Sample Surveys that Use Imagery with Varying Area Coverage

Leigh Harrington, Mark Rivard, William Zink, and *Nancy Cobos*
Environmental Research Institute of Michigan, 1501 Wilson Boulevard, Suite 1105, Arlington, VA 22209

ABSTRACT: Aerial sample surveys may require use of image frames that have varying ground coverage. This may be due to variation in aircraft altitude, deleted land categories (e.g., water), cloud coverage/shadow, and other reasons. A weighted estimate of the population totals was obtained that minimized that variance where the weights were a function of the ground area and the spatial correlation of the items of interest. This estimator included as special cases the two standard estimators that extrapolate sample total and sample density.

INTRODUCTION

WHEN THE AREA OF INTEREST is large and the required scale is large, an aerial survey that images the entire region is often cost prohibitive. Rather, a sample survey is required. If it is part of a well-designed experiment, this sample can then be statistically extrapolated to the entire region.

This paper addresses the problems associated with survey data when images/frames have varying area coverage. That is, the total area imaged at one location differs from that acquired at other locations. This can happen for a number of reasons. For example, if the survey is collected over mountainous terrain at a relatively low altitude, the pilot will have trouble maintaining a constant altitude above ground as the plane flies over valleys and mountain tops. This will cause image frames to have different scales, and the area imaged on the ground will vary with the square of the altitude above the ground.

Even if the image scale is kept constant (as often specified in the survey design), area can vary for other reasons. For example, some part of the frame may be cloud covered or in cloud shadow and must be debited out. Such frames will represent smaller ground area compared to those without cloud cover. Sometimes a particular land-cover category (e.g., water) is of no interest and, if included, might give misleading results. Such adjustments will again induce variation in plot size.

Ground area may also vary because of the survey design itself. Typically an aerial survey stratifies the entire region of interest into two or more homogeneous subareas, for example, high and low density areas. Sometimes, because of navigation errors, or because the ground area associated with a frame is very large, a single image will intersect two or more strata, and the frame must be divided accordingly. As a result, some frames will have smaller (possibly much smaller) area then others.

The effect of varying ground area is to make the statistical analysis more complicated. Standard sample survey methodologies for treating varying sample unit size – such as sampling with probability proportional to size – do not apply here because they require that the ground area associated with each frame or potential frame be known in advance. This is simply not the case for many aerial surveys.

This paper presents alternative approaches for analysis of aerial sample surveys that have varying ground area per frame. In particular, a common way to reduce the survey variance is to standardize the data to a per unit area basis, i.e., densities. A more complex correction was developed that calculates a weight for each frame. These weights are based upon the ground area imaged and spatial correlation of the objects of interest, so that the variance of the survey estimate is minimized.

METHODOLOGY

In order to make the discussion more focused, Table 1 presents a hypothetical aerial survey designed to estimate the total number of agricultural fields in a region of interest. This region had a total area, Q, equal to 179,000 hectares (ha). The value of Q is typically determined from a map. As Table 1 indicates, image frames were acquired at ten random locations within this region. An image analyst then identified the number of agricultural fields, C_i, in each frame and determined the total area exploited, q_i, based upon aircraft elevation data, cloud cover, etc. Note that q_i varied from a minimum of 100 ha (Frame Number 1) to a maximum of 800 ha (Frame Number 7). The average area per frame was 358 ha and the total area exploited was 3580 ha (i.e., a 2 percent sample). In particular, it would require a total of $N = 500$ frames of average size to image the entire region of interest. A total of 150 fields were detected in the sample, as shown in Table 1.

AN ESTIMATE BASED ON FRAME COUNTS

One way to estimate the total number of fields in the region is to simply extrapolate the number of fields in Table 1 to the entire area. Because the sample of ten frames represented 2 percent of the total area, an unbiased estimate of the total number

TABLE 1. HYPOTHETICAL AERIAL SAMPLE SURVEY TO ESTIMATE THE TOTAL NUMBER OF AGRICULTURE FIELDS IN A 179,000 HA REGION

Frame	q_i Area (Hectares)	C_i Field Count	d Density
1	100	4	0.040
2	500	14	0.028
3	250	20	0.080
4	300	6	0.020
5	325	0	0.000
6	330	0	0.000
7	800	74	0.093
8	175	17	0.097
9	425	0	0.000
10	375	15	0.040
Total	$\Sigma q_i = 3580$	$\Sigma C_i = 150$	
Average	$\bar{q} = 358.0$	$\overline{C} = 15.0$	$\bar{d} = 0.0397$
Standard Deviation (SD)	193.6	22.08	0.038

PHOTOGRAMMETRIC ENGINEERING AND REMOTE SENSING,
Vol. 56, No. 2, February 1990, pp. 181–185.

0099-1112/90/5602–181$02.25/0

of fields would be

$$\begin{aligned} Y_1 &= \sum_{i=1}^{n} C_i/f \\ &= 150/0.02 \\ &= 7500 \text{ fields} \end{aligned} \tag{1}$$

where f is the sampling fraction (f = area sampled/total area = 0.02). The standard deviation of the estimate is discussed in a later section and was estimated as 2225. Thus, a 95 percent confidence bound (i.e., 2 standard deviations) would be 7500 ± 60 percent.

This sampling uncertainty would be unacceptably large for many applications. One way to reduce this uncertainty would be to acquire a larger sample. Alternatively, text books on sample survey methodology (e.g., Cochran, 1977) suggest using field densities as a way of reducing the standard deviation. Thus, this approach was also examined as discussed below.

An Estimate Based Upon Densities

Table 1 also shows the density of fields for each frame and also gives the average density, $\bar{d}$, and the standard deviation, where

$$\bar{d} = \sum_{i=1}^{n} d_i/n$$

and

$$SD^2 = \sum_{i=1}^{n} (d_i - \bar{d})^2/(n-1).$$

A second unbiased estimate of the total number of fields in the region of interest would be

$$\begin{aligned} Y_2 &= \bar{d} \times Q \\ &= 0.0397 \times 179{,}000 \\ &= 7115 \text{ fields.} \end{aligned} \tag{2}$$

The estimated variance of Y_2, as shown in Cochran (1977), is given by the equation

$$\begin{aligned} \hat{\text{V}}\text{ar}(Y_2) &= (1-f)\,(SD \times Q)^2/n \\ &= 4{,}534{,}186 \end{aligned} \tag{3}$$

where $SD = 0.038$, as given in Table 1. The total area, Q, was 179,000 ha and $n = 10$ as before. The standard deviation for Y_2 was 2129, and, hence, the estimated 95 percent confidence bounds would be 7115 fields ± 60 percent. Thus, for this case, the use of densities did not reduce the overall uncertainty relative to the total. However, if the population is homogeneously distributed, the use of densities often results in a more precise estimate (e.g., Cochran, 1977).

An Estimate Using Weights

The survey estimate can be further improved if additional information is known regarding the spatial correlation of the fields. Note that the density associated with Frame 1, Table 1, was 0.040 fields per ha. It was based upon a frame that imaged only 100 ha. The density for Frame 7, 0.093 fields per ha, was based upon an area of 800 ha. Of course, the density calculated using Frame 7 is likely to be more reliable than that associated with Frame 1. In fact, if each of the eight potential 100 ha subpatches in Frame 7 were statistically independent of one another, the variance of the density for Frame 7 would be eight times smaller than that for Frame 1. However, such independence is rare. It is more common for spatially distributed features, such as agricultural fields, to exhibit some degree of spatial correlation. This correlation will increase the variance of the density when compared to the independence case. As a result, the variance of a density based upon a frame with 800 ha will be less than one based upon a frame having only 100 ha, but not eight times less.

H. Fairfield Smith (1938) proposed that the variance of plot totals will vary as a power law with plot size. If q_i is the ground area associated with a frame i, then the Smith variance model would imply the variance of the corresponding field counts, C_i, would vary as

$$\text{Var}(C_i\,|q_i) = Bq_i^{b} \tag{4}$$

where $1 \le b \le 2$ and B is the variance of C_i when $q_i = 1$. When $b = 1$, the counts are spatially independent. This follows from the fact that the variance of the sum of independent random variables is just the sum of the individual variances.

If $b = 2$, the counts have perfect spatial correlation. This would imply that a frame with a small ground area was as useful as a frame with a large ground area. It follows that, given the small ground area count, one could predict the count of the surrounding ground area. In practice, one expects to find b values between 1 and 2. For example, the b parameter was estimated to be between 1.5 and 1.7 for wheat field yields in several midwestern states (Hallum and Perry, 1984). This model remained valid over several orders of magnitude in q. There is a large body of sample survey design literature on using this power law; see, for example, Cochran (1977, pp. 243-244). Hansen *et al.* (1953, pp. 306-309), and Jessen (1978, pp. 100-107).

Because density is calculated as counts divided by area (i.e., C/q), Model 4 implies that the variance of the density is

$$\text{Var}(d_i) = Bq_i^{b-2}. \tag{5}$$

Now assume that each frame was a random sample from a super population where $d_i = \mu_d + \epsilon_i$ and the ϵ_i were independent with zero mean and variance given by Equation 5. Further, consider the class of linear weighted estimators of μ_d of the form

$$\hat{\mu}_d = \sum_{i=1}^{n} w_i d_i. \tag{6}$$

It has been shown (Cochran, 1977, p. 160) that if the d_i are independent, and have a variance given by Equation 5, then the best linear unbiased estimate (BLUE) of μ_d has weights that sum to 1.0 which are given by

$$w_i = q_i^{2-b} \Big/ \sum_{j=1}^{n} q_j^{2-b}. \tag{7}$$

The variance of $\hat{\mu}_d$ is

$$\begin{aligned} \text{Var}(\hat{\mu}_d) &= \sum_{i=1}^{n} w_i^2 \,\text{Var}(d_i) \\ &= B \Big/ \sum_{i=1}^{n} q_i^{2-b} \end{aligned} \tag{8}$$

An estimate of B would be (Press, 1972, p. 199)

$$\hat{B} = \sum_{i=1}^{n} (d_i - \hat{\mu}_d)^2\, q_i^{2-b} \Big/ (n-1).$$

So an estimate of the variance of $\hat{\mu}_d$ would be

$$\hat{\text{Var}}(\hat{\mu}_d) = \sum_{i=1}^{n} w_i(d_i - \hat{\mu}_d)^2 \Big/ (n-1). \tag{9}$$

The assumption that the d_i are independent is not acutally correct. The spatial correlations implied by Model 4 persist over long distances. See Harrington (1988) for details concerning the correlations implied by the Smith Power law. However, if sample image frames were sufficiently distant from one another, they would be nearly independent. For this case, the variance estimated by Equation 9 will be approximately correct. The theory can be extended to include correlated d_i, but it requires substantially greater mathematical development, and so it will not be treated here.

Given an estimate of $\hat{\mu}_d$, the estimated total becomes

$$Y_3 = \hat{\mu}_d Q \qquad (10)$$

and the variance of Y_3 is estimated as

$$\hat{\mathrm{Var}}(Y_3) = Q^2 \, \hat{\mathrm{Var}}(\hat{\mu}_d) \qquad (11)$$

as calculated in Equation 9. When $b = 2$ (i.e., perfect correlation), the weights reduce to $w_i = 1/n$. For this case, $\mu_d = \overline{d}$ and Y_3 in Equation 10 reduces to Y_2 (Equation 2) discussed earlier. When $b = 1.0$ (i.e., independence), the weights become

$$w_i = q_i \Big/ \sum_{j=1}^{n} q_j$$

$$\hat{\mu}_d = \sum_{j=1}^{n} C_j \Big/ \sum_{j=1}^{n} q_j$$

Thus, Y_3 reduces to Y_1 (Equation 1). Using Equation 11, then, the associated variance of Y_1 was calculated to be 2225 as stated earlier. Hence, this weighted estimate includes both Y_1 and Y_2 when $b = 1$ and 2, respectively. If this correlation can be characterized (or estimated) by a value of b between 1 and 2, then the spatial correlation can be factored into the analysis explicitly.

Figure 1(a) presents the estimated total using Equation 10 and the data from Table 1, as a function of b. Note that the estimated total was greatest when $b = 1$ (7500 fields), and was minimum when $b = 1.8$ (7101 fields). Figure 1(b) presents the standard deviation as a function of b, Equation 11. It was monotonically decreasing from 2225 to 2129. The important point is that spatial correlation, if known, can be incorporated into the estimate to compensate for the varying ground area. If not known, then the researcher must assume a value of b (typically either 1.0 or 2.0) with unknown consequences.

For the example, the range of values exhibited by both the total estimated number of fields and the associated standard deviation, as shown in Figure 1, is relatively small. Thus, the weighted estimate provided only a modest correction for varying frame size. This need not be the case if the variation in density was larger than that assumed in Table 1. It is quite easy to generate examples where both the estimated total and the associated standard deviation vary much more, as b varies between 1 and 2. To demonstrate this phenomenon, the data in Table 1 were adjusted so that the number of fields in Frame 1 was increased from 4 to 28, while the number of fields in Frame 7 was decreased a corresponding amount from 74 to 50. These two adjustments preserve the total number of fields detected, and the estimate Y_1 (7500 fields) did not change. However, as b increases from 1 to 2, frames with smaller area are given weights that are more nearly equal to those given the largest frames. As a result, the large density newly associated with Frame 1 (0.28 = 28/100) pulled up the estimated total as shown in Figure 2(a). When $b = 2$, the estimated total (Y_2) increased to 10,874 fields, a 45 percent increase. The standard deviation, Figure 2(b), increased by nearly 60 percent over the same interval. This example showed the importance of the spatial correlation effect when there is considerable variation in the frame size or cell-to-cell density.

An Alternative Weighted Estimate

Under the super population model, the total number of fields in the region of interest was a random variable, with a mean value of $\mu_d Q$. The sample survey was based upon a particular realization of this random variable. In many circumstances, however, $\mu_d Q$ is of less interest than the particular realization Y where

$$Y = \sum_{i=1}^{N} d_i q_i$$

$$= \mu_d Q + \sum_{i=1}^{N} q_i \epsilon_i$$

Because the d_i's would be known for the sample, Y can be optimally estimated as

$$Y_4 = (Q - n\bar{q}) \, \hat{\mu}_d + \sum_{i=1}^{n} d_i q_i$$

where $\hat{\mu}_d$ is calculated as in Equation 6 and 7. The first term estimates only those image frames missing from the sample. This term is then added to the observed counts. The variance

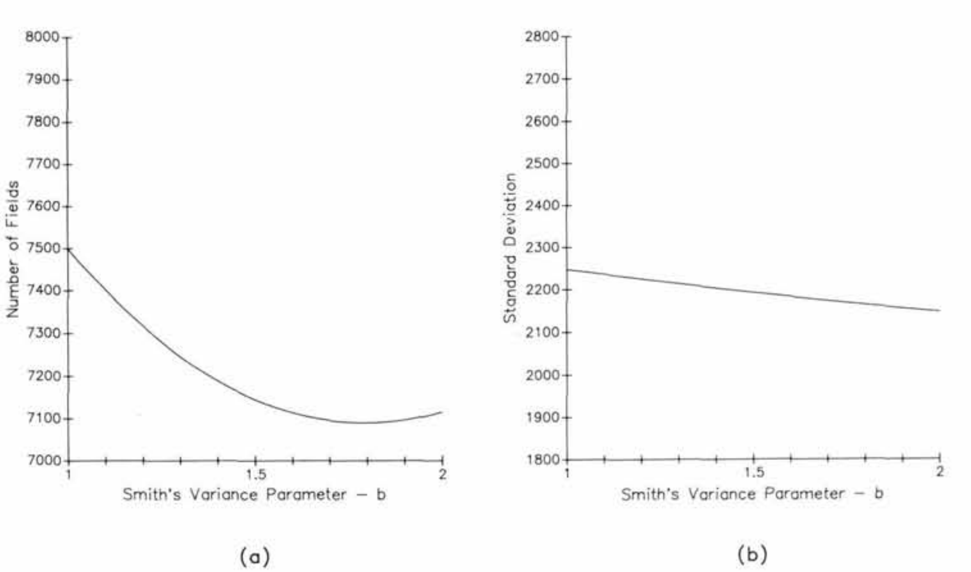

FIG. 1. (a) Estimated number of fields and (b) standard deviation as a function of b.

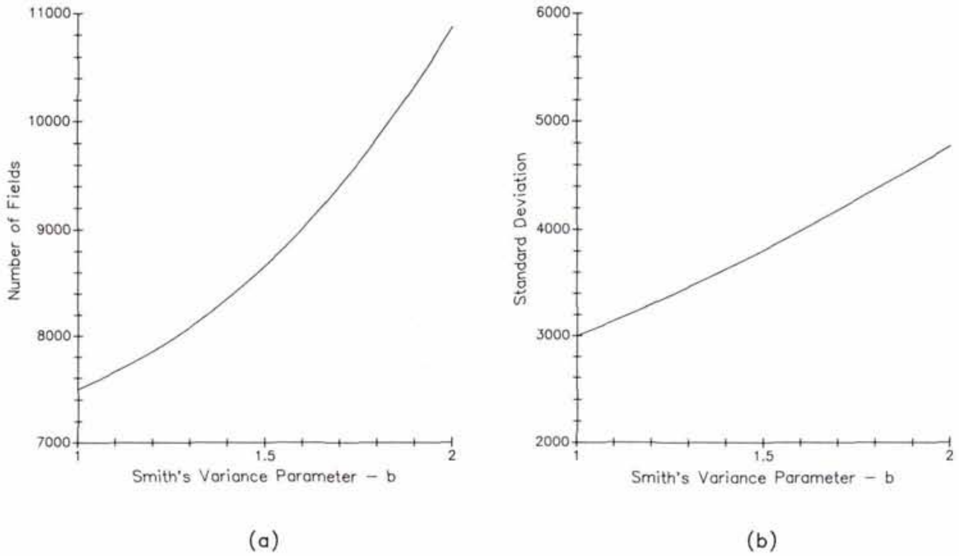

FIG. 2. (a) Estimated number of fields and (b) standard deviation using adjusted sample data.

of $Y_4 - Y$ is

$$\begin{aligned}\mathrm{Var}(Y_4 - Y) &= \mathrm{Var}\left[(Q - n\bar{q})\hat{\mu}_d - \sum_{i=1}^{N-n} d_i q_i\right] \\ &= (Q - n\bar{q})^2\,\mathrm{Var}(\hat{\mu}_d) + \mathrm{Var}\left(\sum_{i=1}^{N-n} d_i q_i\right)\end{aligned}$$

From Equations 8 and 4, this becomes

$$\mathrm{Var}(Y_4 - Y) = \frac{(Q - n\bar{q})^2\, B}{\sum_{i=1}^{n} q_i^{2-b}} + B(Q - n\bar{q})^b. \qquad (12)$$

Note that the variance given by Equation 12 goes to zero as the sample size increases. Thus, the variance of estimator Y_4 behaves similar to variances obtained under more classical assumptions that utilize a finite population correction term, (1-*f*), as in Equation 3 (e.g., Cochran, 1977). When $b = 1$, the above equation reduces to

$$\mathrm{Var}(Y_4 - Y) = B(Q - n\bar{q})Q/n\bar{q}$$

This equation was derived under slightly different assumptions in Cochran (1977, p. 159).

ESTIMATION OF *b*

To use these weighted estimates, the value of *b* must be known or estimated. One approach is to estimate the variance of C for at least two values of q, say q_1 and q_2, where $q_1 > q_2$. Then, from Equation 4, using the method of moments,

$$b = \frac{1\mathrm{n}(\mathrm{Var}(C_i\,|q_1)) - 1\mathrm{n}(\mathrm{Var}(C_i\,|q_2))}{\ln(q_1) - \ln(q_2)}$$

Such variance estimates can be acquired in a pilot survey, or directly from the survey itself, by subdividing frames of size q_1 into subframes of size q_2. Estimation of *b* is discussed in more detail in Proctor (1980), Proctor (1985), and Hatheway and Williams (1958).

DISCUSSION

Aerial sample surveys typically estimate a regional total either by extrapolating the sample total (Estimate Y_1; $b = 1$) or by extrapolating the average density (Estimate Y_2; $b = 2$). If the ground areas associated with each frame in the survey are of equal size, then these two estimates are equivalent. However, if the ground area imaged varies due to changes in the aircraft's altitude, cloud cover, etc, then these two estimates will differ, and they implicitly make different assumptions about the spatial correlation. Y_1 is optimal only if the fields are spatially uncorrelated while Y_2 is optimal only if the fields are perfectly correlated. In practice, most survey data will fall in between these two extremes.

As a first step, it is appropriate to calculate both Y_1 and Y_2, and their associated variances. This will determine the sensitivity of the survey to the two extreme spatial correlation assumptions. If the two estimates are similar, then there is little to be gained by undertaking a more detailed analysis. It has been our experience that, if the frame size, q_i, does not vary by more than a factor of 2 from smallest to largest in moderate sized sample surveys, then Y_1 and Y_2 will be similar.

If the two estimates do differ significantly, then the additional effort needed to estimate b is justified. Again there are two choices. Estimate Y_3, Equation 9, estimates the super population mean, $\mu_d Q$, while Y_4, Equation 11, estimates the particular realization of the total Y. Which is more appropriate depends upon the particular survey objectives. One way to resolve this issue is to ask whether it is desirable for the estimate to have zero uncertainty if the entire population is sampled. If so, then estimate Y_4 is the correct choice.

SUMMARY

Aerial sample surveys often must rely on the use of image frames that vary in ground area. In such cases, assumptions are required concerning the spatial correlation of the objects of interest in order to obtain appropriate estimates of the population totals. The estimates, although unbiased, will be particularly sensitive to these assumptions if the variation in frame sizes is significant. This paper discussed two weighted estimates, the choice of which depends on the nature of the survey. The weights chosen were a function of the ground area and the spatial correlation parameter, b, where $1 \leq b \leq 2$. The estimators include as limiting cases the estimators that extrapolate sample total and sample density, which implicitly assume $b = 1$ (spatial independence) and $b = 2$ (perfect spatial correlation), respectively.

REFERENCES

Cochran, W. G., 1977. *Sampling Techniques*, 3rd Edition, John Wiley and Sons, New York.

Hallum, C. R., and C. R. Perry, 1984. Estimating Optimal Sampling Unit Sizes for Satellite Surveys, *Remote Sensing of Environment*, Vol. 14, pp. 183–196.

Harrington, L., 1988. *Computer Generation of Fractional Brownian Noise*, ERIM Technical Report Number 658301-1-T.

Hansen, M. H., W. H. Hurwitz, and W. C. Madow, 1953. *Sampling Survey Method and Theory*, Vol. I, John Wiley and Sons, New York.

Hatheway, W. H., and E. J. Williams, 1958. Efficient Estimation of the Relationship Between Plot Size and Variability of Crop Yields, *Biometrics* Vol. 14, pp. 207–222.

Jessen, R. J., 1978. *Statistical Survey Techniques*, John Wiley and Sons, New York.

Press, J. S., 1972. *Applied Multivariate Analysis*, Holt Rinehart and Winston, Inc., New York.

Proctor, C. H., 1980. Estimating Smith's b from Sample Data, *American Statistical Association 1980 Proce. of the Section of Survey Research Methods*, pp. 761–765.

———, 1985. Fitting Smith's Empirical Law to Cluster Variances for use in Designing Multi-Stage Sample Suveys, *Journal of the American Statistical Association*, Vol. 80, No. 390, pp. 294–300.

Smith, H. F., 1938. An Empirical Law Describing Heterogeneity in the Yields of Agricultural Crops, *Journal of Agricultural Science*, Vol. 28, pp. 1–23.

(Received 27 September 1988; revised and accepted 13 June 1989)