Monitoring to Detect Changes in Water Quality Series (Proceedings of the Budapest Symposium, July 1986). IAHS Publ. no. 157,1986.

Calibrating SALT: a sampling scheme to improve estimates of suspended sediment yield

ROBERT B. THOMAS
Pacific Southwest Forest and Range Experiment Station, Forest Service, US Department of Agriculture, 1700 Bayview Street, Arcata, California 95521, USA

ABSTRACT SALT (Selection At List Time) is a variable probability sampling scheme that provides unbiased estimates of suspended sediment yield and its variance. SALT performs better than standard schemes which are

estimate variance. Sampling probabilities are based on a sediment rating function which promotes greater sampling intensity during periods of high sediment yield. When preparing to monitor using SALT, the quality of the existing suspended sediment rating data is an important consideration. SALT variance, based on an intensive data set from a river in northwestern California, showed greater sensitivity to the SALT sample size than to the quality of the calibration set.

INTRODUCTION

Probability-based sampling provides several advantages over conventional schemes in collecting concentration data for estimating suspended sediment yield in rivers. For one, the behaviour of the estimator is known. If the estimator is unbiased, no systematic distortion exists between estimated and actual values. Because probability samples estimate variance, it is also possible to estimate errors, set confidence intervals, test hypotheses, and determine required sample sizes. Comparisons between treatments are statistically valid only when estimates of variance are available.

These advantages notwithstanding, most suspended sediment data are not collected according to any probability-based scheme. Estimates often have a large but unknown bias and valid estimates of variance cannot be made (Walling & Webb, 1981). This is due in part to difficulties inherent in measuring suspended sediment and to the sporadic nature of the processes producing it. The chief problem, however, is the lack of a strategy for deciding when measurements should be made so that estimators of the total and variance have acceptable properties.

A new approach to sediment sampling has recently been developed that provides unbiased estimates of total suspended sediment yield and its variance (Thomas, 1985). The technique is called SALT, an acronym for Selection At List Time, which derives from its original application to sampling timber volume (Norick, 1969). The technique is based on variable probability sampling and uses an estimate of the concentration given by a suspended sediment rating to direct sampling. The method emphasizes the sampling of higher, more sediment laden flows which is consistent with our understanding of how such a sampling plan should work. Because the sampling probabilities are known, they can be used to remove bias in the estimators.

A theoretical link between the sampling protocol and the estimating procedures also allows sample size to be set to obtain sufficient precision of the estimators. Establishing sample size usually requires some estimate of the variance. SALT is no exception. Some preliminary sediment rating data are required to set sample size (Thomas, 1985), as well as to operate a SALT scheme. This paper investigates the requirements for this calibration data set.

USING SALT TO SAMPLE SUSPENDED SEDIMENT

Different probabilities can be assigned to the elements of a finite population to make sampling more efficient, i.e., to reduce the sampling variance. If our knowledge of the population allows higher probabilities to be assigned to the more important elements (i.e. those contributing more to the mean or total), we can either spend less effort on the sampling for a given precision or obtain lower variance for the same effort (Murthy, 1967). An easily measured auxiliary variable which is a good predictor of a specified variable can be used to establish the probabilities.

Such a variable is often available for sampling suspended sediment populations. A reasonable prediction of concentration is usually possible from an empirical sediment rating. This prediction can be used as the auxiliary variable to define probabilities. Significantly, the relationship between the variable of interest and the auxiliary variable need not be deterministic. A statistical relationship with appropriate properties is adequate.

The conceptual finite population to be sampled is formed by dividing the monitoring period into N short sampling periods of equal duration. (Period refers here to a portion of the time axis on the suspended sediment discharge hydrograph). The investigator can choose the duration of the periods to ensure that the finite population is sufficiently similar to the continuous population. from which it is formed. The elements of the population to be sampled consist of a measure of suspended sediment yield from each sampling period. For sampling period i, let y_i be the total suspended sediment yield, q_i, the mid-interval water discharge, and c_i the mid-interval suspended sediment concentration determined from an actual physical sample. Then

$$y_i = q_i\, c_i\, \Delta t\ K \qquad (1)$$

where Δt is the time duration of the sampling period and K is a constant to adjust units. Therefore, y_i is a *measure* of the total mass of suspended sediment during sampling period i. If y_i was known for all sampling periods, their sum would approximate the true total mass of suspended sediment in the monitored period. If discharge is measured continuously, q_i can be found for each sampling period, but c_i is available only for those periods actually sampled. Therefore, y_i is also only known for those relatively few sampling periods for which an actual concentration sample was collected.

Now define x_i identically to y_i, except that $\hat{c}_i$ replaces c_i. That is,

$$x_i = q_i \ \hat{c}_i \ \Delta t \ K \tag{2}$$

where $\hat{c}_i$ is an estimate of concentration derived from a sediment rating. Then x_i is an *estimate* of the mass of suspended sediment delivered during the ith sampling period. Because $\hat{c}_i$ is derived from a sediment rating as a function of discharge, it is available for every sampling period, so x_i can be used for defining the probabilities of selection and to adjust the estimators to remove bias.

The value of x_i must be known on a realtime basis and it must be calculated for all sampling periods in the population. For small streams of highly varying concentrations where Δt must be set quite short (e.g. 5-30 minutes),it is necessary to automatically control the sampling process. Stations for measuring sediment in small streams usually have a water stage recorder and a pumping sampler. The SALT procedure additionally requires a transducer to sense stage, a field computer to store and perform the sampling algorithm, and an interface circuit board to connect the equipment. For larger more stable rivers, where Δt is of the order of a day or so, it may be possible to calculate and manually measure discharge and calculate $\hat{c}_i$.

The SALT sampling scheme depends on the magnitude of x_i values to set probabilities. Because the x_i values are not known until real time, a method is required to select random sampling numbers prior to the measurement period. A high estimate is made of the total yield expected during the period to be monitored using an excess of random numbers. Let Y' be an estimate of the upper limit of yield expected in the period to be monitored. Then multiply Y' by a factor W (say W = 10 or so) to obtain Y = W Y'. Y is therefore larger than the expected mass of suspended sediment for the period monitored.

Thomas (1985) describes a method to determine n, the number of random numbers needed for adequate performance of SALT estimators. Consider the Y axis containing n uniform random numbers between zero and Y (Fig. 1). For each of the equal duration sampling periods on the time axis of the suspended transport hydrograph, exactly one sampling interval is placed on this axis, its length equal to the estimated suspended sediment yield, x_i, for the period. Therefore, the Y axis is also the sampling interval axis.

Sampling can begin once Y is established and the n random numbers are selected, sorted, and stored in the calculator. In the middle of each sampling period, the calculator reads the water stage and calculates the values of q_i, $\hat{c}_i$ and x_i. A sampling interval of length x_i is placed on the sampling-interval axis immediately after the sampling interval formed from the

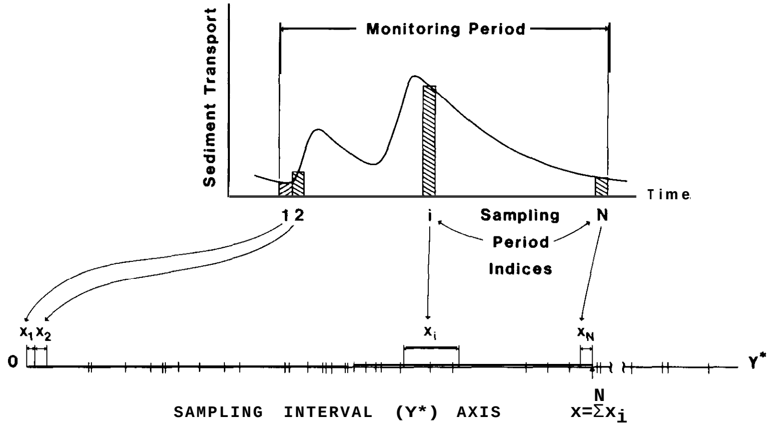

FIG.1 Suspended sediment transport hydrograph and corresponding sampling interval axis for SALT sampling. The correspondence is between the equal duration sampling periods on the time axis and the variable length intervals of estimated suspended sediment discharge on the sampling interval axis. Ticks on the sampling interval axis denote random sampling numbers.

previous period (Fig. 1). If no random number is within this interval, the calculator stores needed information and waits for the next period. If, however, one or more random numbers fall in the interval, the period is sampled by operating the pumping sampler.

At the end of the period monitored, the values of c_i and q_i are known for the periods actually sampled so the corresponding values of y_i can be calculated. Because x_i was calculated for all sample periods, the total estimated suspended sediment yield $X = \Sigma^N x_i$ is also known. Letting p_i indicate the probability that period i is sampled, we have $p_i = x_i/X$, because uniform selection of random numbers over the interval 0 to Y ensures that the numbers are also distributed uniformly over 0 to X, provided X < Y .

Let N be the total number of sampling periods in the period monitored, r_i the number of random values in the ith sampling interval, and n the number of random numbers on the sampling interval axis up to X. The sample size is n, which equals the sum of the r_i across all N sampling periods. The value of n is random, its magnitude depending on the density of random points on the Y* axis and on the level of discharges in the river during the monitored period. Occasionally, more than one random number falls in a sampling interval (i.e. $r_i > 1$). In this case only one concentration sample is taken, but its value is included r_i times in the estimators.

Knowing the p_i for the sampled periods enables the estimators to be adjusted, making them unbiased. Let Y be the true total suspended sediment during the monitored period (i.e., $Y = \Sigma^N y_i$ and $\hat{Y}$ its SALT estimate. Then,

$$\hat{Y} = \frac{1}{n} \sum^N r_i \frac{y_i}{p_i} \tag{3}$$

Denote the SALT estimate of the variance of $\hat{Y}$ by $S^2(\hat{Y})$ Then,

$$S^2(\hat{Y}) = \frac{1}{n(n-1)} \sum^{N} r_i \left(\frac{y_i}{p_i} - \hat{Y} \right) \qquad (4)$$

Equation (4) estimates only the sampling variance. Other sources contributing to the total variance include measurement error and error due to approximating the true population with a finite set of intervals (Thomas, 1983). Estimators are more completely discussed in Norick(1969) and Thomas (1985).

AUXILIARY VARIABLES

Because suspended sediment transport is a sporadic process, selecting sampling periods with equal probabilities would give a very inefficient sample, especially in small, rain dominated streams. High flows are relatively rare; hence, most samples would be collected during frequent, though unimportant low discharges. Simple random sampling estimators would be unbiased, but the variance of the estimates of the total would be high.

The SALT scheme has the advantage of setting probabilities for the sampling periods using an auxiliary variable obtained from a suspended sediment rating. As long as the selection probabilities are known, however they are assigned, the SALT estimates are unbiased. The variance of the estimates of the total, however, are reduced in relation to how well the assignments of probability reflect the true importance of the sampling periods in determining total suspended sediment yield. The SALT scheme permits taking advantage of what we know or can predict about the process to improve sampling in a way that still allows valid estimates of total and variance. Only those rating functions derived using least squares regression are considered.

Presumably, SALT performs well when the variable of interest, y, and the auxiliary variable, x, are highly correlated. This is true in general, but it is possible to construct examples where x and y have a correlation of one, yet variable probability sampling does not perform as well as simple random sampling (Raj, 1968). For SALT to have lower variance than simple random sampling, it is necessary to have Corr $(y^2/x, x) > 0$. This condition seems to hold for most sediment rating data sets, but it should be verified for rating data sets intended for SALT sampling. (The actual values of y and x need not be calculated in order to compute this correlation, because they both have the same factors which cancel out in the formula. It is therefore sufficient to determine the Corr $(c_i^2/\hat{c}_i, \hat{c}_i)$ for the n' rating points, where c_i is the ith measured concentration and $\hat{c}_i$ is the corresponding predicted concentration from the rating regression). Our interest lies not only in obtaining a variance lower than that in simple random sampling, but in having an acceptable probability that the estimates of total suspended sediment are accurate enough for the purpose intended.

The true SALT variance for the estimate of total suspended sediment yield, Y, is given for a sample of size n by

$$\operatorname{var}(\hat{Y}) = \frac{1}{n}\left[X \sum_{i=1}^{N} (y_i^2/x_i) - Y^2\right] \qquad (5)$$

Equation (5) defines the component of variance due solely to random selection of the set of sample periods in the SALT scheme. Var(Y) is smaller as x_i is closer to being proportional to y_i. When the proportion is exact, the variance is zero. This implies that all information in y is contained in x, and because X is known, so is Y.

Another component of variance associated with the SALT scheme derives from determining the rating function. When setting up a SALT sampling scheme, it is important to understand how this additional component compares to the variance in equation (5). It is unclear initially whether the hydrologist should try to improve the rating function or collect a modest set of rating data and reduce overall variance by increasing the SALT sample size. It may be preferable to collect a larger set of rating data having a better distribution of discharges. Conversely, it may be preferable to increase the value of n to achieve the desired precision in the estimators.

Data from a station on the north fork of the Mad River near Korbel in northwestern California will be used to investigate this question. This 10 442 ha basin is in highly erosive terrain and receives about 1 270mm annual precipitation. A large set (933 discharge/concentration pairs) of rating data was collected before a major 4 day storm during 15-19 December, 1983. During the storm a continuous record of turbidity was collected, and estimates of suspended sediment concentration were made at 5 minute intervals using an excellent calibration between concentration and turbidity. Values for y_i are therefore available at 5 minute intervals throughout the storm, and the large set of rating data available for this station enabled high quality estimates of the rating function parameters. The actual SALT variance for this storm (conditional on realized sample size) can be calculated by assuming that the estimated parameters of the rating model are true.

Consider a hypothetical set of n' rating points (q'_i, c'_i), $i = 1, ..., n'$, where

$$\ln c'_i = \ln A + B \ln q'_i + e'_i \qquad (6)$$

in which lnA and B are the coefficients of the underlying model and $e'_i \sim N(0,\sigma^2)$. Regression estimates lna and b for lnA and B give concentration estimates for SALT sampling in the form

$$\ln \hat{c}_i = \ln a + \ln q_i \qquad (7)$$

or alternatively

$$\hat{c}_i = a\, q_i^b \qquad (8)$$

where the q_i are from the SALT population being sampled rather than the calibration rating data set. Remember that X in equation (5) is the sum of the x_i. Substitute equation (8) in equation (2), and the resulting expression for x_i in equation (5). The variance is then

$$\mathrm{Var}(\hat{Y}) = \frac{1}{n} \left(\sum_{i=1}^{N} q_i^{b+1} \sum_{j=1}^{N} \frac{y_j^2}{q_j^{b+1}} - Y^2\right) \qquad (9)$$

or

$$\mathrm{Var}(\hat{Y}) = \frac{1}{n} \left[\sum_{j=i}^{N} y_i^2 \sum_{j=i}^{N} \left(\frac{q_j}{q_i}\right)^{b+l} - Y^2\right] \qquad (10)$$

where Y is the total of the y_i over all N periods.

Because the y_i are assumed fixed (that is, known) for this particular population and we are conditioning on n, the only random variable in equation (10) is b. But b is the rating regression coefficient which has a normal distribution with mean B and variance σ^2/Q', where

$$Q' = \Sigma^{n'} (\ln q'_i - \overline{\ln q}')^2 \qquad (11)$$

and

$$\overline{\ln q}' = \frac{1}{n'} \Sigma^{n'} \ln q'_i \qquad (12)$$

Also, the quantity

$$(q_j/q_i)^{b+1} = \exp[(b+1)R_{ij}] \qquad (13)$$

where,

$$R_{ij} = \ln(q_j/q_i) \qquad (14)$$

Now $\exp[(b + l)R_{ij}]$ is distributed lognormally with the associated normal distribution having mean $(B + 1)R_{ij}$ and variance $R_{ij}^2\sigma^2/Q'$. Taking the expected value of equation (13) gives

$$E[(q_j/q_i)^{b+1}] = \exp[(B+1)R_{ij} + R_{ij}^2\sigma^2/2Q'] \qquad (15)$$

and application to equation (10) yields

$$E(\mathrm{Var}(\hat{Y})] = \left\{\frac{1}{n} \Sigma_{i=1}^{N} y_i^2 \Sigma_{j=1}^{N} \exp[(B+l)R_{ij}+R_{ij}^2 \frac{\sigma^2}{2Q'}] - Y^2\right\} \qquad (16)$$

The variation Var $(\hat{Y})$due to the SALT sample is included the expected value of equation (16) as well as the variation in estimating the regression coefficients A and B; that is, the variation due to calibration. If we accept the values for A,B, and σ^2 determined in the rating regression, we can calculate estimates for $E[\mathrm{Var}(\hat{Y})]$ and $\mathrm{Var}(\hat{Y})$ to evaluate their relative magnitude for this particular stream and storm. To do so we need to choose values for q'_i. The range of the rating discharges, $\ln q'_i$, in the real rating set was partitioned into five classes of equal length. Values of $\ln q'_i$ were selected at equal increments of discharge within each of the five classes. The increment for each class was chosen to approximate four different distributions of rating discharges (Fig. 2). Distribution A is uniform in the ln discharges, B is skewed to the left, C to the right, and distribution D is symmetrical with most values in the center range of ln discharges. Calculations were made for n' of 50, 100, 150, and n of 24, 36 and 48.

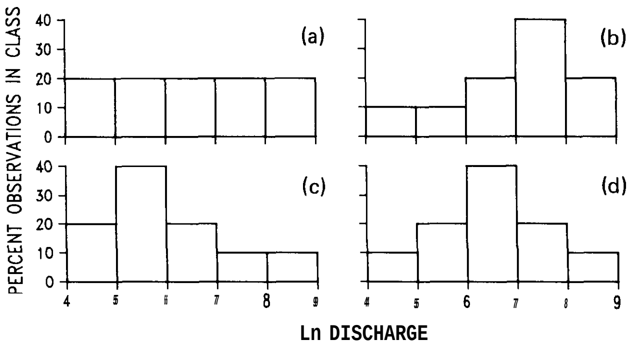

FIG. 2 Distributions of ln discharge values in set of artificial suspended sediment rating data.

The first results of this analysis show the variance due to calibration, i.e., the differences between E[Var($\hat{Y}$)] and Var($\hat{Y}$), as a percentage of E[Var($\hat{Y}$)] plotted as a function of n' (Fig. 3). Different curves are shown for each distribution of rating discharges (Fig. 2). These quantities do not depend on n because the factor 1/n cancels when calculating percentage. As expected, the percentages drop as n' increases. Using distribution A as an example, the percentage when n' = 50 is about 3.65, it drops to 1.86 when n' = 100, and further drops to 1.24 when n'=150. By tripling the size of the calibration rating data set, therefore, we can reduce the variance due to calibration by about two-thirds.

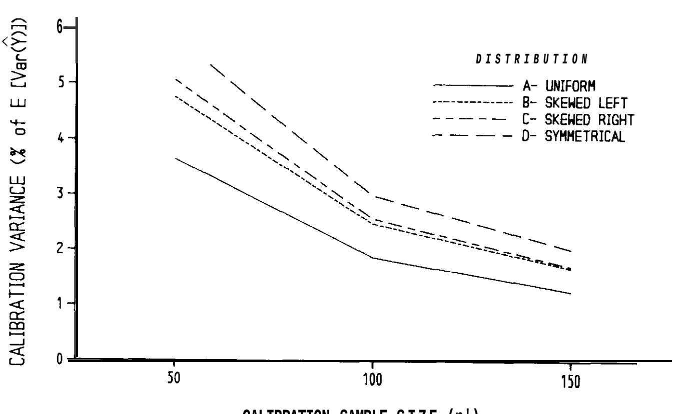

FIG. 3 Calibration variance expressed as percentage of E[Var(Y)]and related to the size of calibration data set n'. Distributions A, B,C and D are as shown in Fig. 2.

The distributional form of the ln rating discharges also has an effect on the calibration variance. Distribution A has the lowest variance of the four distributions considered. The two skewed distributions, B and C, have somewhat larger variance, but are similar to each other. The symmetrical distribution, D, with most data in the center of its range, has the largest variance. At larger values of n' the effect of the lnq'i distributions on calibration variance decreases.

These results are consistent with the fact that collecting regression data towards the extremes of the range provides the best estimate of model parameters. If the regression model form is known to be linear, the data are best collected only at the extremes. Because linearity is not certain in the rating regression case, the more or less uniform distribution (in ln discharges) is probably the best pattern to strive for.

An overwhelming factor, however, is that the variance due to estimating the regression rating function is a small part of $E(Var(\hat{Y})]$. This fact is indicated not only by the small values of the percentages, but also by plotting the values of $E[Var(\hat{Y})]$ (Fig. 4). The variance axis was divided into three sections to expand the vertical axis for the three values of n. In Fig. 4, the scale of the variance axis is constant but the change in magnitude of the variance between the three graphs for different values of n is very large. It is obvious that the change in $E[Var(\hat{Y})]$ due to either the distribution of rating discharges or their number is small relative to the change produced by an increase in the SALT sample size.

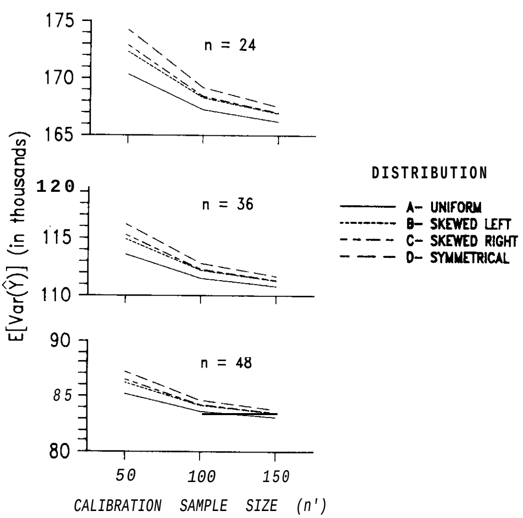

FIG. 4 $E[Var(\hat{Y})]$ as a function of the size of the calibration data set, n', for various values of the SALT sample size, n.

For this river and particular set of storm flows, some reduction in the variance can be realized by improving the size and quality of the rating function. A much greater reduction can be obtained, however, by increasing the size of the SALT sample. It seems reasonable to apply these results to other streams until futher testing suggests otherwise.

ACKNOWLEDGEMENTS I thank James A. Baldwin for his help in formulating the statistical model used in this paper, and the Simpson Timber Company for its co-operation with the Forest Service, U.S. Department of Agriculture, in collecting data at the test site.

REFERENCES

Murthy, M.N. (1967) *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.

Norick, N.X. (1969) SALT (selection at list time) sampling: a new method and its application in forestry. MA Thesis, Univ. of California, Berkeley.

Raj, D. (1968) *Sampling Theory.* McGraw-Hill, New York.

Thomas, R.B. (1983) Errors in estimating suspended sediment. In: (Proc. D.B. Simons Symp. *Erosion and Sedimentation,* Fort Collins, Colorado) 1.162-1.177.

Thomas, R.B. (1985) Estimating total suspended sediment yield with probability sampling. *Water Resour. Res.* 21(9), 1381-1388.

Walling, D.E. & Webb, B.W. (1981) The reliability of sediment load data. In: *Erosion and Sediment Transport Measurement* (Proc. Florence Symp., June 1981), 177-193, IAHS Publ. no. 133.