provides the "true values" by which the quality of survey data is evaluated. In this study neither is available. However, NASS invests very heavily in the quality of data collected in its June Agricultural Survey, and a case can be made that these data may be of better quality than data collected in subsequent quarters. In essence, the June data possess many of the attributes associated with "true value" reinterview data. They are collected through personal enumeration (thought to be the best mode of data collection) by the best trained and most agriculturally experienced enumerators available to NASS. By comparison, most data collected in the follow-on quarters are by telephone from one of NASS' State Statistical Offices (SSOs). Furthermore, the telephone enumerators used in the follow-on quarters are generally hired locally and often have less experience in agricultural surveys than the field enumerators used in June.

Obviously, both the June and follow-on quarter data contain errors. However, if the June data contain fewer errors and can be viewed as a reasonable proxy to the truth, then data quality measures can be estimated directly from the quarterly data. In particular, the random intercept regression approach used in this study yielded estimates of data reliability, intra-group correlation, and the effects of various levels of a grouping variable (e.g., follow-on quarter enumerator assignment or a respondent change indicator). Percentages of model variability (in predicting a follow-on quarter's response with the June response) that were attributable to differential group effects were calculated.

### The Model

Biemer and Atkinson (1995 (1 & 2)) discussed an approach by which measures of data quality and group effects for arbitrary grouping variables could be obtained from two-phase samples, where the second phase sample was selected for reinterview with reconciliation to obtain true values. This paper is an attempt to apply the approach to the situation where reinterview data are not available, but where independent repeated measures on sample units are available through on-going quarterly surveys. In the present case, one response (June's) is expected to be generally superior to the other and to represent a reasonable "proxy to the truth." The underlying model development is described in great detail in the previous references and will be discussed much less rigorously here. In general we fit a model of the form:

$$y_i = \gamma_0 + \gamma \mu_i + z_{gi} \qquad \text{(Eq. 1)}$$

where $\mu_i$ is the June value of the item, $\gamma_0$ and $\gamma$ are constants, and $z_{gi}$ is a random error term. Insofar as $\mu_i$ is the "true" value for the item of interest, the parameter $\gamma_0$ may be interpreted as a constant or absolute bias that is added to all observations, while $\gamma$ is a "proportional" bias. As an example, suppose $\mu_i$ is some measure of farm size (e.g., all land in farm or total acres of cropland). The magnitude of the error in $y_i$ is often proportional to size and is therefore appropriately modeled by $\gamma\mu_i$. The term $z_{gi}$ is the sum of two random components, $d_g$ and $\delta_i$, where $d_g$ is the "bias" or "group effect" associated with group $g$, and $\delta_i$ is an independent unit-level error. We assume that $d_g \sim (0, \sigma_d^2)$ and $\delta_i \sim (0, \sigma_\delta^2 \mu_i^\lambda)$ where $\lambda$ is a known constant. In this study a value of 0 was used for $\lambda$; however, it is possible to estimate $\lambda$ from the data (see, for example, Wright, 1983).

With the above model, further assume the conditional covariance of the errors for $i \in G_g$ is given by

$$\begin{aligned} \text{Cov}\,(z_{gi}, z_{g'i'} \mid i) &= \sigma_d^2 + \sigma_\delta^2 \mu_i^\lambda \quad \text{for } i = i' \\ &= \sigma_d^2 \quad \text{for } i \neq i';\ g' = g \\ &= 0 \quad \text{for } g' \neq g \end{aligned}$$

Let $E(\bullet \mid i)$ denote the conditional expectation given the unit $i$ over the measurement error distribution and $\text{Var}(\bullet)$ denote the unconditional variance with respect to the sampling distribution. If we assume that all the $G_g$, $g = 1,\ldots,J$ are of equal size (say $m$) and that the finite population correction is ignorable, then Biemer and Stokes (1991) show that

$$\text{Var}(\hat{Y}) = N^2 n^{-1} \gamma^2 \sigma_\mu^2 [1 + (m-1)\rho_y] / R \qquad \text{(Eq. 2)}$$

where $R$, referred to as the reliability ratio, is

$$R = \frac{\text{VarE}(y_i \mid i)}{\text{Var}(y_i)} = \frac{\gamma^2 \sigma_\mu^2}{\sigma_y^2}$$

and (when a grouping variable of enumerator assignment is used) $\rho_y$, referred to as the *intra-enumerator correlation*

*coefficient*, is the correlation between pairs of units within an enumerator's assignment.

The reliability ratio, $R$, is the ratio of the variance of the "true" value for the data item -- *viz.*, $\mathrm{Var}(\gamma_0+\gamma\mu_i)$ -- to the variance of the observation $y_i$. Estimation of $R$ usually requires repeated measurements obtained under identical survey conditions and such that the measurement errors associated with each measurement are independent and identically distributed (see Biemer and Stokes, 1991). While these assumptions are perhaps best satisfied with a well-designed and executed reinterview survey, these are costly and create additional respondent burden. The focus of this study was to take an alternate approach and estimate $R$ directly from the quarterly survey data.

Also, under the current model (Eq. 1) with a grouping variable of follow-on survey enumerator assignment, the intra-enumerator correlation coefficient, $\rho_y$, is given by

$$\rho_y = \frac{\sigma_d^2}{\sigma_y^2}.$$

This statistic is widely used in measurement error studies to describe the degree to which the quality of interviewing varies by enumerator (see for example, Groves, 1989). A large estimate of $\rho_y$ indicates that large enumerator effects ($d_g$) are present in the data, and an analysis of the large absolute values of $d_g$ can help identify which enumerator assignments are contributing the most to the enumerator variance. This paper presents estimates of R and $\rho_y$, as well as a distribution of the standardized $d_g$ associated with the enumerators for the 1995-96 and 1996-97 follow-on surveys.

**The Data Analyzed**

To explore the usefulness of this approach in studying quarter-to-quarter reporting variability, survey data sets covering June, September, December, and March were constructed for the 1995-96 and 1996-97 survey years. For the purposes of this study it was necessary to eliminate as thoroughly as possible "real" quarter-to-quarter inventory changes, since the success of this approach for monitoring data quality is predicated on the assumption that data differences between June and the follow-on quarters are indicative of measurement error. Only insofar as this is true, can the approach be used to help identify and quantify sources of error.

Two primary steps were taken to minimize the confounding effects of real change. First, the items analyzed were limited to all land in farm and cropland, items less likely than others to legitimately change during the course of the survey year. Secondly, records for which an operation change was indicated were eliminated.

Separate analysis data sets were created for all land in farm and cropland, based on usability for these items. Records in the data sets represented a usable follow-on quarter's response with an associated usable June response. For a record to be included, the analysis item had to have been reported (not estimated or imputed) in both June and the follow-on quarter. The resulting four-quarter data sets with all States' data contained about 10,000 records for each year and analysis item.

**The Model Selection Process**

Using the Eq. 1 model specification, separate models were generated with two basic grouping variables -- follow-on quarter enumerator assignment and an indicator variable for whether or not there was a respondent change between June and the follow-on survey. A refinement to the basic model to adjust for the differential size of operations in the various levels of the grouping variable (i.e., average size of operation in an enumerator's assignment) was considered. However, this refinement didn't improve the modeling. The variable used to capture this information -- average June acreage (all land in farm or cropland) for a group -- was virtually always both statistically and practically insignificant. The overall proportional adjustment, $\gamma$, appeared to be sufficient to account for differences in the average size of operations in each group, regardless of whether the grouping variable used was enumerator assignment or respondent change indicator.

Another modeling possibility, including random effects for both grouping variables (and their interaction) in the same model, was explored. This approach did not work adequately, however, since crossing enumerator assignment with respondent change indicator resulted in too many empty or sparse cells. In general, there are relatively few respondent changes from June to a follow-on quarter. In the analyzed data sets this occurred only about 25 percent of the time. As a result, crossing this variable with another one was not a viable modeling alternative.

Therefore, our selected models were marginal ones with the two grouping variables incorporated separately. To eliminate any confounding effect on the models, all records with a respondent change were excluded when models using enumerator assignment as the grouping

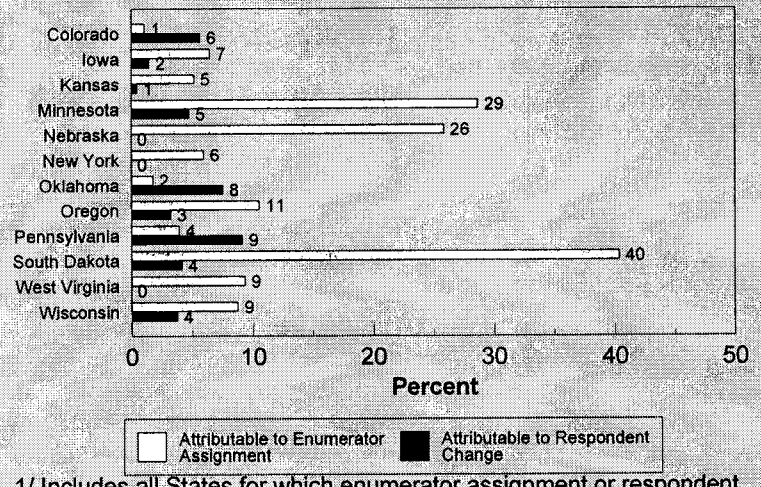
variable were fit.

Finally, to better reflect State-to-State differences in the modeling and to restrict outlier problems to individual States, the models were created at the State level.

## Model Results

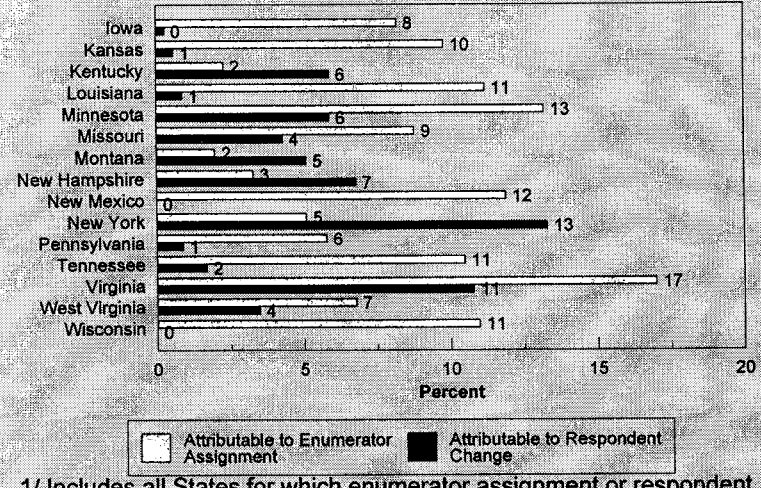
Somewhat surprisingly, although again resulting largely from the rarity of the event, respondent change generally had a small effect. In most States, enumerator assignment effect was larger, and in some States it was substantial. Figures 1 and 2 compare the percentage of model variability attributable to enumerator assignment vs. respondent change for all land in farm and cropland, respectively. Some caution is needed in interpreting the percentages in Figures 1 and 2, since in a few cases (most notably Minnesota and Nebraska for all land in farm) the large percentages of variability accounted for were on very small bases. Some States showed substantial overall model variability while other States showed very little.

**Figure 1: Estimated Impact of Enumerator Assignment and Respondent Change on Quarter to Quarter Variability for the 1995 Crop Year – All Land 1/**




1/ Includes all States for which enumerator assignment or respondent change accounted for at least 5 percent of the indicated variability.

**Figure 2: Estimated Impact of Enumerator Assignment and Respondent Change on Quarter to Quarter Variability for the 1995 Crop Year – Cropland 1/**




1/ Includes all States for which enumerator assignment or respondent change accounted for at least 5 percent of the indicated variability.

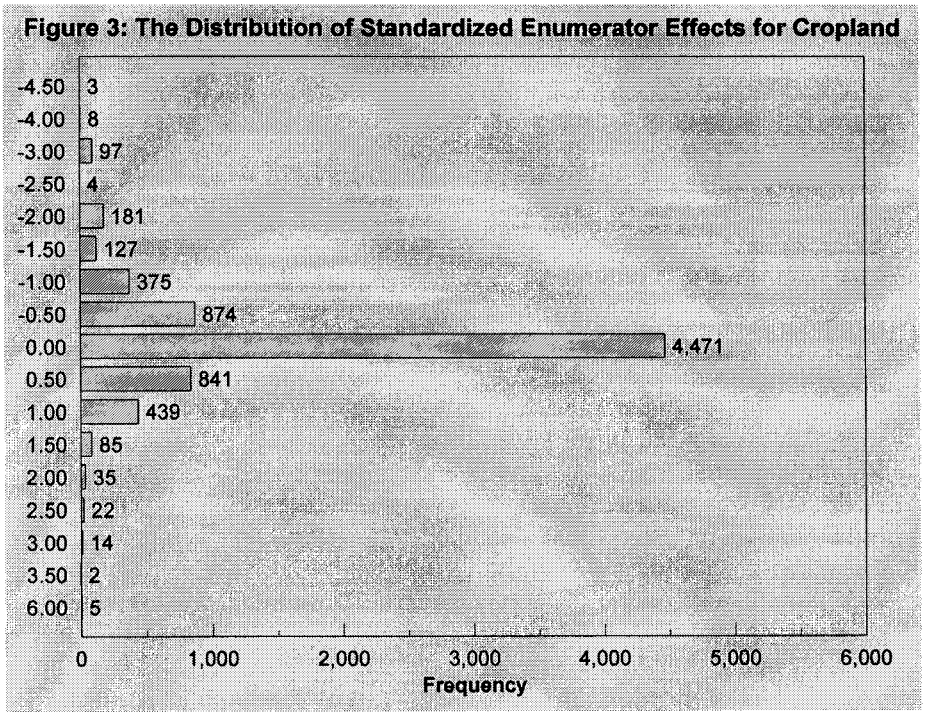



Figure 3 shows a distribution of estimated enumerator effects for cropland, standardized to reflect within-State variability. The tails of the distribution indicate enumerator assignment effects that are abnormally large. A review of these can help identify groups where nonsampling error problems may exist.

Table 1 shows our estimates of data quality at the U.S. level for all land in farm and cropland for both the 1995 and 1996 survey years. The tabled results indicate that both acreage items improved in reliability between 1995 and 1996. The estimate of intra-enumerator correlation for cropland also decreased noticeably between the two years, indicating data quality improvement.

Estimates of reliability and intra-enumerator correlation at even the 1996 level, however, indicate the potential for substantial variance inflation. Based on the average enumerator assignment size in the quarterly surveys analyzed in this study, an intra-enumerator correlation of .028 could cause an increase in variance in an estimate of about {[1+(6-1)(.028)]/.71}-1=.606, or 60.6% (using Eq. 2).

Estimates of the intra-enumerator correlation varied widely by State. In some States a large variance inflation could be expected, while in others the problem is considerably smaller. Individual State outliers played a substantial role in the calculated estimates at the State level, but a somewhat lesser role at the U.S. level. U.S. estimates were computed as a weighted average of State estimates, using total acreage of that type (cropland or all land) in the State as the weight. The standard errors of the estimates, as shown in Table 1, were calculated through bootstrapping.

In general, State estimates of both reliability and intra-enumerator correlation looked reasonable relative to the State results that had earlier been obtained from NASS' reinterview program from 1987-90.

**Table 1: Estimates of Rho and R**

| Item | Survey Year | $\rho_y$ (s.e.) | $\hat{R}$ (s.e.) | Group Size (m) |
|---|---|---|---|---|
| All Land in Farm | 1995 | .026 (.013) | .68 (.028) | 7 |
| | 1996 | .029 (.019) | .80 (.023) | 6 |
| Cropland Acres | 1995 | .041 (.018) | .68 (.024) | 7 |
| | 1996 | .028 (.018) | .71 (.024) | 6 |

## CONCLUDING REMARKS

This paper documents an attempt to mine existing data to satisfy three basic objectives -- to obtain an indication of the relative quality of items collected in an ongoing survey program, to assess the relative impact of two potential sources of nonsampling error and to provide a tool to help target areas where additional enumerator training may be needed.

Like most situations where data are put to use in a way for which they're not specifically designed, the validity of some of the underlying assumptions may be questionable. In particular, the assumption that the June value represents truth, which provides the underpinning of the interpretation of our calculated statistics as estimates of reliability and intra-enumerator correlation, can be debated. Also, the assumption of equal group sizes does not strictly hold in NASS' survey workload assignments.

While violations of these assumptions caused problems in a few specific instances, the process generally provided very useful information. Because of unequal size groups and a confounding of errors, the enumerator assignments identified as problematic were not always indicative of poor enumeration at all, but sometimes a combination of a small assignment size and a serious key-entry error. Whatever the reason for their identification, however, the groups with large, absolute group effects were generally ones that should be examined for some type of nonsampling error. Also, estimates of survey quality from this approach were comparable to those previously produced at much higher cost through earlier reinterview surveys.

Finally, the modeling used in this research would be simple to implement as a standard survey analysis tool. Since the procedure was implemented through SAS' PROC MIXED in this study, there was no special programming code necessary to run the models. The code as written produced review listings of all outlier and leverage point samples, and all samples in enumerator assignments whose effects were statistically significant. The procedure appears to perform well in identifying problematic groups of data, indicating that it may indeed have potential for operational use.

Finally, from an operational point of view the most challenging part of the whole study was to properly link the survey responses from several different files, into one combined file. However, with the advent of a data warehouse in NASS, the data access and reformatting activities required to support the modeling effort should be greatly simplified for future applications of this approach.

## REFERENCES


Biemer, P. and Atkinson, D. (1995). "An Integrated Approach for Estimating Measurement Error Bias and Variance in Two-Phase Samples," *Proceedings of the 1995 Annual Research Conference*, pp. 355-357.

Biemer, P. and Atkinson, D. (1995). "Estimating Measurement Error Bias and Variance in Two-Phase Samples," *American Statistical Association - Proceedings of Section on Survey Research Methods, Volume II*, pp. 775-780.

Biemer, P. and S.L. Stokes (1991). "Approaches to the Modeling of Measurement Errors." In P.P. Biemer, R.M. Groves, L.E Lyberg, N.A. Mathiowetz, S. Sudman (Eds.) *Measurement Errors in Surveys.* New York: John Wiley & Sons, pp. 487-516.

Groves, R.M. (1989). *Survey Errors and Survey Costs*, John Wiley & Sons, N.Y.

Wright, R.L. (1983). "Finite Population Sampling with Multivariate Auxiliary Information," *Journal of the American Statistical Association,* 78, pp. 879-884.