ON THE ESTIMATION OF PHOSPHORUS
FROM THE NIAGARA RIVER TO LAKE ONTARIO

by

M.E. Thompson[1] and A.H. El-Shaarawi[2]

[1]University of Waterloo
Waterloo, Ontario

[2]Rivers Research Branch
National Water Research Institute
Canada Centre for Inland Waters
Burlington, Ontario, L7R 4A6

ABSTRACT

A predictive or model based approach to making inferences about the loading of a contaminant from a point source is presented and illustrated by estimating the total phosphorus, TP, loading from the Niagara River to Lake Ontario for the years 1967 to 1982. Information about explanatory variables and/or autodependence among successive observations can be easily included in the model and this allows more accurate inferences to be made. The results of the application to the Niagara River show a major decrease in TP loadings has occurred during the study period.

RÉSUMÉ

    Un modèle permettant de prédire la charge d'un polluant émis à partir d'une source ponctuelle est présenté et illustré à l'aide d'une estimation de l'accumulation du phosphore total, TP, de la rivière Niagara jusqu'au lac Ontario entre 1967 et 1982. Des informations relatives aux variables indépendantes ou à l'autodépendance existant entre les observations successives peuvent être facilement incluses dans le modèle, ce qui permet de réaliser des prédictions plus précises. L'application de notre méthode dans la rivière Niagara a montré une diminution importante des charges polluantes du TP au cours de la période durant laquelle nous avons réalisé notre étude.

MANAGEMENT PERSPECTIVE

External information and autodependence are used to develop an empirical approach for load estimation. The approach is applied successively to model total phosphorus loading from the Niagara River to Lake Ontario. The results show a major decrease in TP has occurred during the period 1967 to 1982.

PERSPECTIVE-GESTION

Une méthode empirique d'estimation de la charge polluante a été
mise au point par l'utilisation d'informations externes et de
l'autodépendance. Cette méthode est appliquée de façon successive
pour modéliser l'accumulaton du phosphore total de la rivière Niagara
jusqu'au lac Ontario. Les résultats montrent qu'il y a eu une
diminution importante du PT de 1967 à 1982.

INTRODUCTION

Water quality of many lakes and rivers are managed by controlling the inputs (loads) of nutrients and toxic substances in both point (municipal and industrial) and nonpoint (agricultural, urban and rural) sources of pollution. For example, the 1978 Agreement between Canada and the United States has the target that the waters of the Great Lakes be "free from nutrients directly or indirectly entering waters as a result of human activity in amounts that create growth of aquatic life that interfere with beneficial uses". The achievement of the objectives of the control require a sound monitoring strategy and a reliable method of load estimation.

This paper presents a statistical approach for modelling and estimating the load from a point source. As an illustration, the technique is used to estimate the phosphorus load from Niagara River to Lake Ontario. The basic statistical issue arises from the fact that although the daily flows of the river are usually available, the concentrations of pollutants are only available on a small number of days each year.

An estimate the mean daily load, $\bar{L}$, of a chemical past a certain location in a river over a period T is formally expressed as

$$\bar{L} = \frac{1}{T} \int_0^T CQdt,$$

where Q and C are the instantaneous water flow and concentration respectively. This is usually approximated by

$$\bar{L} = \frac{1}{N} \sum_{i=1}^{N} C_i Q_i \qquad\qquad (1)$$

where N is the number of days in T and Q_i and C_i are the flow and the concentration in day i respectively. Typically the Q_i's are known, but the C_i's are known for a sample s of N days.

The ratio estimator (Cochran, 1977), corrected for sampling bias, is currently used by the International Joint Commission (IJC) to estimate $\bar{L}$. However, from recent advances in sampling theory this choice can be criticized. First of all, the sample of days for which concentrations would be available is not necessarily random. Second, there are estimators besides the ratio estimator which would use the flow data fully in estimating $\bar{L}$. The choice among these ought ideally to depend on a model for the daily concentration. The ratio estimator is optimal if the variance of the concentration varies as 1/flow (Royall, 1971). From the Niagara River 1975 to 1982 data, this inverse relationship does not appear to be attained, although in the winter months flow decreases and concentration fluctuates a little more widely. Thus the most widely accepted justification for the ratio estimates does not apply in this case.

Since the sample is not random and some serial autocorrelation seems likely, a predictive or model based approach to the estimation of loading is suggested in this paper. In this approach, $C_1, C_2, \ldots, C_N$ are jointly distributed random quantities and the

sampled C_i are used to "predict" the unsampled C_i and hence $\bar{L}$. It is also hoped the model will yield suitable estimates of uncertainty in the prediction of $\bar{L}$. Note that it is assumed that the sampled C_i are determined without error, and that therefore the uncertainty arises only from the fact that some data are missing.

THE MODELLING APPROACH

Suppose that a transformation $Z_i = h(C_i)$ exists such that the distribution of Z_i is approximately normal with a constant variance σ^2 and a mean

$$E(Z_i) = M_i = \sum_{j=1}^{p} a_{ij} \theta_j \tag{2}$$

where the a_{ij}'s represent measurements made on a set of p explanatory variables associated with Z_i and the θ_j's represent a set of p unknown parameters. In most water quality applications $h(C_i) = \ln C_i$, that is, the concentrations follows a log normal distribution. It is more appropriate to use matrix notation and to partition the vector Z of variables $Z_1,\ldots,Z_N$ into Z_s and Z_r which classifies the Z's into those included in the sample and the remainder. Let n be the sample size and without loss of generality, $Z_1,\ldots,Z_n$ be the elements of Z_s, and the remaining $Z_{n+1},\ldots,Z_N$ be the $N-n$ elements of Z_r. In matrix notation we have

$$Z = \begin{bmatrix} Z_s \\ Z_r \end{bmatrix}$$

$$E(Z) = A\theta = \begin{bmatrix} A_s \\ A_r \end{bmatrix}\theta$$

where A is the matrix of order Nxp with elements a_{ij}'s while A_s and A_r represent the matrices corresponding to the vectors Z_s and Z_r respectively. Furthermore, let the variance-covariance matrix of Z be V which also is partitioned as

$$V = \begin{bmatrix} V_{ss} & V_{sr} \\ V_{rs} & V_{rr} \end{bmatrix}$$

Given V it can be shown that the best linear predictor of Z_r is given by:

$$\hat{Z}_r = A_r\hat{\theta} + V_{rs}V_{ss}^{-1}(Z_s - A_s\hat{\theta}) \tag{3}$$

where

$$\hat{\theta} = (A_s' V_{ss}^{-1} A_s)^{-1} A_s' V_{ss}^{-1} Z_s$$

and A_s' refers to the transpose of A_s. The matrix of the mean square error (MSE) of the prediction is given by

$$\delta_{ss} = (V_{rr} - V_{rs} V_{ss}^{-1} V_{sr}) + (A_r - V_{rs} V_{ss}^{-1} A_s)(A_s' V_{ss}^{-1} A_s)^{-1}$$
$$(A_r - V_{rs} V_{ss}^{-1} A_s)' \tag{4}$$

Since for the Niagara River data the logarithm of the concentration appears to be the appropriate transformation, then $\hat{Z}_r$ is an unbiased estimate for Z_r and has a Gaussian distribution with variance-covariance matrix δ_{ss}. From the work of Thompson and Bischoping (1986), nearly unbiased estimates $\hat{C}_r$ and var $(\hat{C}_r)$ for the mean and variance of C_r are easily obtained. This leads to estimating the mean daily load by

$$\hat{\bar{L}} = (\underline{C}_s{}'\underline{Q}_s + \hat{C}_r'\underline{Q}_r)/N \qquad (5)$$

where $\underline{Q}_s$ is the vector of the flow values that corresponds to the days where the concentrations were measured and $\underline{Q}_r$ is the vector of the remainder flow values.

The variance of $\hat{\bar{L}}$ is given by

$$\text{var } (\hat{\bar{L}}) = \underline{Q}_r' \text{ Var}(\hat{C}_r)Q_r/N^2 \qquad (6)$$

The above expressions can be easily derived by noting that if Z_i and Z_j are normally distributed with means μ_i and μ_j and variances $\sigma^2{}_i$ and $\sigma^2{}_j$, then $C_i = \text{Exp } Z_i$ and $C_j = \text{Exp } Z_j$ are log normally distributed with means $\eta_i = \text{Exp } (\mu_i + \sigma^2{}_i/2)$ and $\eta_j = \text{Exp } (\mu_j + \sigma^2{}_j/2)$, and variances $\eta^2{}_i (\text{Exp } \sigma^2{}_i - 1)$

and η^2_i (Exp σ^2_j - 1) respectively. The covariance between c_i and c_j is

$$\eta_i \eta_j \; e^{\rho_{ij}\sigma_i\sigma_j} \{Exp \; (\sigma^2_i + \sigma^2_j + 2\rho_{ij} \; \sigma_i\sigma_j) - 1\}$$

where ρ_{ij} is the correlation between Z_i and Z_j.

APPLICATION

The data used to illustrate the methods of this paper consist of the daily flow and total phosphorus (TP) concentrations for the period from June 1975 to December 1982. The flow values are available for each day while the TP concentrations were not measured for some days. Figure 1 displays the monthly means for the flow and TP series. The plot shows a clear seasonal pattern and a slight downward trend in TP. Also the fluctuations in TP appear to increase with the increase in the level of TP. As a result, it was decided to perform the statistical analysis after transforming TP values to logs. Examination of the plots (not given in this paper) of the deviations of the daily flow within a month from its monthly mean and the corresponding plots for log TP showed that:

    i) within months, both series appeared roughly stationary;

    ii) variability was somewhat higher for some months than others, but there appeared to be no clear seasonal pattern in variability;

iii) there were several 'outliers' in TP concentration; the lowest value of -1.94 for log TP was eliminated from the latter and considered as missing.

The autocorrelation and partial autocorrelation functions for sample three months periods for the phosphorus series were computed using SAS. Table 1 gives an example of these functions. In each case, the output was compatible with a stationary AR(1) or a white noise model. The estimated lag 1 autocorrelations are given in Table 2. These three month periods were chosen because of the presence of long stretches in which no TP values were missing.

Finally, for a sample of months, an AR(1) model was assumed and the lag 1 autocorrelation was estimated for the log phosphorus series by the method of maximum marginal likelihood (Ramakrishnan, 1985). For these computations only, the six highest log TP values were designated 'outliers' and removed. The results are given in Table 3. From the Table it may be noted that the autocorrelation appears to be somewhat stronger for the winter months. Although some of the autocorrelation values are suspiciously high, no association was found between estimated lag 1 autocorrelation and number of missing days.

Assuming the AR(1) model for the log TP values, the estimates of the monthly and yearly mean daily phosphorus loadings are given in Table 4 along with their estimated standard errors. It is clear from the Table that a major reduction in the total TP load to Lake Ontario has occurred between 1976 and 1982. All the standard error are

comparable from year to year with the exception of 1978 where the highest load has occurred along with the largest standard error.

CONCLUSION

This paper describes a method for estimating the input load from a source of pollution and illustrated its use for estimating the total phosphorus load from the Niagara River to Lake Ontario between 1976 to 1982. One major advantage of the method is it allows the utilization of any relevant available information which can lead to a more precise load estimate. Most available methods assume that the concentrations are a realization of a sequence of independent random variables and hence they do not account for the presence of serial correlation which is likely to be present since the data represent a time series. The method given here is well adapted to the nature of data by not only including the dependence among successive concentrations, but also allowing the utilization of any quantitative information in the estimation process, and this was clearly shown in the Niagara River example.

REFERENCES

Cochran, W.G., 1977. Sampling techniques. Wiley, New York, 428 pp.

Ramakrishnan, V., 1985. Marginal likelihood analysis of growth curves. M. Phil. Thesis, University of Waterloo.

Royall, R.M., 1971.   Linear regression models in finite population
    sampling theory.   In: Foundations of Statistical Inference, ed.
    V.P. Crodambe and D.A. Sprott.   Holt, Rinehart and Winston,
    Toronto.

Thompson, M.E. and K. Bischoping, 1986.   On the estimation of monthly
    mean phosphorus loadings.   Developments in Water Science, V27,
    Elsevier, pp 460-468.

Table 1. Autocorrelation and partial autocorrelation functions and their standard errors (June-August 1975)

Lag k	Autocorrelation r_k	St. Error $\sigma(r_k)$	Partial Autocorrelation ϕ_k	St. Error of $\sigma(\phi_k)$
1	0.57	0.10	0.57	0.10
2	0.47	0.13	0.21	0.10
3	0.32	0.15	-0.02	0.10
4	0.19	0.16	-0.06	0.10
5	0.04	0.16	-0.15	0.10
6	-0.06	0.16	-0.08	0.10
7	-0.12	0.16	-0.04	0.10
8	-0.25	0.16	-0.17	0.10
9	-0.25	0.17	-0.01	0.10
10	-0.24	0.17	-0.01	0.10
11	-0.24	0.17	-0.05	0.10
12	-0.16	0.18	0.07	0.10
13	-0.07	0.18	0.07	0.10
14	0.09	0.18	0.17	0.10
15	0.14	0.18	0.04	0.10
16	0.27	0.18	0.11	0.10
17	0.31	0.18	0.06	0.10
18	0.21	0.19	-0.18	0.10
19	0.22	0.19	0.00	0.10
20	0.16	0.20	0.00	0.10
21	0.11	0.20	-0.02	0.10
22	0.05	0.20	-0.12	0.10
23	-0.13	0.20	-0.11	0.10
24	-0.17	0.20	0.05	0.10

Table 2. Lag 1 autocorrelations for the three month periods

Period	Autocorrelation
June to August 1975	0.57
September to November 1975	0.49
April to June 1977	0.33
July to September 1982	0.25
June to August 1983	0.20*
September to November 1983	0.35

*not significant

Table 3. Estimated lag 1 autocorrelations for <u>lnphos</u> (number of missing/outlier observations)

Month \ Year	1976	1977	1978	1979	1980	1981	1982
JAN	0.36(5)	0.81(13)	0.86(28)	(0)	0.73(14)	0.55(1)	0.50(4)
FEB	0.50(9)	(0)	0.96(10)	0.23(6)	0.71(7)	-0.19(11)	0.55(4)
MAR	0.15(6)	0.56(8)	0.70(0)	(0)	0.60(7)	0.13(3)	(0)
APR	(0)	(0)	0.19(12)	0.76(13)	(0)	0.82(3)	0.43(1)
MAY	0.21(8)	0.44(2)	0.79(9)	0.39(7)	-0.15(1)	0.46(4)	-0.25(1)
JUN	0.01(8)	(0)	0.15(18)	0.41(13)	0.43(2)	0.20(5)	0.55(3)
JUL	0.04(4)	0.49(7)	0.40(10)	0.61(8)	0.23(1)	0.00(6)	0.08(1)
AUG	0.11(5)	(0)	(0)	0.33(7)	0.28(2)	0.50(2)	(0)
SEP	(0)	0.89(1)	-0.05(12)	0.51(19)	0.55(2)	(0)	0.00(3)
OCT	0.55(9)	0.79(1)	0.58(1)	0.12(14)	0.66(4)	0.61(3)	0.45(1)
NOV	0.59(10)	0.73(1)	0.71(4)	0.40(0)	0.54(1)	0.86(10)	(0)
DEC	0.36(3)	0.56(10)	0.83(15)	0.63(11)	(0)	0.66(3)	0.05(1)

Table 4. Estimated mean daily TP loadings in tonnes for each month and the yearly loading for the years 1976 to 1982

Month \ Year	1976	1977	1978	1979	1980	1981	1982
JAN	12.12(0.324)*	8.75(0.222)	15.91(1.548)	18.04(0)	20.26 (1.055)	7.155(0.036)	18.771(0.767)
FEB	12.21(0.602)	9.12(0)	12.32(0.329)	11.52(0.76)	17.25 (1.067)	7.692(0.289)	7.488(0.134)
MAR	15.82(0.281)	11.64(0.107)	16.43(0)	11.07(0)	10.178(0.360)	6.655(0.097)	10.825(0)
APR	18.26(0)	11.57(0)	14.84(0.672)	15.31(0.120)	11.554(0)	8.686(0.181)	10.583(0.086)
MAY	14.66(0.360)	11.59(0.106)	13.05(0.498)	10.88(0.39)	10.438(0.148)	10.758(0.121)	13.799(0.102)
JUN	11.73(0.254)	9.98(0)	14.42(3.553)	8.74(0.56)	12.874(0.253)	9.959(0.322)	8.942(0.196)
JUL	10.88(0.237)	9.42(0.282)	13.74(0.615)	10.67(0.44)	8.768(0.097)	7.037(0.051)	6.777(0.065)
AUG	13.83(0.820)	8.10(0)	17.67(0)	7.08(0.181)	8.874(0.101)	6.255(0.088)	6.765(0)
SEP	8.17(0)	18.14(0.385)	10.79(0.840)	6.69(0.46)	10.119(0.285)	7.518(0)	7.015(0.091)
OCT	10.20(0.297)	10.66(0.120)	9.52(0.115)	11.15(0.60)	10.173(0.343)	10.846(0.249)	8.507(0.116)
NOV	15.70(1.122)	11.34(0.104)	11.15(0.347)	9.10(0)	12.069(0.196)	11.698(0.567)	12.098(0)
DEC	17.83(0.264)	22.48(0.950)	25.73(1.315)	10.72(0.45)	11.908(0)	13.380(0.351)	15.307(0.267)
YEARLY	4916.31(51.30)	4350.41(43.23)	5357.17(131.11)	3987.92(56.04)	4424.85 (49.69)	3278.04 (26.28)	3874.69 (27.62)

*Values in brackets are estimates of standard error due to sampling. A value of 0 means no data for that month were missing.