L. Daniel Maxim
Leigh Harrington
Everest Consulting Associates
Princeton Junction, Nj 08550

To Mix or Match: On Choosing Matched Samples in Comparative Aerial Surveys

Equations are presented to characterize trade-offs in choosing between matched samples and unmatched samples.

Abstract: *The problem of sampling design in repetitive aerial surveys is addressed. Specifically, the trade-offs in sampling design to maximize either the precision of an estimate of the change in a population (between two surveys) or of an estimate of the current level of a survey variable are examined. Optimal designs for change detection require that a matched set of quadrats be selected. But optimal designs for estimation of levels of the survey variable often involve taking unmatched quadrats. Practical designs attempt to strike a balance between these objectives. This design problem is described and illustrated.*

Introduction

Aerial surveys are often designed with a multiplicity of objectives. A common purpose is to develop quantitative estimates of some population characteristic (e.g., number of houses, extent of vegatative damage, amount of forested area, the areal extent of various crops, etc.) at a given time period. But often estimates of the *change* in the survey variate between (among) survey periods are required. For example, in a longitudinal study of the impact of "slash and burn" cultivation in countries where this is practiced, it might be as or more important to measure the change in forested area between surveys as it is to estimate this quantity at any time (see the work of Royal Thai Forestry Department (1977) using Landsat imagery of Thailand).

An understanding of the relative importance of these objectives is essential for proper design of the survey because, in general, sampling plans that are optimal for one objective are not optimal for the other. Specifically, sampling plans that are best for estimating the *change* in a population differ from those that are efficient for estimating *levels* of a survey variate. Thus, a trade-off among survey objectives must be made. This article explores and illustrates this trade-off.

Problem Statement and Notation

The survey population consists of N quadrats or cells. A sample of n of these is selected in each of two time periods, the imagery is exploited, and the observed number of objects of interest, x_{it}, in each quadrat $i = 1, \ldots, n$ in each time period $t = 1, 2$ is recorded. To be concrete, suppose that the first survey has been completed and a design for the second period's survey is required. It is assumed that the true number of objects in each quadrat is identically distributed with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 in time periods 1 and 2, respectively. (If the quadrats are not identically distributed, the population can be partitioned into strata that satisfy this assumption.)

An important design feature is whether or not the same quadrats (i.e., matched samples) are selected in each of the surveys. Denote by f_m the fraction of the quadrats in the sample that are the same (i.e., matched) in the two time periods. That is, the sample of n quadrats is partitioned into two groups; a sample of size $n \cdot f_m$ that is matched over the time periods and a sample of size $n(1 - f_m)$ that is not matched (i.e., selected at random from the remainder of the population in time period 2). The design problem is to select a value for f_m. If f_m is set equal to unity, the same sample quadrats are

Photogrammetric Engineering and Remote Sensing,
Vol. 48, No. 12, December 1982, pp. 1863-1867.

0099-1112/82/4812-1863$02.25/0

selected in each period. If f_m is set equal to zero, a second random sample of quadrats is taken in period 2. Values of f_m between 0 and 1 yield hybrid designs—hence, the terms mix or match.

Design Considerations for Estimating Change

For simplicity, assume that only detection and sampling errors exist; see Maxim *et al.* (1981c) for a discussion of misclassification errors. Let p be the detection probability, assumed known and common to both time periods. The simple mean-per-unit or expansion estimate of the population total in period 1, $\hat{T}_1$, (based on all n samples) is given by

$$\hat{T}_1 = \frac{N}{np} \sum x_{i1}, \tag{1}$$

and has variance

$$\sigma^2_{\hat{T}_1} = \frac{N^2 K_1^2}{n}, \text{ where } K_1^2 = \left[\frac{\mu_1(1-p)}{p} + \sigma_1^2 \right]. \tag{2}$$

(See Maxim *et al.* (1981b) for proof.) The terms in the quantity K_1 are assumed known or are estimated from the sample mean and variance. Likewise the detection probability, p, is assumed known. (See discussions in Maxim *et al.* (1981a,c) for remarks on the estimation of this quantity.)

Two statistics are immediately available to estimate the change in the total from time period 1 to time period 2. The first is to base the estimate of change *only* on the set of matched quadrats. This estimate, denoted $\hat{\Delta}T_m$, is given by

$$\hat{\Delta}T_m = \frac{N}{npf_m} \sum (x_{i2} - x_{i1}), \tag{3}$$

where the sum is taken over the matched quadrats. Equation 3 is the simple difference equation together with the mean-per-unit estimate of the population totals in each time period. The variance of this estimate, σ_m^2, is (neglecting the finite population correction)

$$\sigma_m^2 = \frac{N^2}{nf_m} \left[K_1^2 + K_2^2 - 2\rho K_1 K_2 \right], \tag{4}$$

where

$$K_2^2 = \left[\mu_2(1-p)/p + \sigma_2^2 \right],$$

and

ρ = correlation coefficient between quadrat totals in each time period.

Note that (before the fact) μ_2, σ_2, and ρ are not known and so surrogates for these quantities may have to be used. Earlier studies, for example, may be used to estimate ρ. Likewise, trend extrapolation can often be used to estimate μ_2 and σ_2.

Alternatively, the estimate of change can be based on only the unmatched set. This estimate, $\hat{\Delta}T_u$, is given by

$$\hat{\Delta}T_u = \frac{N}{n(1-f_m)p} (\sum x_{i2} - \sum x_{i1}) \tag{5}$$

where the sums are taken over the unmatched quadrats. The variance of this estimate, σ_u^2, (the subscript u denoting unmatched) is

$$\sigma_u^2 = \frac{N^2}{n(1-f_m)} \left[K_1^2 + K_2^2 \right]. \tag{6}$$

These estimates, $\hat{\Delta}T_m$ and $\hat{\Delta}T_u$, can be optimally combined to form another estimate, $\hat{\Delta}T$, as

$$\hat{\Delta}T = w_m \hat{\Delta}T_m + (1 - w_m)\hat{\Delta}T_u \tag{7}$$

and the optimal choice for w_m to minimize the variance in $\hat{\Delta}T$, w_m^*, is given by (see Hodges *et al.* (1970) for a discussion of weighted linear estimates)

$$w_m^* = \sigma_u^2/(\sigma_m^2 + \sigma_u^2). \tag{8}$$

After some manipulation, the variance of this estimate, σ_d^2, reduces to

$$\sigma_d^2 = \frac{N^2 \left[K_1^2 + K_2^2 - 2\rho K_1 K_2 \right] \left[K_1^2 + K_2^2 \right]}{n \left[K_1^2 + K_2^2 - 2\rho(1-f_m) K_1 K_2 \right]}. \tag{9}$$

Inspection of Equation 9 shows how the optimal weighted linear estimate of the period to period change depends upon f_m. For positive values of ρ (as might be expected in practice), it follows that σ_d^2 is minimized when f_m is set equal to unity. That is, to maximize the precision of an estimate of change, the survey design should consist exclusively of matched quadrats. For this design, $w_m^* = 1.0$ (i.e., only the estimator $\hat{\Delta}T_m$ is used) and the variance is

$$\sigma_d^2(f_m = 0.1) = \frac{N^2}{n} \left[K_1^2 + K_2^2 - 2\rho K_1 K_2 \right]. \tag{10}$$

Note also that the estimate becomes more precise as ρ increases. This parallels a well known (see Cochran (1977)) result in sampling theory when detection errors are not considered.

The relative efficiency of the optimal matched design compared to a design where there is no matching ($f_m = 0.0$) can be calculated by contrasting Equations 10 and 9. Denoting $\theta = K_2/K_1$, the relative sample size of a matched experiment, compared to that for an unmatched experiment of equal precision, is given by

$$1 - 2\rho(\theta/(1 + \theta^2)), \tag{11}$$

and is a function of both ρ and θ. Efficiency gains from matching are greatest when $\theta = 1.0$ and $\rho = 1.0$ and less elsewhere. Table 1 shows how the relative sample size required depends upon θ and ρ (the results are symmetric in θ, i.e., values for $\theta =$

TABLE 1. RATIO OF SAMPLE SIZE OF MATCHED TO UNMATCHED SAMPLING PLANS FOR ESTIMATING PERIOD-TO-PERIOD CHANGE

	$\theta = K_2/K_1$			
ρ	1.0	2.0	5.0	10.0
0.0	1.0	1.0	1.0	1.0
0.1	0.9	0.920	0.962	0.980
0.2	0.8	0.840	0.923	0.960
0.3	0.7	0.760	0.885	0.941
0.4	0.6	0.680	0.846	0.921
0.5	0.5	0.600	0.808	0.901
0.6	0.4	0.520 *	0.769	0.881
0.7	0.3	0.440	0.731	0.861
0.8	0.2	0.360	0.692	0.842
0.9	0.1	0.280	0.654	0.822
1.0	0.0	0.200	0.615	0.802

* Example in text.

0.2 are identical to those for $\theta = 5$). As can be seen, matched designs can be very much more efficient than unmatched designs. For example, when $\theta = 2.0$ and $\rho = 0.6$, a matched design requires a sample size of only 52 percent of that for an unmatched design—a 48 percent savings in sample size.

DESIGN CONSIDERATIONS FOR ESTIMATING LEVELS

If the survey objective is simply to estimate the total in period 2, however, a different design may be appropriate. This problem has been considered at length (see Cochran (1977) or Kulldorf (1963), for example) for the case where detection errors are not present. Results are summarized here with appropriate modification to include detection errors. As above, the sample is partitioned into matched and unmatched portions. The mean-per-unit or expansion estimate in period 2 from the unmatched portion, $\hat{T}_{u2}$, is given by

$$\hat{T}_{u2} = \frac{N}{n(1 - f_m)p} \sum x_{i2}, \tag{12}$$

with variance equal to $N^2K_2^2/n(1 - f_m)$. The estimate in period 2 from the matched portion, $\hat{T}_{m2}$, is given by the regression estimate,

$$\hat{T}_{m2} = \frac{N\sum x_{i2}}{np\, f_m} + b\left[\hat{T}_1 - \frac{N\sum x_{i1}}{np\, f_m}\right], \tag{13}$$

where

b = coefficient of the regression equation for period 2 estimates on period 1 estimates. b would, of course, have to be estimated from the data.

This estimate has variance

$$N^2\left(\frac{K_2^2(1 - \rho^2)}{nf_m} + \rho^2\left[\frac{K_2^2}{n}\right]\right). \tag{14}$$

As before, an optimal linear combination $\hat{T}_2 = w\hat{T}_{u2} + (1 - w)\hat{T}_{m2}$ can be developed to estimate the survey total in period 2. The variance of this best estimate, denoted $\sigma_{\hat{T}2}^2$, is given by

$$\sigma_{\hat{T}2}^2 = \frac{N^2K_2^2(1 - (1 - f_m)\rho^2)}{n(1 - (1 - f_m)^2\rho^2)}. \tag{15}$$

Inspection of Equation 15, however, indicates that this variance is not minimized for $f_m = 1$, but rather (upon differentiation) at a value, f_m^*, equal to

$$f_m^* = (1 - \rho^2)^{1/2}/(1 + (1 - \rho^2)^{1/2}). \tag{16}$$

It is interesting to note that the optimal matching fraction is independent of the detection probability, quadrat means or variances, etc., and is solely a function of ρ. Likewise, the efficiency gains from optimal designs depend only on ρ. Optimal designs in this case are, therefore, identical to those for sampling in cases where detection errors do not occur. Table 2 shows how the optimal value for f_m and the relative efficiency of the optimal designs (compared to the case where f_m is unity) vary with p. For example, if p is 0.5, 46.4 percent of the quadrats should be matched. The resulting plan offers a 6.7 percent reduction in sampling effort in comparison to a plan where $f_m = 1.0$. Note that, for this purpose, sampling economics are greatest in cases where the correlation coefficient is near unity.

THE TRADE-OFF ILLUSTRATED

Thus, there is a trade-off that is required among design objectives. A design that maximizes the precision of the estimate of change in the population requires that *all* survey quadrats be matched ($f_m = 1.0$). But, a survey designed to maximize the precision of the current period's estimate requires that only *some* of the quadrats be matched.

Table 3 provides a numerical illustration of this designer's dilemma. A sample of 100 quadrats is taken from a population of size 1000 on two occasions. The assumed period-to-period correlation is moderate and positive ($\rho = 0.5$), the detection probability is 0.8 and other quantities are as shown on the bottom on Table 3. The various columns show σ_u^2, σ_m^2, w_m^*, σ_d^2, and $\sigma_{\hat{T}2}^2$ as a function of f_m. The table also calculates a measure of the relative precision of the estimates of change and level in period 2 as the proportional half-width of a 95 percent confidence level. Its use is illustrated in an example to follow. These relative precision values are plotted in Figure 1.

For this example, the choice of f_m that minimizes the current period's variance is 0.46—i.e. out of the sample of 100, 46 are to be matched and 54 are left unmatched. The relative precision of the period 2 total estimate is ±2.65 percent. The design that maximizes the precision of the estimate of the period 1, period 2 change

TABLE 2. OPTIMAL MATCHING PLANS AS A FUNCTION OF THE PERIOD-TO-PERIOD CORRELATIONS

Period-to-Period Correlation Coefficient, ρ	Optimal Fraction Matched f_m	Relative Sample Size Required for Optimal Plan to Plan with $f_m = 1.0$	Savings as a Percent
0	0.500	1.000	0.0
0.1	0.499	0.997	0.3
0.2	0.495	0.990	1.0
0.3	0.488	0.977	2.3
0.4	0.478	0.958	4.2
0.5	0.464	0.933	6.7
0.6	0.444	0.900	10.0
0.7	0.417	0.857	14.3
0.8	0.375	0.800	20.0
0.85	0.345	0.763	23.7
0.90	0.304	0.718	28.2
0.92	0.282	0.696	30.4
0.94	0.254	0.671	32.9
0.96	0.219	0.640	36.0
0.98	0.166	0.599	40.1
1.0	~0	0.500	50.0
Equation	$(1 - \rho^2)^{1/2}/(1 + (1 - \rho^2)^{1/2})$	$0.5\,(1 + \sqrt{1 - \rho^2})$	

Source: Suggested from a Table in Cochran (1977).

requires that all 100 samples are matched. The relative precision of the estimate of change is ± 4.88 percent. But, as is shown in Figure 1, the precision of one estimate can only be improved at the expense of the other. Note that, in this example, it would not be rational to consider values of f_m less than 0.44, because this would reduce the precision of *each* estimate. This "area of irrationality" is shown on Figure 1. For values of f_m greater than 0.44, however, real trade-offs are required, and a rational choice can only be made by balancing or weighing the relative importance of each design objective. Note that, in this example, the overall sample size ($n = 100$) is sufficient to estimate both quantities quite accurately, so that the trade-off is not dramatic, but for smaller n each of the estimates is less precise and the trade-off is of greater relevance.

CONCLUDING REMARKS

In repetitive surveys the systems analyst often has to balance the twin objectives of characterizing the current population with detecting and estimating change in the population. Optimal de-

TABLE 3. DETAILS OF NUMERICAL EXAMPLE

1	2	3	4	5	6	7	8	9	10
$n(1 - f_m)$	nf_m	f_m	σ_u^2	σ_m^2	w_m^*	σ_d^2	Relative Precision	σ_{T2}^2	Relative Precision
0	100	1.00	∞	237,900	1.0	237,900	4.88	300,000	2.74
10	90	0.90	4,500,000	264,333	0.945	249,667	5.00	293,233	2.71
20	80	0.80	2,250,000	297,375	0.883	262,660	5.13	287,879	2.68
30	70	0.70	1,500,000	339,857	0.815	277,079	5.26	283,887	2.66
40	60	0.60	1,125,000	396,500	0.739	293,173	5.41	281,250	2.65
50	50	0.50	900,000	475,800	0.654	311,252	5.58	280,000	2.65
54	46	0.46	833,333	517,174	0.617	319,123	5.65	279,905	2.65
60	40	0.40	750,000	594,750	0.558	331,707	5.76	280,219	2.65
70	30	0.30	642,857	793,000	0.448	355,039	5.96	282,051	2.66
80	20	0.20	562,500	1,189,500	0.321	381,903	6.18	285,714	2.67
90	10	0.10	500,000	2,379,000	0.174	413,164	6.43	291,536	2.70
100	0	0.00	540,000	∞	0	450,000	6.71	300,000	2.74

Other Assumptions:

$N = 1{,}000$ $\quad \mu_1 = 20$ $\quad \sigma_1^2 = 10$ $\quad K_1^2 = 15$

$n = 100$ $\quad \mu_2 = 40$ $\quad \sigma_2^2 = 20$ $\quad K_2^2 = 30$

$p = 0.80$ $\quad \rho = 0.5$

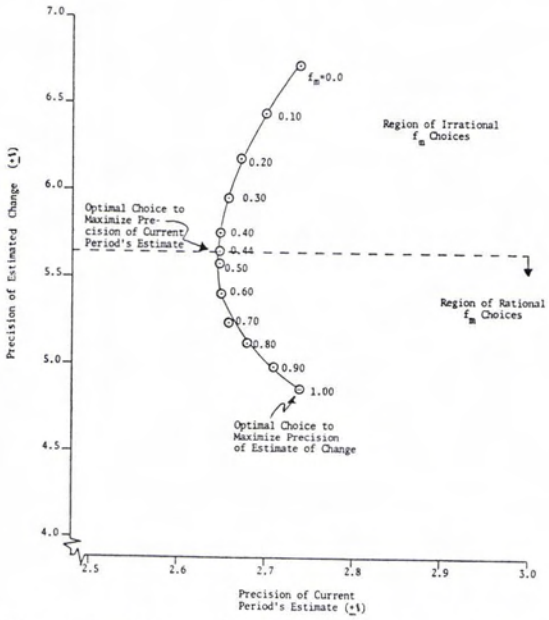

FIG. 1. Precision tradeoffs for fixed imagery budget.

signs for change detection necessitate using a matched sample of quadrats. Any departure from this design decreases the precision of the estimate.

But, if estimates are required for the current time period, optimal designs involve choosing unmatched samples—the optimal degree of matching is a function of the period-to-period correlation in the survey variate. Equations are presented to characterize trade-offs in these two measures of precision.

This is written on the assumption that observation is "passive." If changes in the population occur due to the survey itself, e.g., crops are sprayed, reforestation initiated, illegal stills confiscated, channels dredged, etc., then additional considerations are relevant.

Finally, it should be remembered that the specific equations presented here are developed for the specific case of two time periods. These results are easily generalized to more than two time periods—with broadly similar results—though at some increase in complexity.

ACKNOWLEDGMENTS

The authors would like to thank Drs. Nancy David, Mary Kennedy, and Harrison Weed for stimulating discussions on this problem. The referees also made several constructive comments regarding this manuscript.

REFERENCES

Cochran, W. G., 1977. *Sampling Techniques,* 3 ed., John Wiley and Son, New York.

Hodges, J. L., Jr., and E. L. Lehmann, 1970. *Basic Concepts of Probability and Statistics,* Holden-Day, San Francisco, p. 288.

Kulldorff, G., 1963. Some Problems of Optimal Allocation for Sampling on Two Occasions, *Review International Statistical Institute,* Vol. 31, pp. 24-57.

Maxim, L. D., L. Harrington, and M. Kennedy, 1981a. A Capture-Recapture Approach for Estimation of Detection Probabilities in Aerial Surveys, *Photogrammetric Engineering and Remote Sensing,* Vol. 47, June 1981, pp. 779-788.

Maxim, L. D., H. D. Weed, L. Harrington, and M. Kennedy, 1981b. Intensity Versus Extent of Coverage, *Photogrammetric Engineering and Remote Sensing,* Vol. 47, June 1981, pp. 789-797.

Maxim, L. D., L. Harrington, and M. Kennedy, 1981c. Alternative 'Scale Up' Estimates for Aerial Surveys Where Both Detection and Classification Errors Exist, *Photogrammetric Engineering and Remote Sensing,* Vol. 47, August 1981, pp. 1227-1239.

Royal Thai Forestry Department, 1977. *The Assessment of Forest Area from Landsat Imagery.*

(Received 5 December 1981; accepted 12 May 1982; revised 9 June 1982)