WATER RESOURCES RESEARCH, VOL. 26, NO. 9, PAGES 2069–2077, SEPTEMBER 1990

Mean Square Error of Regression-Based Constituent Transport Estimates

E. J. GILROY, R. M. HIRSCH, AND T. A. COHN

U.S. Geological Survey, Reston, Virginia

Estimates of long-term transport of constituents commonly are obtained by summing retransformed estimates from regressions of logarithmically transformed response variables. Typical explanatory variables for these regressions include functions of flow, change in flow, time, and time of year. The mean and mean square error of four estimators of long-term transport at periodically measured stations are presented as a function of the observed values of the explanatory variables from the long-term record and summary statistics of the regression data. Estimates of the mean square errors can be used in designing sampling strategies to attempt to minimize the uncertainty in the estimation of long-term transport subject to a constraint on the number of samples to be taken. This uncertainty is expressed in terms of the explanatory variables in the long-term record, the regression coefficients and standard error of the regression and the mean and covariance structure of the explanatory variables used in the regression.

INTRODUCTION

The study of surface water quality problems often involves making estimates of the transport of various constituents past some cross section of a river. This is frequently of interest in the examination of sediment or nutrient budgets for water bodies with long residence times, such as lakes, reservoirs, or estuaries. It is also useful in the evaluation of the fate and transport of a constituent. There may be estimates of the inputs of the constituent to the river basin from point sources, land application, erosional loss, or atmospheric deposition. Comparison of the transport out of the river basin with these estimated inputs can be useful in understanding the relative importance of storage, transformation, or other unexpected losses or gains of the constituent in question.

The specific problem considered in this paper is the accuracy of estimates of constituent transport past a station at which there is a continuous record of streamflow and periodic measurements of concentration. The approach considered here is to use the flow record (or other continuously measured variables that are available and useful) to make estimates, using a regression model, of the load on the individual days of the record and then summing these daily estimates to form an estimate of the total transport for the period of interest. The dependent variable is the logarithm of the transport rate. Daily loads are estimated by exponentiating the regression estimates. The explanatory variables in the regression model typically include one or more of the following: functions of discharge, typically the logarithm; other variables which incorporate the concept of hysteresis [see *Hirsch*, 1988] or the effect of bed forms, or temperature as it affects viscosity and shear stress; variables to incorporate seasonal variations, using trigonometric functions of time of year or categorical variables for seasons; and variables to incorporate long-term trends, using time in years or categorical variables for different multiyear periods.

There are two important issues about such regression models that are not addressed in this paper. The first is model selection: the determination of what variables to include in the model. The second is the problem of estimating the regression using instantaneous data as an explanatory variable and then applying it using a function of the daily average flow as a predictor variable. The first issue is one that is common to regression problems and many of the modern diagnostic tools are probably appropriate for use in making these decisions. The second issue can lead to substantial errors in results if two conditions are met: the variation in flow during the day is large and the relationship between transport and flow is highly nonlinear. This problem can always be overcome by subdividing the day into sufficiently small periods that flow variation within each period is quite small. Unfortunately, much of the available streamflow data is permanently stored as daily averages and the information at finer time steps is not readily available.

In addressing the question of accuracy of the transport estimates, it is necessary to begin with an examination of the accuracy of the individual daily load estimates that are to be summed to make the estimate of the long-term transport. The sampling properties of such individual retransformed regression estimators of load have been discussed recently in the hydrologic literature. *Ferguson* [1986, 1987] and *Koch and Smillie* [1986] have discussed the problems of bias in a retransformed logarithmic estimator. *Cohn et al.* [1989], assuming normally distributed residuals about the logarithmic model, give the exact minimum variance unbiased estimator (MVUE) and its variance, based on the work of *Finney* [1941] and *Bradu and Mundlak* [1970], and discuss the differences among several estimators over a hydrologically reasonable range of conditions. The results given by *Cohn et al.* [1989] and some results due to *Likeš* [1980] are used in this paper to give the variance of the sum of individual retransformed regression estimators of daily loads using the MVUE. For completeness the first two moments of three other estimators considered in the long-term transport problem are presented. The results have significant implications for the problem of determining the design of long-term water quality monitoring programs. These are explored in a simple example.

MATHEMATICAL MODEL AND ESTIMATORS

For many constituents the relationship between concentration (or equivalently load) and flow is taken to be linear in

This paper is not subject to U.S. copyright. Published in 1990 by the American Geophysical Union.

Paper number 90WR00240.

the logarithms [*Gregory and Walling*, 1973]. The model in this paper considers the same logarithmic structure as by *Cohn et al.* [1989]. Let

$$Y(i) = \ln [L(i)] = \beta_0 + \beta_1 X_1(i) + \beta_2 X_2(i) + \cdots + \beta_p X_p(i) + \varepsilon(i) \qquad i = 1, 2, \cdots, M \quad (1)$$

where $L(i)$ is the ith value of the load of the constituent of interest, $X_j(i)$ is the ith concurrent value of the jth explanatory variable, the β_j variables are regression coefficients, the $\varepsilon(i)$ variables are error terms and M is the number of concurrent values of response and explanatory variables. In matrix notation these M values are represented by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2)$$

where $\mathbf{Y}$ is an $M \times 1$ vector of $Y(i)$ variables, $\mathbf{X}$ is an $M \times (p + 1)$ matrix of explanatory variables, the first column being a column of 1s, $\boldsymbol{\beta}$ is a $(p + 1) \times 1$ vector of regression coefficients and $\boldsymbol{\varepsilon}$ is an $M \times 1$ vector of error terms.

Using these M concurrent values of the Y and X variables, ordinary least squares estimates, $\hat{\boldsymbol{\beta}}$, of $\boldsymbol{\beta}$, are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (3)$$

If $\mathbf{X}_i' = (1, X_1(i), \ldots, X_p(i))$ is a vector of explanatory variables at time i then an estimate of $Y(i)$, the logarithm of the load at time i, is given by $\hat{Y}(i) = \mathbf{X}_i'\hat{\boldsymbol{\beta}}$ with the properties of the estimator determined by the properties of the noise term $\boldsymbol{\varepsilon}$ in (2). If $\boldsymbol{\varepsilon}$ is taken to be distributed as multivariate normal with mean 0 and covariance matrix $\sigma_\varepsilon^2 I$, where I is the $M \times M$ identity matrix, then $\hat{Y}(i)$ is normal with mean $\mathbf{X}_i'\boldsymbol{\beta}$ and variance $V_{ii}\sigma_\varepsilon^2$, where

$$V_{ii} = X_i'(X'X)^{-1}X_i$$

and σ_ε^2 can be estimated by $s^2 = [1/(M - p - 1)] \sum_{i=1}^{M} e^2(i)$ as in ordinary least squares where $e(i) = Y(i) - \hat{Y}(i)$. The diagonality of the covariance matrix of ε is a reasonable assumption in light of the fact that for most periodic stations fewer than 12 measurements a year are made. It can be argued that approximating the near zero off-diagonal terms of Var (ε) with zeros will give acceptable results. The properties of the estimation error term $e(i)$ are completely determined by these normality and covariance structure assumptions. However, the estimation error of interest is the error in estimating the load, $L(i)$, the antilogarithm of $Y(i)$. *Cohn et al.* [1989] consider three estimators of $L(i)$. The "rating curve" estimator is simply the exponentiation of $\hat{Y}(i)$, denoted by $\hat{L}_{RC}(i) = \exp(\hat{Y}(i))$. As is well known, see references in the work by *Cohn et al.* [1989], this rating curve estimator is biased under the assumption of normally distributed regression residuals. The question arises as to the existence of a bias correction factor that can be applied to the rating curve estimator. Each of the estimators considered in this paper can be written as a product of the rating curve estimator and some bias correction factor. Under the assumed error distribution each of the bias correction factors is statistically independent of the rating curve estimator.

The quasi-maximum likelihood estimator (QMLE) as proposed by Ferguson is given by multiplying the "rating curve" estimator by the correction factor $\exp(s^2/2)$ and is denoted by

$$\hat{L}_{QMLE}(i) = \hat{L}_{RC}(i) \exp\left(\frac{s^2}{2}\right) \quad (4)$$

If in place of s^2 one used $\hat{\sigma}^2 = ((M - p - 1)/M)s^2$, then this estimator would be the maximum likelihood estimator.

The rationale for using this bias correction factor stems from the assumption of normality of the regression residuals. *Cohn et al.* [1989] show how the moments of $\hat{L}_{RC}$ and $\hat{L}_{QMLE}$ can be written in terms of the moment generating functions of the normal and chi-square distribution functions and thus show that both these estimators are biased. As an alternative estimator they present the minimum variance unbiased estimator [*Finney*, 1941; *Bradu and Mundlak*, 1970] $\hat{L}_{MVUE}(i)$. The MVUE estimator is given by

$$\hat{L}_{MVUE}(i) = \hat{L}_{RC}(i) g_m\left(\frac{m + 1}{2m}(1 - V_{ii})s^2\right) \quad (5)$$

where

$$g_m(t) = \sum_{k=0}^{\infty} \frac{m^k(m + 2k)}{m(m + 2)\cdots(m + 2k)} \left(\frac{m}{m + 1}\right)^k \left(\frac{t^k}{k!}\right) \quad (6)$$

and $m = M - p - 1$.

A fourth estimator not considered by *Cohn et al.* [1989] is the smearing estimator of *Duan et al.* [1982] and *Duan* [1983]. The bias correction factor that yields the smearing estimator is the average of the exponentiated log regression residuals, $e(j)$, so that the smearing estimator is given by

$$\hat{L}_{SM}(i) = \hat{L}_{RC}(i) \frac{1}{M} \sum_{j=1}^{M} \exp(e(j)) \quad (7)$$

The properties of the first three estimators for estimating individual load events are thoroughly discussed in *Cohn et al.* [1989]. The properties of the smearing estimator are given by *Duan et al.* [1982] and will be summarized below. Interest most often centers on estimating sums of such load events rather than individual events. This is discussed in the following section.

Mean Square Error of Sums of Estimated Load Events

Let $\mathbf{X}_i$, for $i = M + 1, M + 2, \cdots, M + N$ be N values of the explanatory variables for which the N corresponding values of the constituent loads, $L(i)$ are unavailable. One quantity associated with the explanatory variables $\mathbf{X}_i$ for $i = M + 1, M + 2, \cdots, M + N$ is the total load, LTOT(θ), defined as

$$\text{LTOT}(\theta) = \sum_{i=M+1}^{M+N} \hat{L}_\theta(i) \quad (8)$$

where θ denotes one of the estimators (RC, QMLE, SM, or MVUE). The mean value, $\mu(\theta)$, of LTOT(θ), is easily obtained from the results given by *Cohn et al.* [1989] for θ = RC, QMLE, or MVUE and from *Duan et al.* [1982] for θ = SM. Only μ(MVUE) is equal to the mean value of the sum of the actual loads. Determination of the mean square error, MSE(θ), of LTOT(θ) necessitates evaluating the expected value of terms of the form $\hat{L}_\theta(i)\hat{L}_\theta(j)$ which can be obtained

directly from the results of *Cohn et al.* [1989] and *Duan et al.* [1982] for θ = SM, for $i = j$. For $i \neq j$ the expectation can be obtained from *Cohn et al.* [1989] for the RC or QMLE estimators. For the MVUE the required expectation follows from a result due to *Likeš* [1980] which gives the expected value of the product of two $g_m(t)$ functions in which $t_i = a_i W$ and $t_j = a_j W$ where W is a chi-square random variable and the a variables are real numbers.

For each of the estimators, let $\hat{L}_\theta(j) = \hat{L}_{RC}(j)$ BCF(θ) where BCF(θ) is the bias correction factor for estimator θ. BCF(RC) = 1 and BCF(MVUE) = BCF(MVUE, j); i.e., the MVUE bias correction factor changes with j because of a dependence on the value of the explanatory variable.

Then from *Cohn et al.* [1989] the mean values are

$$\mu(\mathrm{RC}) = \sum_{i=M+1}^{M+N} \exp(\mathbf{X}_i\boldsymbol{\beta} + V_{ii}\sigma_\epsilon^2/2) = \sum_{i=M+1}^{M+N} \mu(\mathrm{RC}, i) \tag{9}$$

$$\mu(\mathrm{QMLE}) = \mu(\mathrm{RC})\, E(\mathrm{BCF(QMLE)}) \tag{10}$$

$$\mu(\mathrm{SM}) = \mu(\mathrm{RC})\, E(\mathrm{BCF(SM)}) \tag{11}$$

$$\mu(\mathrm{MVUE}) = \sum_{i=M+1}^{M+N} \exp(\mathbf{X}_i\boldsymbol{\beta} + \sigma_\epsilon^2/2) = \sum_{i=M+1}^{M+N} \mu(\mathrm{RC}, i)\, E(\mathrm{BCF(MVUE}, i)) \tag{12}$$

where

$$E(\mathrm{BCF(QMLE)}) = \left(1 - \frac{\sigma_\epsilon^2}{m}\right)^{-m/2} \tag{13}$$

$$E(\mathrm{BCF(MVUE}, i)) = \exp((1 - V_{ii})\sigma_\epsilon^2/2) \tag{14}$$

and from *Duan et al.* [1982]

$$E(\mathrm{BCF(SM)}) = \frac{1}{M}\sum_{j=1}^{M} \exp((1 - V_{jj})\sigma_\epsilon^2/2) \tag{15}$$

Note that the MVUE bias correction factor is the only correction factor that depends on the values of the explanatory variables used in the estimate.

The expected value of the sum of the $L(i)$ values is just μ(MVUE). The mean square error of LTOT(θ) about μ(MVUE) for any θ is given by

$$\mathrm{MSE}(\theta) = \sum_{i=M+1}^{M+N}\sum_{j=M+1}^{M+N} \mathrm{Cov}(i, j, \theta) + (B(\theta))^2 \tag{16}$$

where Cov (i, j, θ) = Cov $(\hat{L}_\theta(i), \hat{L}_\theta(j))$ and $B(\theta) = \mu(\theta) - \mu$(MVUE).

The covariance terms are obtained from the following expected values:

$$E(\hat{L}_{RC}(i)\, \hat{L}_{RC}(j)) = \mu_i\mu_j \exp\left(\frac{(V_{ii} + V_{jj} + 2V_{ij})\sigma_\epsilon^2}{2}\right) \tag{17}$$

$$E(\hat{L}_{QMLE}(i)\, \hat{L}_{QMLE}(j)) = E(\hat{L}_{RC}(i)\, \hat{L}_{RC}(j))\, E(\mathrm{BCF^2(QMLE)}) \tag{18}$$

$$E(\hat{L}_{SM}(i)\, \hat{L}_{SM}(j)) = E(\hat{L}_{RC}(i)\, \hat{L}_{RC}(j))\, E(\mathrm{BCF^2(SM)}) \tag{19}$$

$$E(\hat{L}_{MVUE}(i)\, \hat{L}_{MVUE}(j)) = E(\hat{L}_{RC}(i)\, \hat{L}_{RC}(j))\, E(\mathrm{BCF(MVUE}, i)\, \mathrm{BCF(MVUE}, j)) \tag{20}$$

where

$$\mu_i = \exp(\mathbf{X}_i\boldsymbol{\beta}) \tag{21}$$

$$V_{ij} = \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_j \tag{22}$$

$$E(\mathrm{BCF^2(QMLE)}) = \left(1 - \frac{2\sigma_\epsilon^2}{m}\right)^{-m/2} \tag{23}$$

$$E(\mathrm{BCF^2(SM)}) = \frac{1}{M^2}\left(\sum_{k=1}^{M} \exp(2(1 - V_{kk})\sigma_\epsilon^2) + 2\sum_{l=1}^{M-1}\sum_{k=l+1}^{M} \exp((2 - V_{kk} - V_{ll} - 2V_{kl})\sigma_\epsilon^2/2)\right) \tag{24}$$

$$E(\mathrm{BCF(MVUE}, i)\, \mathrm{BCF(MVUE}, j)) = \exp\left(\frac{(2 - V_{ii} - V_{jj})\sigma_\epsilon^2}{2}\right) \cdot g_m\left(\frac{(1 - V_{ii})(1 - V_{jj})\sigma_\epsilon^4(m + 1)}{2m^2}\right) \tag{25}$$

Equations (9)–(25) are sufficient to compute the biases and mean square errors of the four estimators. All the elements in these equations are either known or estimable from the data.

APPLICATIONS

The utility of these formulas for mean square error is twofold. The first application arises in estimating, after the fact, the MSE of a particular estimate of total load. The second application is in the development of sampling designs to achieve a low MSE at a given sampling cost or conversely a low sampling cost for a given MSE.

In either case the application of the formulas requires the assumption of a particular model form and values of the coefficients ($\boldsymbol{\beta}$ and σ^2). This paper does not consider the error in estimates of the MSE that are the result of using fitted model forms and coefficients rather than the true, but unknown, model form and coefficients. However, it should be recognized that in cases where substantial amounts of extrapolation are used (i.e., where the regression model is applied to data beyond the limits of the X data used for calibration) the resulting MSE estimates can be severely biased. The formulas developed in this paper all assume that the model form is the correct one and that all of the error is due to sampling error and not lack of fit. In practice this could lead to serious errors. For example, if the data used in fitting the model are all from days of low to moderate

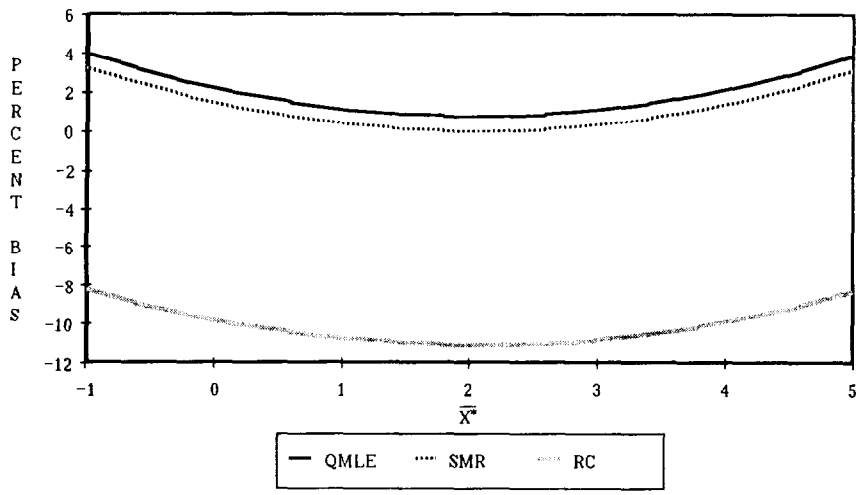

Fig. 1. Bias, in percent of the mean of the true total load, of the RC, QMLE and SM estimators of total load as a function of $\bar{X}^*$ for $M = 36$, $N = 1000$, $\beta = 2.0$, $S_x^* = 1.0$ and $\sigma_\varepsilon = 0.5$.

discharge, the estimates made at high discharge will be much more uncertain than the formulas would indicate. This is because the formulas assume that only the intercept and slope of the line are unknown whereas, in general, the functional shape must also be viewed as unknown. Days of high discharge typically contribute a large fraction of the load and hence a large fraction of the MSE, and thus when substantial amounts of extrapolation are involved, the formulas presented here should only be viewed as an approximation to the MSE of total load.

The quantities necessary for computing the MSE after the fact are the following: the σ_ε, β coefficients, the means and covariances of the M values of X in the calibration data set, and the N actual values of X for which the estimates are made. To simplify the presentation, it is useful to make some assumptions. First we will assume that X is univariate (for example, X is the logarithm of discharge). By making this assumption the characteristics of the M sample calibration data set can be described by the statistics $\bar{X}$ and S_X: the sample mean and standard deviation of the calibration X values. Another simplification can be made by assuming that N is very large in comparison to M (a realistic assumption), and that the estimate of total load that is being made excludes the known values of load for the M days on which measurements exist. This assumption makes the MSE independent of the particular values of X in the calibration data set (it depends only on the two sample moments $\bar{X}$ and S_X^2).

Furthermore, if N is much larger than M, the MSE, when expressed as a percentage of the expected value of total load, is insensitive to the actual value of N and insensitive to the specific values of X that exist in the sample given that the distribution for the X data is fixed. Thus one can select a large N (1000 was used here), and a distribution of X (normal was used here), and take a random sample of 1000 observations from that distribution and use them as the data set for which estimates are to be made. Selecting a different random sample of 1000, or more, observations will have only a very slight effect on the computed MSE.

One further simplification, to aid in presenting the results, is to express the moments of the M calibration values of X in terms of the moments of the N values of X used to make the estimates. If the sample mean and standard deviation of these N values are μ_x and σ_x, respectively, then the stan-

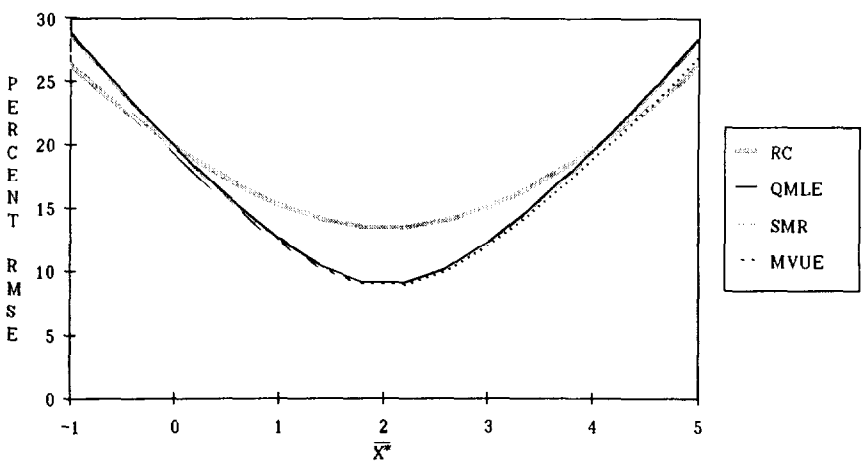

Fig. 2. RMSE, in percent of the mean of the true total load, of the RC, QMLE and SM estimators of total load as a function of $\bar{X}^*$ for $M = 36$, $N = 1000$, $\beta = 2.0$, $S_x^* = 1.0$ and $\sigma_\varepsilon = 0.5$.

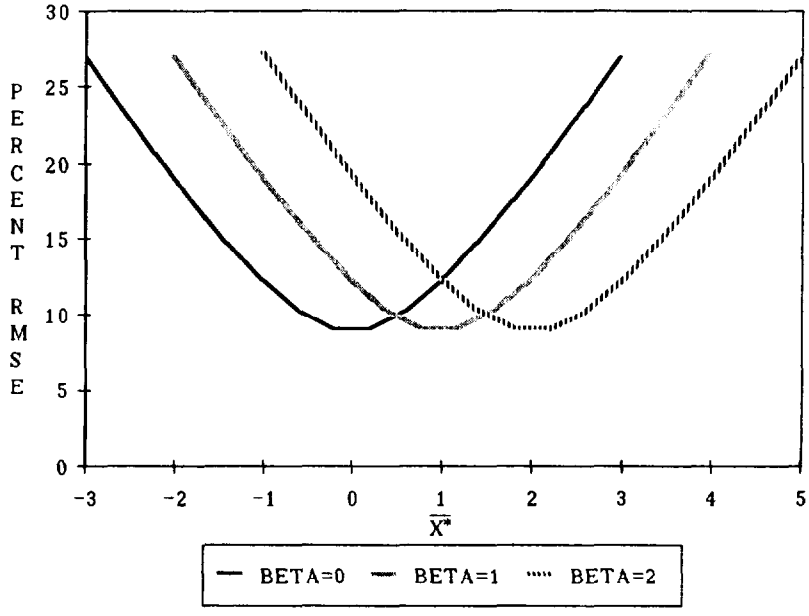

Fig. 3. RMSE, in percent of the mean of the true total load, of the MVUE estimator of total load as a function of $\bar{X}^*$ for $M = 36$, $N = 1000$, $\sigma_\varepsilon = 0.5$, $S_x^* = 1.0$ and $\beta = 0, 1, 2$.

dardized statistics to characterize the M calibration values are

$$\bar{X}^* = (\bar{X} - \mu_x)/\sigma_x^2$$

$S_x^* = S_x/\sigma_x$. It should be noted that the results given in the next section based on these simplifying assumptions do not depend on the distribution of the M calibration values (the results depend only on their moments) but the results do depend on the distributional characteristics of the N estimation values.

The other application of the formulas mentioned above is for the design of sampling strategies. However, because the X data arise from a stochastic process, it is never possible to design the sampling plan in the usual sense of the word "design." What is possible is to specify an algorithm to be used to make decisions about taking a sample at a specific time. The difficulties faced in developing a good algorithm are the need to avoid exceeding budgets by taking too many samples, the need to avoid taking too few samples because the flow conditions which occur do not lead to many decisions to sample, the uncertainty as to the flow conditions that will actually exist at the time the sample will be taken, and logistical difficulties with limited manpower being called upon to take many samples in a short period of time at an entire network of stations. Examination of suitable algorithms is an important research topic but is beyond the scope of this paper. However, the computations described below are potentially useful in that they identify the relative importance of the three key elements in the design of the calibration sample: the central tendency ($\bar{X}^*$), the variability (S_x^*), and the sample size (M). In addition to demonstrating the relatively obvious advantages of having large values of S_x^* and M, these examples illustrate the influence of $\bar{X}^*$ and the idea that the ideal value of $\bar{X}^*$ is a function of the model coefficient β_1.

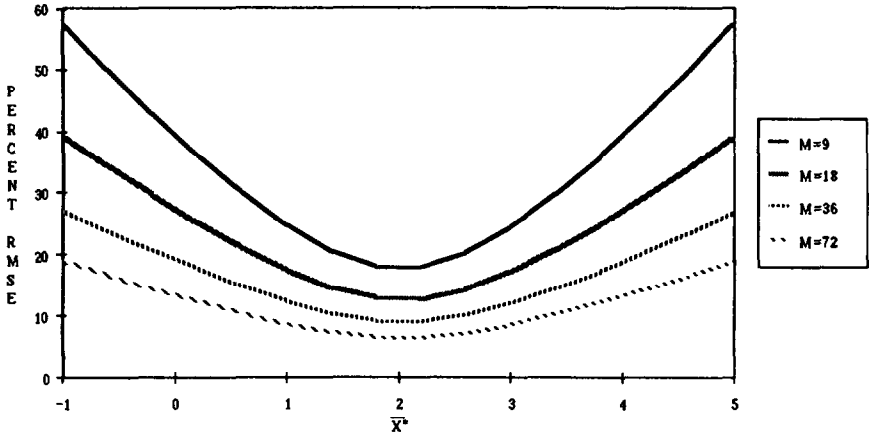

Fig. 4. RMSE, in percent of the mean of the true total load, of the MVUE estimator of total load as a function of $\bar{X}^*$ for $N = 1000$, $\beta = 2.0$ and $\sigma_\varepsilon = 0.5$, $S_x^* = 1.0$ and $M = 9, 18, 36, 72$.

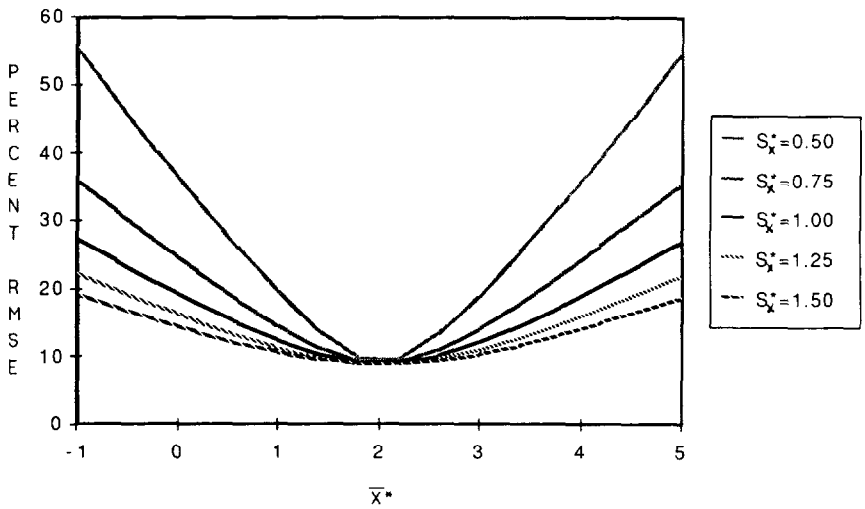

Fig. 5. RMSE, in percent of the mean of the true total load, of the MVUE estimator of total load as a function of $\bar{X}^*$ for $M = 36$, $N = 1000$, $\beta = 2.0$ and $\sigma_\varepsilon = 0.5$, $S_x^* = 0.50, 0.75, 1.00, 1.25$ and 1.50.

EXAMPLES

The examples described in this section provide some insight into the characteristics of the estimators and the way that the error of transport estimates may be influenced by changes in the makeup of the calibration data set. As mentioned above, for these examples it is assumed that only one explanatory variable is used. In practice, the one explanatory variable might be the logarithm of discharge. In generalizing these results to the case with multiple explanatory variables the errors would be a function of a multivariate measure of distance between the calibration data and the full data set for the period of record.

Figures 1 and 2 show the bias and root mean square error (RMSE) in percent, respectively, for the four estimators: RC, QMLE, MVUE, and SM for $M = 36$, $N = 1000$, $\beta_1 = 2.0$, $\sigma_\varepsilon = 0.5$ as a function of $\bar{X}^*$. The results shown in Figures 1 and 2 are computed in the following fashion. A single sample of N (=1000) observations is drawn from a standard normal distribution. They are subsequently rescaled to have a mean of exactly zero ($\mu_x = 0$) and a standard deviation of exactly one ($\sigma_x = 1$). Then the mean of the calibration data set ($\bar{X}$) is set to an arbitrary value (in the range -1.0 to $+5.0$) and the standard deviation of the calibration data set (S_x) is set to 1. Having fixed M, N, β_1, σ_ε^2, the actual values of the 1000 observations of X for which there are no Y data, and the first two moments of the 36 calibration X data, the bias $B(\theta)$ can be computed for each of the four methods.

Note that if the explanatory variable is the logarithm of discharge, then $\beta_1 = 2$ implies that load is proportional to the square of discharge, which further implies that concentration is proportional to discharge. The standard error in log base e units of the regression in Figures 1 and 2 is 0.5. Expressed as standard error in percent, this is a standard error of 53% = $100 * (\exp(0.5^2) - 1)^{1/2}$.

The MVUE has zero bias, regardless of the value of $\bar{X}^*$ and so is not shown in Figure 1. As in the case of individual load estimates, the rating curve estimator of long-term transport is (generally) biased downward while the QMLE estimator overcompensates. The nonparametric smearing estimator is less biased than the QMLE estimator in this case. Note that the bias in each of the cases shown has a minimum at approximately $\bar{X}^* = 2.0$.

Figure 2 shows that the QMLE and SM estimators are comparable with respect to root mean square error and both do almost as well as the MVUE estimator. The RC estimator is the poorest performer over most of the range of $\bar{X}^*$. Only in the extremities of the $\bar{X}^*$ range does the RC estimator approach the root mean square errors of the other three estimators. Note in Figure 2 that the minimum root mean square error seems to occur near $\bar{X}^* = 2$. Figure 3 shows the root mean square error curves for the MVUE estimator for $\beta_1 = 0, 1, 2$. In all three cases the minimum value of the curves occur near $\bar{X}^* = \beta_1$. On the basis of an approximation to the sum of the N estimated values of load, this optimal value of $\bar{X}^*$ can be shown to be equal to β_1, the slope of the regression line, for the case $p = 1$. A presentation of this result, along with the approximation to the RMSE of the QMLE, SM and MVUE estimators is given in the appendix.

Figures 4 and 5 show the root mean square error of the MVUE estimator as a function of $\bar{X}^*$, M and the parameter, S_x^* which is the ratio of the standard deviation of the logs of the flows used in calibrating the regression to the standard deviation of the logs of the N flows used in the estimation of total load. Figure 4 shows (1) the decrease in RMSE as M increases from 9 to 72 and S_x^* stays constant at unity and (2) the increase in RMSE as $\bar{X}^*$ moves away from the optimal value $\bar{X}^* \approx 2$. Figure 5 shows the decrease in the RMSE function as S_x^* increases while the other parameters of the model stay fixed at $M = 36$, $N = 1000$, $\beta = 2.0$ and $\sigma_\varepsilon = 0.5$. The more widely the flows used in the calibration of the regression equation are spread relative to the remaining flows of interest, the better the estimate of total load will be. For $S_x^* \geq 1.0$ the RMSE of the MVUE is relatively flat with respect to $\bar{X}^*$.

CONCLUSIONS

Under the assumption of normality and independence of the residuals of the log linear multiple regression model, exact formulations of the first two moments of four reasonable estimators of the total load are presented. Both the bias

and root mean square error of each of the four estimators are functions of M, the calibration sample size, p, the number of explanatory variables, N, the number of additional values of the p explanatory variables not in the calibration set, σ_ε^2, the mean square error of the regression, β, the regression coefficients and the mean and covariances of the explanatory data used in the calibration set.

As in the case of individual load estimates the QMLE estimator overcompensates for the bias in the RC estimator. Interestingly, the nonparametric smearing estimator does a better job of correcting for the bias than does the QMLE for the case shown here. Of the three biased estimators, RC, QMLE and SM, the QMLE and SM have RMSEs comparable to the MVUE while the RC estimator has greater RMSE over a reasonable range of $\bar{X}^*$, a standardized measure of how far the calibration data set is from the remainder of the data set of interest. For fixed values of N and β the RMSE of the MVUE is a concave upward function of $\bar{X}^*$, with a minimum occurring near $\bar{X}^* = \beta_1$ in the case of $p = 1$. Increasing the variability (S_x^*) and or the size (M) of the calibration data set lowers the RMSE and makes the RMSE less sensitive to $\bar{X}^*$.

The results presented in this paper provide the means to estimate the RMSE of a given estimate of total load based on a log linear regression at a periodic station. However, these estimates are predicated on assumptions as to the true form and parameters of the regression relationship which are usually not known in practice. Nevertheless, the results can be useful in the process of designing or improving stream sampling programs for estimating transport.

APPENDIX

The mean square error (MSE) of the three total load estimators, QMLE, SM and MVUE, are quite similar over the range of interest if N, the number of individual loads estimated, is large. Considered as functions of $\bar{X}$, these three MSEs all appear to have a minimum near the value $\bar{X} = \mu_x + \beta_1\sigma_x^2$. The exact MSEs are complicated double sums for which derivations of extrema are not easily performed. However, a certain simplification of the MSEs yields a close approximation to the actual MSEs and at the same time gives a mathematically tractable form.

For the case $p = 1$ and the logarithm of flow being the explanatory variable, a good approximation to the MSE of total load, under the assumption of normally distributed logarithms of flows, can be obtained as follows.

As above, let

$$\text{LTOT(QMLE)} = \sum_{i=M+1}^{M+N} \exp\left(\bar{Y} + b_1(X_i - \bar{X})\right) * \exp\left(\frac{s^2}{2}\right)$$

$$= \exp\left(\frac{s^2}{2}\right) * \exp(\bar{Y}) \sum_{i=M+1}^{M+N} \exp\left(b_1(X_i - \bar{X})\right) \quad \text{(A1)}$$

Then if N is large, this last summation can be approximated by an integral with appropriate weights assigned to the relative occurrences of the X_i variables not used in the calibration of the regression. Remembering that the X_i variables correspond to the logarithms of flow, it is not unreasonable to take these weights to be proportional to the relative frequency of occurrence of a normal random variable. Hence the summation can be approximated by N times the expected value of a lognormal random variable. The terms b_1 and $\bar{X}$ are constant with respect to this approximating expectation. If the distribution of the X_i variables is taken to be normal with mean μ_x and variance σ_x^2 then $b_1(X_i - \bar{X})$ is normal with mean $b_1(\mu_x - \bar{X})$ and variance $b_1^2\sigma_x^2$. Hence $\exp(b_1(X_i - \bar{X}))$ is lognormal and the summation is approximated by $N \exp(b_1(\mu_x - \bar{X}) + (b_1^2\sigma_x^2/2))$. The total load is then approximated by

$$\text{LTOT}(\theta) \approx \bar{L} = N \exp\left(\frac{s^2}{2} + \bar{Y}\right) \cdot \exp\left(b_1(\mu_x - \bar{X}) + \frac{b_1^2\sigma_x^2}{2}\right) \quad \text{(A2)}$$

where θ can be QMLE, SM or MVUE because N is large.

Under the assumptions of independent, normally distributed ε variables for the calibration data set, the random variables $\bar{Y}$, s^2 and b_1 are distributed independently of one another with the following distributions:

$$\bar{Y} : N\left(\beta_0 + \beta_1\bar{X}, \frac{\sigma_\varepsilon^2}{M}\right) \quad \text{(A3)}$$

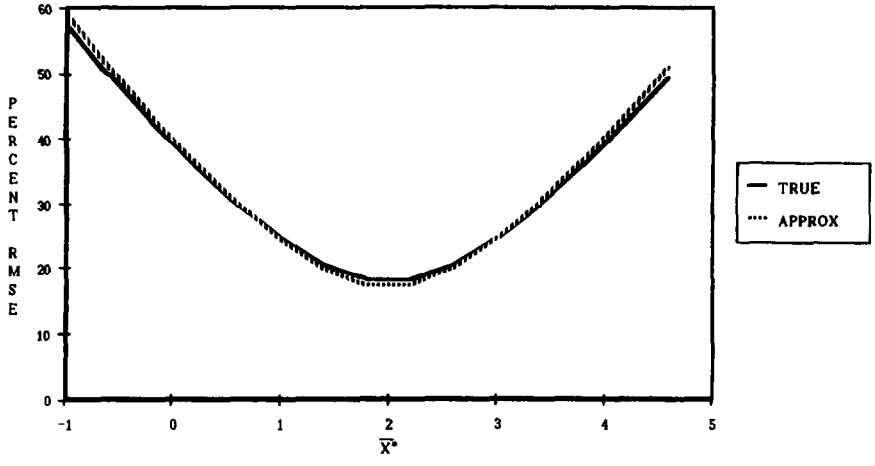

Fig. A1. Comparison of approximation and true RMSE for $M = 9$, $N = 1000$, $\beta = 2.0$, $\sigma_\varepsilon = 0.5$ and $S^* = 1.0$.

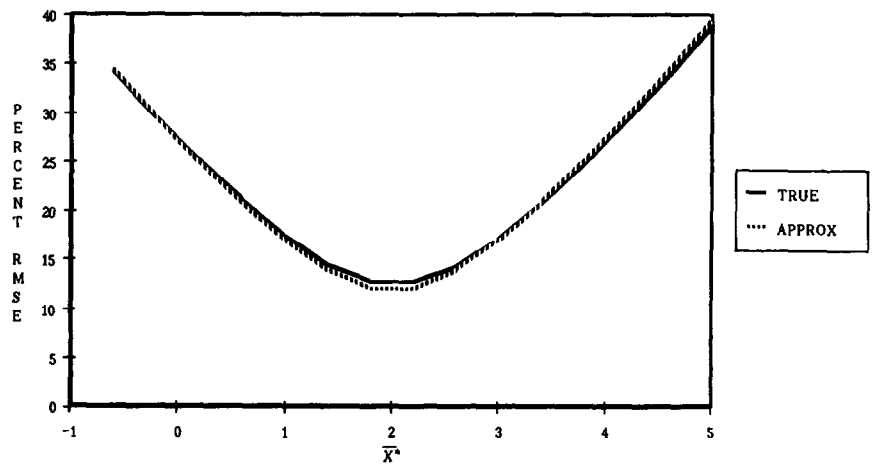

Fig. A2. Comparison of approximation and true RMSE for $M = 18$, $N = 1000$, $\beta = 2.0$, $\sigma_\varepsilon = 0.5$ and $S^* = 1.0$.

$$\frac{ms^2}{\sigma_\varepsilon^2} : \chi_m^2 \tag{A4}$$

$$b_1 : N(\beta_1, \sigma^2(b_1)) \tag{A5}$$

where $\sigma^2(b_1)$ is the sampling variance of the slope in simple linear regression.

Hence all the moments of $\bar{L}$ are easily obtained. If the term in $\bar{L}$ containing b_1^2 is reexpressed by completing the square, the form associated with a noncentral χ^2 distribution with one degree of freedom is easily recognized. $\bar{L}$ can be rewritten as

$$\bar{L} = N \exp\left(\frac{-(\mu_x - \bar{X})^2}{2\sigma_x^2}\right) \exp\left(\frac{s^2}{2}\right) \exp(\bar{Y}) \cdot \exp\left(\frac{(\sigma_x b_1 + (\mu_x - \bar{X})/\sigma_x)^2}{2}\right) \tag{A6}$$

To see that the argument of the exponential in the last term on the right in this last expression is distributed as a constant, φ, times a noncentral χ^2 with noncentrality parameter, λ, note that b_1 is distributed as

$$\sigma(b_1)U + \beta_1$$

where U is distributed as $N(0, 1)$. The moment generating function of a noncentral χ^2 is given by *Johnson and Kotz* [1970, p. 134]. The constants φ and λ are given by

$$\varphi = \sigma_x^2\sigma^2(b_1)/2 \qquad \lambda = \left(\beta_1 + \frac{(\mu_x - \bar{X})}{\sigma_x^2}\right)^2 / \sigma^2(b_1) \tag{A7}$$

Using these results the kth moment of $\bar{L}$ is given by

$$E(\bar{L}^k) = N^k \exp\left(\frac{-k(\mu_x - \bar{X})^2}{2\sigma_x^2}\right) \cdot \left(1 - \frac{k\sigma_\varepsilon^2}{m}\right)^{-m/2} \exp\left(k(\beta_0 + \beta_1\bar{X}) + \frac{k^2\sigma_\varepsilon^2}{2M}\right) \cdot (1 - 2k\varphi)^{-1/2} \exp\left(\frac{\lambda k\varphi}{1 - 2k\varphi}\right) \tag{A8}$$

Note that the following two inequalities must be satisfied

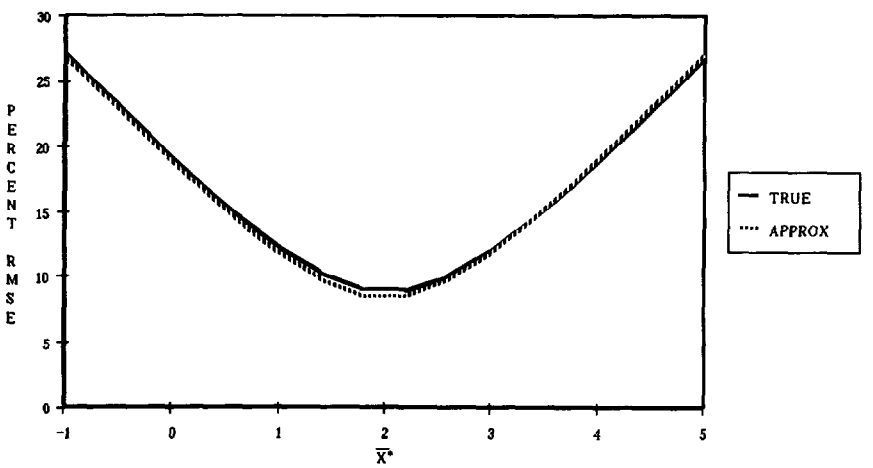

Fig. A3. Comparison of approximation and true RMSE for $M = 36$, $N = 1000$, $\beta = 2.0$, $\sigma_\varepsilon = 0.5$ and $S^* = 1.0$.

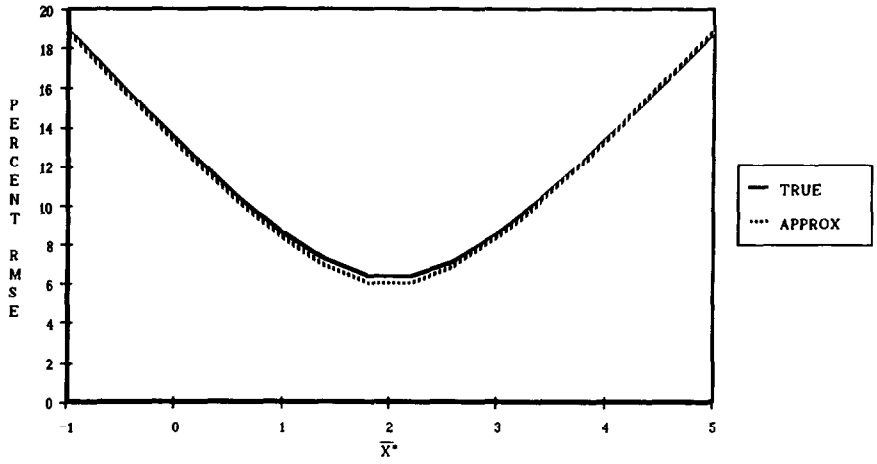

Fig. A4. Comparison of approximation and true RMSE for $M = 72$, $N = 1000$, $\beta = 2.0$, $\sigma_\varepsilon = 0.5$ and $S^* = 1.0$.

$$\frac{k\sigma_\varepsilon^2}{2m} < 1$$

$$2k\varphi < 1$$

Hence for k large enough these moments will fail to exist and the approximation will not be valid.

The mean square error as a fraction of total load is then approximated by

$$\text{MSE} = \frac{E(\bar{L}^2) - E^2(\bar{L})}{E^2(\bar{L})} = \left(1 - \frac{\sigma_\varepsilon^2}{m}\right)^m \left(1 - \frac{2\sigma_\varepsilon^2}{m}\right)^{-m/2} \cdot (1 - 2\varphi)(1 - 4\varphi)^{-1/2} \exp\left(\frac{\sigma_\varepsilon^2}{M} + \frac{4\varphi^2\lambda}{(1 - 4\varphi)(1 - 2\varphi)}\right) - 1$$

The value of $\bar{X}$ that minimizes this approximation to the relative mean square error of the total load estimate is easily found by noting that the only term that depends on $\bar{X}$ is λ. (Note that the normality assumption allows the choice of $\bar{X}$ independently of S_x^2.) This minimizing value, $\bar{X}_{\min}$, of $\bar{X}$ is given by

$$\bar{X}_{\min} = \mu_x + \beta_1\sigma_x^2$$

Figures A1–A4 show how closely the simplified relative MSE approximates the true relative MSE of the MVUE estimator for $N = 1000$, $\beta_1 = 2.0$, $\sigma_\varepsilon = 0.5$, $S_x^* = 1.0$, and $M = 9, 18, 36, 72$. Another approximation can be obtained for the expected value of $\text{LTOT}^2(\theta)$ by first squaring $\text{LTOT}(\theta)$ in (A1) and then approximating the double sum by a double integral. This will result in an approximation to the MSE differing only by an additive term proportional to $1/N$. Because the approximation is obtained for large N and the approximation using (A8) lends itself more easily to differentiation with respect to $\bar{X}$, the simpler approximation is shown here.

REFERENCES

Bradu, D., and Y. Mundlak, Estimation in lognormal linear models, *J. Am. Stat. Assoc.*, *65*(329), 198–211, 1970.

Cohn, T. A., L. L. DeLong, E. J. Gilroy, R. M. Hirsch, and D. Wells, Estimating constituent loads, *Water Resour. Res.*, *25*(5), 937–942, 1989.

Duan, N., Smearing estimate: A nonparametric retransformation method, *J. Am. Stat. Assoc.*, *78*(383), 605–610, 1983.

Duan, N., W. G. Manning, C. N. Morris, and J. P. Newhouse, A Comparison of Alternative Models for Medical Care, *Rep. R-2754-HHS*, The Rand Corporation, Santa Monica, Calif., 1982.

Ferguson, R. I., River loads underestimated by rating curves, *Water Resour. Res.*, *22*(1), 74–76, 1986.

Ferguson, R. I., Accuracy and precision of methods for estimating river loads, *Earth Surf. Processes Landforms*, *12*, 95–104, 1987.

Finney, D. J., On the distribution of a variate whose logarithm is normally distributed, *J. R. Stat. Soc.*, *7*, 155–161, 1941.

Gregory, K. J., and D. E. Walling, *Drainage Basin Form and Process*, John Wiley, New York, 1973.

Hirsch, R. M., Statistical methods and sampling design for estimating step trends in surface-water quality, *Water Resour. Bull.*, *24*(3), 493–503, 1988.

Johnson, N. L., and S. Kotz, *Distributions in Statistics: Continuous Univariate Distributions-2*, 306 pp., John Wiley, New York, 1970.

Koch, R. W., and G. M. Smillie, Bias in hydrologic prediction using log-transformed regression models, *Water Resour. Bull.*, *22*(5), 717–723, 1986.

Likeš, J., Variance of the MVUE for lognormal variance, *Technometrics*, *22*(2), 253–258, 1980.

T. A. Cohn, E. J. Gilroy, and R. M. Hirsch, U.S. Geological Survey, 410 National Center, Mail Stop 410, Reston, VA 22092.

(Received May 1, 1989;
revised December 1, 1989;
accepted January 9, 1990.)