Protocols for the Optimal Measurement and Estimation of Nutrient Loads

Error Approximations

Technical Report

Prof. David R. Fox
Australian Centre for Environmetrics

Report 03/05

April 2005

THE UNIVERSITY OF MELBOURNE
CSIRO
MELBOURNE UNIVERSITY PRIVATE

This page intentionally blank

Acknowledgements

Financial support for this work has been provided under the Victorian Government’s Regional Catchment Investment Plan (Project No. WG-0405-10.45); Queensland EPA and the Department of Environment and Heritage (Coastal Catchment Initiative); the University of Melbourne (Centre for Water and Landscape Management); Southern Rural Water; and Goulburn-Murray Water. Additional support has been provided by the University of Melbourne (Department of Civil and Environmental Engineering and the School of Anthropology, Geography, and Environmental Science), CSIRO Land and Water, and Melbourne University Private.

The author is particularly grateful to Mr. Geoff Gordon of the NSW EPA who made many valuable contributions to the derivation of equation 6 and the material in Appendix A. The contributions of Prof. Barry Hart, Dr. K.S. Tan and Dr. Teri Etchells who reviewed an earlier version of this report are also acknowledged.

Table of Contents

1. Introduction

The accurate estimation of total loads of sediments and nutrients is a problem that is attracting considerable attention among natural resource managers, environmental protection agencies, governments, landowners, and the general community. The delivery of sediments from Queensland catchments has been identified as a threat to the ecosystem of the Great Barrier Reef, while point and diffuse sources of land-based nutrients are implicated in the increased frequency and severity of algal blooms in water bodies around the country. Accordingly, there has been a growing trend towards the expression of aspirational and compliance targets for nutrients and sediments in terms of either a relative or absolute reduction in total *load*. For example, a 20% nutrient reduction target has been imposed on Queensland catchments impacting the Great Barrier Reef while the Victorian EPA has required a 40% reduction in the total phosphorous load from the McAlister Irrigation District by 2005 and a commensurate 40% reduction in total nutrient loads to the Gippsland Lakes by 2022. As noted by Henderson and Bui (2004), the quantification of errors and uncertainty is particularly important in the context of ecological risk assessments as a failure to do so may lead to risks being significantly under or over-estimated.

This report focuses on the quantification of errors associated with a number of common load estimation techniques. We also point out the duality between simple mean-based load estimators and ratio estimation techniques.

2. Load Estimation

A list of some 24 computational techniques for estimating a load was provided in Letcher et al. (2002). Most of these formulae can be classified as belonging to one of the groupings: mean-based estimators; ratio estimators; and regression estimators. In this paper we consider a class of load estimators given by equation 1.

$$\hat{L} = K\left(\sum_{i=1}^{n_c} w_i c_i\right)\left(\sum_{j=1}^{n_q} v_j q_j\right) \tag{1}$$

where c_i is a measured concentration on the i^{th} occasion; q_j is a measured flow on the j^{th} occasion and w_i and v_j are weights[1]. *K* is a constant that reconciles the sampling time-step with the period of interest (eg. if concentrations and flows represent daily values and an annual load estimate is required, then *K=365*).

3. Theoretical mean and variance

Before turning our attention to the properties of load estimators, it will be useful to develop some theoretical results for the expected value and variance of a load under certain distributional assumptions. In what follows we assume (not unreasonably), that the distribution of concentration (C) and flow (Q) are well described by the bivariate lognormal distribution given by equation 2 and that load, $L = CQ$.

$$f_{C,Q}(c,q) =$$

$$\frac{1}{2\pi cq\sigma_C\sigma_Q\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{\ln(c)-\mu_C}{\sigma_C}\right)^2-2\rho\left(\frac{\ln(c)-\mu_Q}{\sigma_C}\right)\left(\frac{\ln(q)-\mu_Q}{\sigma_Q}\right)+\left(\frac{\ln(q)-\mu_{2Q}}{\sigma_Q}\right)^2\right]\right\}$$

(2)

where μ and σ are the mean and standard deviation of the log-transformed data and ρ is the correlation between *log* concentration and *log* flow.
Fox (2004) showed that the expected load is given by equation 3.

$$E[L] =$$

$$\exp\left\{\left(\mu_C+\mu_Q\right)+\frac{1}{2\left(1-\rho^2\right)}\left[\left(\rho\sigma_Q+\sigma_C\right)^2+\left(\rho\sigma_C+\sigma_Q\right)^2-2\rho\left(\rho\sigma_Q+\sigma_C\right)\left(\rho\sigma_C+\sigma_Q\right)\right]\right\}$$

(3)

Furthermore, it can be established that the second (uncorrected) moment is:

[1] The weights are somewhat arbitrary although values are usually determined by the nature of the sampling scheme. For example, a constant weight of 1/n implies a simple average while flow weighted averaging implies weights are determined on the basis of observed flow (higher flow implying higher weight).

$$E\left[L^2\right]=$$
$$\exp\left\{2\left(\mu_C+\mu_Q\right)+\frac{2}{\left(1-\rho^2\right)}\left[\left(\rho\sigma_Q+\sigma_C\right)^2+\left(\rho\sigma_C+\sigma_Q\right)^2-2\rho\left(\rho\sigma_Q+\sigma_C\right)\left(\rho\sigma_C+\sigma_Q\right)\right]\right\} \tag{4}$$

and so the variance is given as

$$Var\left[L\right]=E\left[L^2\right]-\left(E\left[L\right]\right)^2 \tag{5}$$

4. Uncertainty in load estimates

We next turn our attention to sampling properties of the estimator given by equation 1. In particular, it can be shown that an approximation[2] to the variance is:

$$Var\left[\hat{L}\right]=K^2\left(\sum_{i=1}^{n_c}w_i^2\right)\left(\sum_{j=1}^{n_q}v_j^2\right)Var\left[L\right] \tag{6}$$

For suitable choices of the weights w_i and v_j we can obtain variance approximations for a number of common load estimators. Furthermore, the duality between a ratio estimator of load and one obtained using flow-weighted mean concentrations can be established. These issues are covered under special cases 1-3 below.

Special Case #1 – The Naïve estimator (average flow *x* average concentration)

The simplest of all load estimators is a scaled product of the mean concentration and the mean discharge (flow). We refer to this as the 'naïve' estimator – its attractiveness lies in its computational simplicity, although serious biases (typically > 30%) result

[2] It is recognised that this approximation does not take into account autocorrelation between the *c* and *q* data, nor the cross-correlations between them. See Appendix A for a derivation.

(Fox, 2004). The naïve estimator is readily seen to be obtained by letting $w_i = \frac{1}{n_c}$ and $v_j = \frac{1}{n_q}$ giving

$$\hat{L}_1 = K\bar{C}\,\bar{Q} \tag{7}$$

and

$$Var\left[\hat{L}_1\right] = \frac{K^2 Var\left[L\right]}{n_c\, n_q} \tag{8}$$

Special Case #2 – Load estimator using flow-weighted mean concentrations and unknown total discharge

Unlike the naïve estimator which assigns equal weight to each observed concentration, the flow-weighted mean concentration (*fwmc*) uses weights that are proportional to the magnitude of the associated flow. In this sense, the naïve estimator may be thought of as a time-based average whereas the *fwmc* is a flow-based average. It is implicit in flow-weighted averaging that the flow and concentration data are contemporaneous whereas no such assumption was previously made. Thus, $n_c = n_q = n$ and the weights for *fwmc* are

$$w_i = \frac{q_i}{\sum_{i=1}^{n} q_i}; \quad v_i = 1 \quad \forall\, i$$

Thus,

$$\hat{L}_2 = K' \frac{1}{\sum_{i=1}^{n} q_i} \left(\sum_{i=1}^{n} c_i\, q_i \right) \left(\sum_{i=1}^{n} q_i \right)$$

$$\hat{L}_2 = K' \sum_{i=1}^{n} c_i\, q_i \tag{9}$$

where $K' = \frac{K}{n}$ (eg. if one month of daily concentration data are available to estimate an annual load using equation 9, then *K=365* and *n=30*). The K' factor is needed in this case because the total discharge, $\sum_{i=1}^{n} q_i$ is only known for the *sample* and not the entire period of interest.

Furthermore,

$$Var\left[\hat{L}_2\right] = K'^2 \sum_{i=1}^{n} \left(\frac{q_i}{\sum q_i^2} \right) n \, Var\left[L\right]$$

$$= \frac{K^2 Var\left[L\right]}{n\left(\sum_{i=1}^{n} q_i\right)^2} \sum_{i=1}^{n} q_i^2 \qquad (10)$$

<u>Aside</u>
It can be readily established that in the case $n_c = n_q = n$, the variance of $\hat{L}_2$ is greater than the variance of $\hat{L}_1$. To see this, we look at $Var\left[\hat{L}_2\right] - Var\left[\hat{L}_1\right]$.

$$Var\left[\hat{L}_2\right] - Var\left[\hat{L}_1\right] = \frac{K^2}{n\left(\sum_{i=1}^{n} q_i\right)^2} \sum_{i=1}^{n} q_i^2 \, Var\left[L\right] - \left\{ \frac{K^2 Var\left[L\right]}{n^2} \right\}$$

$$= K^2 Var\left[L\right] \left\{ \frac{\sum_{i=1}^{n} q_i^2}{n\left(\sum_{i=1}^{n} q_i\right)^2} - \frac{1}{n^2} \right\}$$

Hence, $Var\left[\hat{L}_2\right] > Var\left[\hat{L}_1\right]$ if

$$\frac{\sum_{i=1}^{n} q_i^2}{n\left(\sum_{i=1}^{n} q_i\right)^2} - \frac{1}{n^2} > 0$$

$$\Rightarrow \sum_{i=1}^{n} q_i^2 - \frac{\left(\sum_{i=1}^{n} q_i\right)^2}{n} > 0$$

which is *always* true since the last expression is the sample variance of the measured flows.

Special Case #3 – Load estimator using flow-weighted mean concentrations and known total discharge

This case is identical to special case #2 with the exception that the *fwmc* is applied to the total (annual) discharge, $\sum_{i=1}^{K} q_i$. Thus, the weights are as before except that the $\left\{v_j\right\}$ weights span the period of interest *(j=1,..,K)* rather than the sample *(j=1,..,n).*

Thus,

$$\hat{L}_3 = \frac{\sum_{i=1}^{n} c_i q_i}{\left(\sum_{i=1}^{n} q_i\right)} Q \tag{11}$$

where *Q* is the total (annual) discharge.

Furthermore,

$$Var\left[\hat{L}_3\right] = \frac{K\sum_{i=1}^{n} q_i^2}{\left(\sum_{i=1}^{n} q_i\right)^2} Var\left[L\right] \tag{12}$$

Note, if we have sampling fraction $f = \frac{n}{K}; \quad 0 < f < 1$ then equation 10 can be written as $Var\left[\hat{L}_2\right] = \frac{Var\left[\hat{L}_3\right]}{f}$ and it is evident that $Var\left[\hat{L}_3\right] < Var\left[\hat{L}_2\right]$.

5. The duality of the *fwmc* load estimator and a ratio estimator

Ratio estimation is a well known technique for potentially reducing the error (increasing the precision) of the estimate when an auxiliary variable that is correlated with the variable of interest is available. A full treatment of ratio estimators is given in Cochran (1977). In the present context, a ratio estimator is formed by assuming the ratio of the total load for the *sample* to the total discharge for the *sample* is the same as the corresponding quantities over the period of interest. That is

$$\frac{l}{q} = \frac{L}{Q}$$

Where *l (L)* is the sample (population) load and *q (Q)* is the sample (population) discharge. The ratio estimator is then

$$\hat{L}_{ratio} = \left(\frac{l}{q}\right)Q \tag{13}$$

Expanding equation 13, we have

$$\hat{L}_{ratio} = \frac{\left(\sum_{i=1}^{n} w_i c_i\right)\left(\sum_{i=1}^{n} v_i q_i\right)}{\sum_{i=1}^{n} q_i} \cdot \sum_{j=1}^{K} q_j \tag{14}$$

and letting $v_i = 1 \quad \forall i$ and $w_i = \frac{q_i}{\sum_{i=1}^{n} q_i}$ we see that $\hat{L}_{ratio} = \hat{L}_3$.

6. An Example

We consider the estimation of the total phosphorous (TP) load in a drain (designated CG3) in Gippsland, Victoria during the 2004 irrigation season[3]. The availability of daily flow and TP measurements enables us to compute the ‘true’ load as 5,517.10 kg. A random sample of *n*=29 observations were taken and the results used to demonstrate the methods outlined in this paper. The parameters given in table 1 were estimated from the *log*-transformed flow and concentration data.

Table 1. Parameters for log-flow and log-concentration

	Log-Flow	*Log-Concentration*
μ	2.5561	-0.02834
σ	0.6706	0.8008
ρ	0.482	

[3] Data courtesy of Southern Rural Water

By substituting the parameter estimates in table 1 into equations (3) and (4) we obtain (using equation (4)) estimate the load variance to be $Var\left[L\right] = 3132.863$. We next obtain load estimates using methods 1-3.

Method#1

Our data yield: $n = 29$, $\overline{c} = \frac{1}{n}\sum_{i=1}^{n} c_i = 1.17225$, and $\overline{q} = \frac{1}{n}\sum_{i=1}^{n} q_i = 11.4015$. The duration of the irrigation season is such that K=279 days. Thus

$$\hat{L}_1 = (279)(1.17225)(11.4015) = 3728.95\,\text{kg}$$

Compared to the ‘true’ load of 5517.10kg, $\hat{L}_1$ is seen to *underestimate* the true load by 33%. This overestimation is a consequence of the high (positive) correlation between log-concentration and log-flow. A bias correction factor (Fox 2004) can be applied in attempt to reduce this effect. In this case an improved estimate is obtained by multiplying $\hat{L}_1$ by $\exp\left\{Cov\left[\ln C, \ln Q\right]\right\} = \exp\left(\rho\sigma_c\sigma_q\right) = 1.2954$. This gives a modified total load of 4830.5kg which has reduced the bias to 13%.

From equation (8) we have

$$Var\left[\hat{L}_1\right] = \frac{279^2}{(29)(29)} Var\left[L\right] = 350220.28$$

and hence $SE\left[\hat{L}_1\right] = \sqrt{Var\left[\hat{L}_1\right]} = 591.8$.

Method#2

From equation (9)

$$\hat{L}_2 = K'\sum_{i=1}^{29} c_i\, q_i$$

$$= \frac{279}{29}(434.410) = 4179.32\,\text{kg.}$$

Compared to the ‘true’ load of 5517.10kg, $\hat{L}_2$ is seen to *underestimate* the true load by 24%. From equation (10) we have

$$Var\left[\hat{L}_2\right] = \frac{279^2 Var\left[L\right]}{(29)\left(\sum_{i=1}^{29} q_i\right)^2} \sum_{i=1}^{29} q_i^2$$

$$= \frac{279^2}{29} \frac{(4553.11)(3132.863)}{330.643^2} = 350220.28$$

and hence $SE\left[\hat{L}_2\right] = \sqrt{Var\left[\hat{L}_2\right]} = 591.8$.

Method#3

From equation (11)

$$\hat{L}_3 = K \frac{\sum_{i=1}^{29} q_i^2}{\left(\sum_{i=1}^{29} q_i\right)^2} Var\left[L\right]$$

$$= \frac{279(4553.11)(3132.863)}{330.643^2} = 36402.82 \text{ kg.}$$

Compared to the 'true' load of 5517.10kg, $\hat{L}_2$ is seen to *underestimate* the true load by 24%. From equation (10) we have

$$Var\left[\hat{L}_2\right] = \frac{279^2 Var\left[L\right]}{(29)\left(\sum_{i=1}^{29} q_i\right)^2} \sum_{i=1}^{29} q_i^2$$

$$= \frac{279^2}{29} \frac{(4553.11)(3132.863)}{330.643^2} = 36402.82$$

and hence $SE\left[\hat{L}_3\right] = \sqrt{Var\left[\hat{L}_3\right]} = 190.8$.

Appendix A – Derivation of Equation 6

Equation (1) can be written in matrix notation as $\hat{L} = K\left(1^{\mathbf{T}}Wc\right)\left(1^{\mathbf{T}}Vq\right)$. Observe that the term inside each bracket is a scalar and hence $\hat{L}^T = \hat{L}$. Thus

$$\begin{aligned}\hat{L} &= K\left(c^T W^T 1\right)\left(1^T Vq\right)\\ &= Kc^T\left(W^T 11^T V\right)q\\ &= Kc^T Aq\end{aligned}$$

where $A = W^T 11^T V$. Since the trace of a scalar is the scalar itself, we have

$$c^T Aq = tr\left(c^T Aq\right) = tr\left(Aqc^T\right) = tr\left(AB\right)$$

where $B = qc^T$.

Now *A* can be written as the product of two vectors, $A = wv^T$ where the vectors *w* and *v* are each of length n and are zero except for the sampled days, (for concentration and flow respectively), when they contain the respective weights for those sampled days. So the (i, j) element of *A* is the product of the sample weights when i is in I and j is in J, and 0 otherwise. The matrix *B* is also the product of two vectors, so $B_{i,j} = q_i c_j$.

Now, $tr\left(AB\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\, b_{ji}$ and so $tr\left(AB\right) = \sum_{i\in I}\sum_{j\in J} w_i v_j c_i q_j$ hence

$$Var\left[tr\left(AB\right)\right] = \sum_{i\in I}\sum_{j\in J}\left(w_i v_j\right)^2 Var\left[c_i q_j\right] + 2\sum_{i\in I}\sum_{j\in J}\sum_{i'\in I}\sum_{j'\in J} Cov\left[c_i q_j, c_{i'} q_{j'}\right]$$

Equation (6) is obtained by assuming the covariances in the expression above are zero. While this is not unreasonable for daily loads well separated in time, it is unlikely to be true on short time scales, in which case equation (6) will most likely underestimate the true variance (since loads will tend to be positively correlated).

References

Cochran, W.G. (1977) Sampling Techniques (third edition). Wiley.

Fox, D.R. (2004) Statistical Considerations for the modelling and analysis of flows and loads – Components of load. Technical report 02/04, Australian Centre for Environmetrics.

Henderson, B. and Bui, E. (2004) Case Study: Sediment and Transport Models in ERA. Unpublished, CSIRO Canberra.

Letcher, R. A., Jakeman, A. J., Calfas, M., Linforth, S., Baginska, B., and Lawrence, I. (2002). "A comparison of catchment water quality models and direct estimation techniques." *Environmental Modelling and Software*, 17, 77-85.