A SCREENING PROCEDURE FOR IMPROVING LARGE AREA CROP ACREAGE ESTIMATES

R. S. Chhikara, Lockheed Electronics Company, Inc.[1]

ABSTRACT

When crop acreage estimates for large areas are made, several factors may affect their accuracy. If one or a few small-size area segments are sampled from a stratum and if the crop acreage determination in a sample segment is subject to measurement error, the stratum crop acreage estimate may deviate considerably from the actual crop acreage. This paper presents a screening procedure that evaluates the stratum crop acreage estimates and detects the strata for which significant deviations from the expected acreages in the strata are observed. A significance test based on the extreme studentized deviate statistics is used to detect potential multiple outliers. The procedure is applied to evaluate the stratum crop acreage estimates obtained for a crop survey conducted by the National Aeronautics and Space Administration using satellite data.

1. INTRODUCTION

In agricultural sample surveys, it is fairly common to experience nonsampling errors because of nonresponse and measurement errors. The measurement errors are often due to a fallible measuring method, incorrect reporting, or mistakes in recording the information.

When the survey data are subject to both measurement and sampling errors, the stratum crop acreage estimates may become quite unreliable. Therefore, it is desirable to screen the data before they are utilized in making stratum estimates and then, in turn, in obtaining a large area crop acreage estimate. Stratum estimates afflicted with relatively higher measurement and sampling errors can be given lesser weights than others in obtaining the estimate for the entire area of interest.

Most often, survey data screening has been used when the observation for a sample unit is partially or completely lacking or when the reporting of data is unreliable and can be verified from other sources of information [Hocking et al. (1974), Pregibon (1977), Freund and Hartley (1967)]. In these studies, among others, screening has been at the level of the small-size area segment used as the sampling unit. However, when the measurement error for the sample segment is random and the within-stratum variance cannot be reliably estimated, a valid evaluation at the segment level may not be feasible.

Presently, we consider the evaluation of stratum crop acreage estimates when the number of strata is large and the crop acreage in a sample segment is estimated and therefore subject to measurement error. The information on crop acreage in the past is generally available at the stratum level (when the stratum is fairly large), and a significant correlation between the crop acreage during the current year and the crop acreage in a previous year can be expected at this level. Based on these assumptions, a statistical procedure has been developed to evaluate the stratum crop acreage estimates against their historical crop acreages and to detect the strata for which significant deviations from the expected ratio of the stratum acreage estimate to the historical acreage across the strata are observed. The procedure is described in the next section. It is applied to screen the stratum wheat acreage estimates obtained in a crop survey conducted by the National Aeronautics and Space Administration (NASA) using satellite data.

2. PROCEDURE

2.1 Approach

Let L be the number of strata and n_i be the number of (fixed-size) area segments randomly sampled from the *ith* stratum, i = 1, 2, ···, L. For stratum i, let $\overline{Y}_i$ be the average crop acreage and $\overline{X}_i$ be the corresponding value in a previous year.[2] Suppose y_{ij} is the actual crop acreage and $\hat{y}_{ij}$ is its estimate for the *jth* sample segment of stratum i. The sample mean

$$\overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \qquad (2.1)$$

is an unbiased estimate of $\overline{Y}_i$, and its precision is influenced by the rate of sampling alone; whereas

$$\hat{\overline{y}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \hat{y}_{ij} \qquad (2.2)$$

as an estimate of $\overline{Y}_i$ is subject to both sampling and measurement errors. In practice, the y_{ij} are unknown, and thus $\overline{y}_i$ cannot be computed.

Consider the ratio

$$\hat{z}_i = \frac{\hat{\overline{y}}_i}{\overline{X}_i} \quad , \quad i = 1,2,\cdots,L \qquad (2.3)$$

Then, $\hat{z}_i$ can be written as

$$\hat{z}_i = Z_i + (z_i - Z_i) + (\hat{z}_i - z_i) \qquad (2.4)$$

where $Z_i = \frac{\overline{Y}_i}{\overline{X}_i}$ and $z_i = \frac{\overline{y}_i}{\overline{X}_i}$

The first error component, $\varepsilon_i = (z_i - Z_i)$, is due to sampling and the second error component, $\delta_i = (\hat{z}_i - z_i)$, is due to measurement error. Considering that the measurement error may cause bias in the stratum estimate, let $E(\delta_i | \overline{X}_i) = B_i$.

It will be assumed that ε_i and δ_i are uncorrelated; and, given $\overline{X}_i$, these errors do not depend upon Z_i. Accordingly, the conditional mean and variance of $\hat{z}_i$ are

$$\left.\begin{aligned} E(\hat{z}_i|\overline{X}_i) &= Z_i + B_i \\ \text{and} \qquad \operatorname{Var}(\hat{z}_i|\overline{X}_i) &= \sigma_i^2 + \sigma_{0i}^2 \end{aligned}\right\} \qquad (2.5)$$

where $\sigma_i^2 = \operatorname{Var}(\varepsilon_i|\overline{X}_i)$ and $\sigma_{0i}^2 = \operatorname{Var}(\delta_i|\overline{X}_i)$,

$$i = 1,2,\cdots,L$$

When the stratification is based primarily on the crop size in the past and if all strata that had about the same crop size in the past are homogeneous and have quite similar cropping practices, the bias and variance in equation (2.5) for the group of strata are likely to be the same. Therefore, it is possible to divide the strata into a set of groups and assume equality for the stratum variances and biases within each group of strata. Accordingly,

$$\left.\begin{aligned} \sigma_i^2 &= \sigma_s^2 \\ \sigma_{0i}^2 &= \sigma_c^2 \\ B_i &= B \end{aligned}\right\} \qquad (2.6)$$

for all strata in a group.

Assume that $\overline{Y}_i$ is proportional to $\overline{X}_i$, so that $Z_i = Z$ for all i and $\operatorname{Var}(Z_i) = 0$. Thus, the unconditional mean and variance are

$$\left.\begin{aligned} E(\hat{z}_i) &= Z + B \\ \text{and} \qquad \operatorname{Var}(\hat{z}_i) &= \sigma_s^2 + \sigma_c^2 \end{aligned}\right\} \qquad (2.7)$$

for the group of strata. It should be noted that the mean and variance in equation (2.7) are still conditional on the year of $\overline{X}_i$. The year-to-year variance component of $\hat{z}_i$ is assumed to be much smaller than $\sigma_s^2 + \sigma_c^2$ and, hence, is ignored.

If the assumption of $\overline{Y}_i$ being proportional to $\overline{X}_i$ does not hold and Z_i varies across strata, the mean and variance of $\hat{z}_i$ are $(\overline{Z} + B)$ and $\sigma_0^2 + \lambda\left(\sigma_s^2 + \sigma_c^2\right)$ in place of those given in equation (2.7), where $0 < \lambda < 1$ and $\overline{Z}$ and σ_0^2 are the mean and variance of Z_i for strata in a group. However, $\lambda\left(\sigma_s^2 + \sigma_c^2\right)$ can be expected to dominate σ_0^2 whenever the groups are judiciously chosen, with $\overline{X}_i$ varying as little as possible within a group.

The $\hat{z}_i$ for a group are examined for any significant deviations from their mean. It will be assumed that, for each group, the variable $\hat{z}_i$ is normally distributed, with mean and variance of the form discussed above. The parameters, of course, are unknown and will be estimated from the observed data. A significance test based on the extreme studentized deviate (ESD) statistics and discussed below is applied to detect multiple outlying observations (outliers) in the data. The strata for which the observed values of $\hat{z}_i$ are declared outliers are flagged as having unreliable crop acreage estimates.

2.2 Screening of Stratum Estimates

Consider a group of h strata with observations $\hat{z}_1, \hat{z}_2, \cdots, \hat{z}_h$ for the variable $\hat{z}_i$ as defined in equation (2.3) for the screening of stratum estimates. There is no fixed rule to decide the possible number of outliers in the data. However, a certain percentage of the data can be considered for potential outliers. The fixed percentage rule seems impractical, inasmuch as it will lead to testing for too many outliers when the number of observations is large. A more suitable rule may be to consider $\sqrt{h}$ (to the nearest integer) for the number of possible outliers, as it will result in testing for a reasonable number of outliers; e.g., not more than 10 for 100 data points.

Consider the sequence of data sets $A_0, A_1, \cdots, A_k$, where A_0 is the full set of data and the set A_i is formed by deleting from A_{i-1} the data point farthest away from the mean of A_{i-1}, $i = 1, 2, \cdots, k$ (k denotes the number of potential outliers). Then, consider the absolute value of the maximum studentized deviate for test statistics in each set. If one of these test statistics exceeds its critical value, the data are declared to have outliers. The outliers are those data points excluded from the set following the last set for which the test statistics exceed their critical values. For details of the multiple outlier detection test, see Rosner (1975). However, see Chhikara and Feiveson (1980) if the number of outliers to be tested is between 3 and 10, since Rosner gives critical values of the ESD test statistics only for the cases of k = 1 and k = 2. These critical values are for the 5-percent level of significance.

2.3 Improving the Acreage Estimation

The strata flagged as having outlying observations need to be treated differently from other strata in the group. Although it is desirable to make use of the full sample data and thereby to utilize the estimates of the flagged strata, it is equally important to improve upon their estimates as well as those for the entire area of interest.

One approach to prevent "bad" estimates for strata is to discard their sample estimates and, instead, to obtain their estimates using a ratio estimation technique. A ratio estimate for a flagged stratum, derived by multiplying the acreage estimate of all the unflagged strata in the group by the ratio of the historical acreage for the flagged stratum to that for the unflagged strata, will be an improvement over the direct estimate obtained from the sample data, provided that $\overline{X}_i$ is proportional to $\overline{Y}_i$. If the grouping of strata is done judiciously, so that $\overline{X}_i$ is approximately proportional to $\overline{Y}_i$, a better set of estimates of flagged strata will be obtained by replacing their direct estimates by the ratio estimates.

The screening of all strata estimates requires the above procedure to be repeated for each group of strata. When the improved estimates for all the flagged strata are used in computing the stratified estimate for a large area, an improved crop acreage estimate is obtained for the entire area of interest.

3. AN APPLICATION

To illustrate this technique, a real survey data application is described here. A large area crop survey experiment was conducted by NASA in cooperation with the U.S. Department of Agriculture (USDA) and the National Oceanic and Atmospheric Administration (NOAA) for estimating the 1977 wheat production in the U.S. Great Plains[3] using satellite data. The sampling unit was a 5- by 6-nautical-mile area segment. Sample segments were selected using a stratified random sampling technique. The counties were considered as strata. None to a few segments were allocated to a county. The wheat acreage determination for a sample segment was afflicted with a measurement error resulting from a fallible classification method. Under this method, an image analyst labeled the spectral classes, and a statistical discriminant analysis of the spectral data was performed to classify each data point as wheat or nonwheat and thereby to estimate the proportion of wheat acreage in the segment. For details, see Chhikara and Feiveson (1978).

Large sampling and measurement errors were experienced in strata where wheat cultivation was sparse. These phenomena occurred because, at most, one sample segment was allocated to a county with low wheat density and because of the higher uncertainty involved in classifying data for a segment with a low wheat acreage proportion.

For the winter wheat region, the observed data for $\hat{z}_{ij}$ are given in Figure 1. ($\hat{z}_{ij}$ is the ratio of the segment wheat acreage estimate to $\overline{X}_i$ for the county in which the segment lies, where the 1974 county wheat acreages from the Agricultural Census reports were used in determining $\overline{X}_i$.) Clearly, when the proportion of wheat in a county, P_i, is very small, $\hat{z}_i$ is highly variable[4] and the stratum acreage estimate $\hat{\overline{y}}_i$ may be quite unreliable and inaccurate.[5] Figure 1 also indicates that the distribution of $\hat{z}_i$ depends upon P_i. However, if counties are grouped on the basis of P_i, such dependence can be eliminated for the strata within a group.

Counties were divided into four groups: low, marginal, medium, and high wheat density, as given, respectively, by the following.

$$G_1 = \{i:\ 0 < P_i < 0.05\}$$

$$G_2 = \{i:\ 0.05 < P_i \leq 0.15\}$$

$$G_3 = \{i:\ 0.15 < P_i \leq 0.30\}$$

$$G_4 = \{i:\ 0.30 < P_i \leq 1\ \}$$

Table 1 lists the number of counties in each group.

The distribution of $\hat{z}_i$ for each group was skewed. A plot of data using a logarithmic scale made the distribution fairly symmetrical. Thus, the logarithmic transformation was applied to the observed data of $\hat{z}_i$ to obtain the normal approximation for the primary distribution. The transformed data (that is, the logarithm of $\hat{z}_i$ from each group) were evaluated using the screening procedure given in section 2.2. The counties for which observations were detected as outliers were flagged. The number of counties flagged in each group is also listed in Table 1.

The group G_1 contained 33 counties, of which 4 counties were flagged to have unreliable acreage estimates. Table 2 lists the details showing how the outlier test procedure was carried out. The fact that the computed test statistic T_i exceeds its critical value λ_i for the last time for set A_3 indicates the presence of four outliers.

The wheat acreage estimates for the flagged counties were obtained by the ratio estimation technique discussed in section 2.3. When the large area estimates obtained using these ratio estimates and those obtained using the corresponding direct estimates were compared with the USDA wheat acreage estimates for the year, the use of ratio estimates led to a slightly better wheat acreage estimate for the U.S. Great Plains.

REFERENCES

Chhikara, R. S., and Feiveson, A. H. (1978), "Landsat-Based Large Area Crop Acreage Estimation — An Experimental Study," *Amer. Statistical Assn. 1978 Proceedings of the Section on Survey Research Methods*, pp. 153-159.

________ (1980), "Extended Critical Values of Extreme Studentized Deviate Test Statistics for Detecting Multiple Outliers, *Communications in Statistics*, B9, no. 2, to appear.

Freund, R. J., and Hartley, H. O. (1967), "A Procedure for Automatic Data Editing," *J. of Amer. Statistical Assn.*, 62, 341-352.

Hocking, R. R., Huddleston, H. F., and Hunt, H. H. (1974), "A Procedure for Editing Survey Data," *J. of the Royal Statistical Soc.*, Series C (Applied Statistics), 23, 121-133.

Pregibon, D. (1977), "Typical Survey Data: Estimation and Imputation," paper presented at the Conference on the Analysis of Large Data Sets, sponsored by the Inst. of Math. Statistics, Dallas, Texas.

Rosner, B. (1975), "On the Detection of Many Outliers," *Technometrics*, 17, 221-227.

FOOTNOTES

[1] Under Contract NAS 9-15800 to the National Aeronautics and Space Administration, Lyndon B. Johnson Space Center, Houston, Texas 77058.

[2] Instead of historical crop acreage, if another auxiliary variable that is significantly correlated with the crop acreage is considered, the approach and the subsequent screening procedure will still be applicable. Preferably, consider the auxiliary variable that is used in stratifying the area.

[3] Although estimates were made for the 1975, 1976, and 1977 crop years, only 1977 is considered here.

[4] When $n_i = 1$, $\hat{z}_i = \hat{z}_{ij}$.

[5] Since $P_i = \frac{\overline{X}_i}{\text{segment size}}$, and the segment size is fixed, either P_i or $\overline{X}_i$ can be used in the discussion.

TABLE 1. DISTRIBUTION OF COUNTIES BY GROUPS FOR THE WINTER WHEAT REGION

Group	Number of counties	
	Total	Flagged as outliers
G_1	33	4
G_2	80	2
G_3	82	7
G_4	44	1
Total	239	14

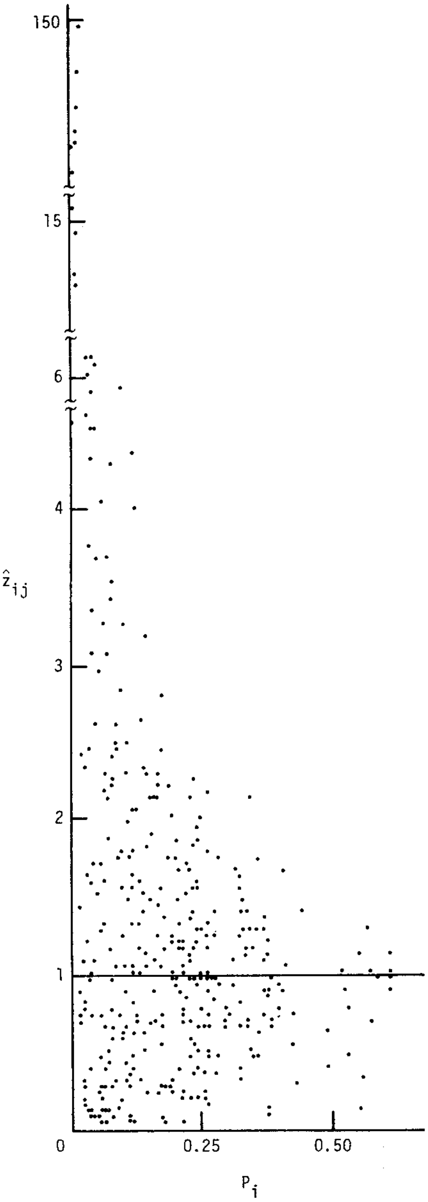

FIGURE 1. PLOT OF $\hat{z}_{ij}$ VERSUS P_i.

TABLE 2. DETECTION OF OUTLIERS IN THE DATA SET FROM GROUP G_1

State code	County code	Wheat acreage in 1974, %	Estimated wheat acreage, %	Log $\hat{z}_i$				T_i	λ_i
				Data set	Mean	Standard deviation	Extreme outlying observation		
31	123	4.69	0.00	A_0	-0.298	1.521	-3.848	2.333	3.33
46	71	4.50	0.00	A_1	-0.187	1.404	-3.806	2.578	2.84
31	157	3.92	0.00	A_2	-0.071	1.259	-3.670	2.859	2.67
31	175	2.94	0.00	A_3	0.049	1.085	-3.382	3.161	2.57
46	53	0.76	0.00	A_4	0.168	0.886	-2.025	2.475	2.51
31	141	3.11	0.58	A_5	0.246	0.794	-1.600	2.426	2.45