RRB 89-04

STATUS OF THE NIAGARA RIVER POINT SOURCE DISCHARGE INFORMATION: SAMPLING DESIGN AND ESTIMATION OF LOADING

by

A.H. El-Shaarawi[1] and D.J. Williams[2]

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

[2]Inland Waters Directorate
Ontario Region

January 1989
NWRI Contribution #89-78

ABSTRACT

Two basic requirements, low bias and high precision, are necessary for generating reliable estimates for the load from point and nonpoint sources of pollution. Biases and low precision can be the result of using a bad sampling design and/or an inadequate method of estimation. The effects of biases can be reduced at the design stage prior to the data collection or at the data anlysis stage. This paper discusses the statistical issues involved in generating adequate load estimations using recently published point source discharge data from the Niagara River to illustrate these issues.

RÉSUMÉ

Une erreur systématique faible et une grande précision sont les deux exigences de base pour obtenir des estimations fiables de la charge polluante originant de sources de pollution ponctuelles et non ponctuelles. Un plan d'échantillonnage mal conçu ou une méthode d'estimation non appropriée peuvent être responsables des erreurs systématiques et de la faible précision des résultats. Les erreurs systématiques peuvent être corrigées en partie au moment de l'établissement du plan d'échantillonnage avant la collecte des données ou au stade d'analyse des données. Cet article examine les traitements statistiques nécessaires à l'obtention d'estimations précises de la charge polluante. Les problèmes de statistiques relatifs à ces estimations seront illustrés à l'aide de données de déversements localisés prises dans la rivière Niagara et publiées récemment.

MANAGEMENT PERSPECTIVE

This paper presents the necessary statistical conditions involved in generating precise and accurate load estimation and uses recently published point source discharge data from the Niagara River to illustrate the issues involved. The findings are useful for designing both a sampling design and choosing the approach for load estimation.

PERSPECTIVE-GESTION

À partir de données de déversements localisés prises dans la rivière Niagara et publiées récemment, cet article présente les traitementsstatistiques nécessaires à l'obtention d'estimations précises de la charge polluante. Les résultats obtenus s'avèrent utiles pour la conception d'un plan d'échantillonnage et pour le choix d'une méthode appropriée d'estimation de la charge polluante.

INTRODUCTION

Toxic pollutants are discharged to the Niagara River from a number of municipal and industrial point sources. According to the 1984 Report of the Niagara River Toxics Committee (NRTC, 1984), ninety-five percent of the total point source load of EPA priority pollutants was contributed by 37 of the 188 known discharges. Two reports summarizing the most recent (1985/86) loading data were subsequently released in August and September 1987 by the New York State Department of Environmental Conservation (DEC) and the Ontario Ministry of the Environment (MOE) respectively (McMahon, 1987; MOE, 1987). Both reports identified major reductions between 1981/82 and 1985/86 in both the organic and inorganic priority pollutant loads from these same facilities. The DEC Report, in particular, cited a number of causative factors for these reductions including plant closings, process shutdowns, the completion of wastewater treatment plants, and the successful implementation of remedial programs.

Closer scrutiny, however, suggests that there are also statistical considerations related to (1) the sampling design used to generate the data and (2) the method used for computing the loads in both these reports, that were overlooked and cause some concern about the reliability of the reported load estimates. The statistical factors which limit the usefulness of the data include: (a) the presence of systematic error; (b) the low level of precision; and (c) the unavailability of a measure of uncertainty in the data (either for individual facilities or the total load estimate).

The objectives of this paper are to discuss the roles of the statistical issues involved in designing an efficient monitoring program capable of reliably estimating the loads to the river and detecting real year to year changes in these load estimates. Data from the DEC and MOE reports are used to illustrate the issues involved. In addition, a method for sampliing the effluents is proposed which is useful for measuring both the total load to the river and the loads from individual major dischargers.

REQUIREMENTS FOR A RELIABLE LOAD ESTIMATE

A reliable estimate of the load must satisfy a number of requirements which need to be kept in mind prior to planning the data collection and during analysis and interpretation of the data. Briefly, these are: (1) absence of systematic error (bias), (2) a high level of precision and (3) some estimate of the uncertainty surrounding the load estimate generated from the data. The role of each requirement is discussed briefly below with examples.

1. Bias: It is important that load estimates be free of bias. Bias will lead to over- or underestimation of the load. One of the main difficulties is not knowing the direction, magnitude and/or even the causes of the bias. The sources of bias include: (1) contribution from unknown dischargers (will cause an underestimate of unknown magnitude), (2) no data for known dischargers (an estimate of these

loads can, perhaps, be attempted using external information), (3) lack of knowledge or complete ignorance about the nature and structure of the variability in both concentrations and flows, (4) changes in sampling and analytical techniques, and (5) the method used to estimate the loads.

Tables 1 to 4 illustrate the effect of ignoring the data structure when estimating the loading.

Table 1 presents measurements of flow and total phosphorus concentration (TP) taken by MOE over a consecutive three day period on four separate occasions at the Niagara Falls WWTP. The data indicate large differences in TP concentrations between the different samples and hence, a sampling plan which does not take this into account will lead to a biased estimate for the load. Samples taken on each of the three days within each of the four sampling occasions were treated as replicates in a one way analysis of variance (ANOVA) to evaluate the significance of differences among the concentrations on different occasions. Table 2 shows that when the four separate sampling occasions are compared, both the TP concentrations and TP loads exhibit strong significant differences.

Table 3 presents the monthly average total priority pollutant loads for the ten most significant U.S. dischargers to the Niagara River for 1985/86 (McMahon, 1987). A similar analysis to that noted above was carried out by dividing the 12 months of data into four groups of three months each. The results, provided in Table 4, show that the loads to the river from some facilities exhibit significant "seasonal" differences.

The method of load estimation itself provides yet another possible source of bias. For example, the DEC Report compared the median load estimates based on the dischargers monthly data to the mean based on one day sampling by DEC. Since dischargers have more frequent data, the median was used as the basis for the comparison. As is commonly known, many environmental data have a typical non-symmetric distribution with a very long right tail due to the presence of extremely high values. As a result, the median will tend to underestimate the mean load. This is illustrated in Table 5 which shows that the mean exceeds the median for the majority of dischargers. Indeed, if the log loads follow a normal distribution with mean μ and variance σ^2, then the load will have a log-normal distribution with mean Exp $\{\mu + \sigma^2/2\}$ and median e^μ. This shows that the mean equals the median times Exp $(\sigma^2/2)$ and hence, the degree of bias in estimating the mean by the median depends on the magnitude of σ^2.

2. <u>Precision</u>: The precision of the estimate depends on the sampling design and the method used for estimation. The number of samples and the spacing of the sampling dates are among the most important components of the sampling design. For a given sample size, the gain in the increase of precision is related to the structure of the variability within the sampling period. If the input of a discharger is homogeneous, then the spacing of the sampling dates plays no role in determining the precision of the estimate. In this case, the variance of the mean load is proportional to 1/sample size.

There are a large number of methods available for estimating the load and the one that will produce the highest precision depends to a very large extent on the development of a model for representing the variability of the load.

3. Estimation of Uncertainty: It is not sufficient to report only an estimate for loads without producing from the data a measure of the reliability of such estimates. When the distribution of the estimator is highly variable, then a single estimate is useless and can even be misleading.

A DESIGN FOR ESTIMATING THE LOAD FROM MAJOR DISCHARGERS AND THE TOTAL LOAD

Suppose it is required to estimate the input of a large number, p, of point sources to a river such as the Niagara River and suppose that the flow information is reasonably well known or can be easily or cheaply obtained but the input concentrations are unknown. Suppose further, that out of the p point sources, s are classified as significant. Due to financial and technical constraints, it is only possible to perform chemical analyses on n samples, with $n<p$. Realizing that it is not possible under these conditions to estimate the mean concentration for each point source, the objectives are modified to: (1) estimating the mean concentration for each significant point source, (2) estimating the mean concentration for the non-significant point

sources, (3) estimating the overall mean concentration of the input to the river, and finally (4) testing the significance of the difference between the mean concentrations of two point sources.

Here we describe an approach for achieving the above objectives. To do this, let μ_i be the true but unknown mean concentration for the i<u>th</u> significant point source ($i = 1,2,\ldots,s$) and μ_{s+1} be the true mean concentration for the remaining point sources. The overall mean concentration of the input to the river is then given by:

$$\mu = \{\sum_{i=1}^{s} \mu_i + (p - s)\mu_{s+1}\}/p$$

or the flow weighted mean is given by:

$$\mu_F = \{\sum_{i=1}^{s} \mu_i F_i + \mu_{s+1} (\sum_{i=s+1}^{p} F_i)\}/\sum_{i=1}^{p} F_i$$

We are interested in estimating the μ's and their standard errors. The approach is to perform chemical analyses on samples which are composites of subsamples from the different point sources. Let y_j be the concentration of the j<u>th</u> samples ($j = 1,2,\ldots,n$), let x_{ji} be an indicator variable which takes the value 0 or 1 depending on whether or not the j<u>th</u> sample includes a subsample from the i<u>th</u> source. Hence, the mean θ_j of y_j is given by:

$$\theta_j = \sum_{i=1}^{s+1} x_j \mu_i$$

It is more convenient to represent this set-up in the matrix form:

$$\underline{y} = X\underline{\mu} + \underline{\epsilon}$$

where $\underline{y}$ is the vector of observations of length n, X is the binary matrix of order (n x s+1) with x_{ji} as its elements, $\underline{\mu}$ is a vector of length (s+1) with elements $(\mu_1,\ldots,\mu_{s+1})$, and $\underline{\epsilon}$ is a vector of n independent identically distributed random variables with zero mean and variance σ^2. The least squares estimates of $\underline{\mu}$ and σ^2 are given by:

$$\hat{\underline{\mu}} = (X'X)^{-1}X'\underline{y}$$

and

$$\hat{\sigma}^2 = \frac{1}{n-s-1} \sum_{j=1}^{n} (y_j - \sum_{i=1}^{s+1} x_{ji} \hat{\mu}_i)^2$$

where $\hat{\mu}_i$ is the i<u>th</u> element of $\hat{\underline{\mu}}$, and X' is the transpose of the matrix X.

The analysis is greatly simplified if the matrix X is as the design matrix for a Balanced Incomplete Block Design (BIBD). In the BIBD case the matrix (X'X) takes the simple form,

$$(X'X) = \begin{matrix} a & b \ldots\ldots b \\ b & a \ldots\ldots b \\ \cdot & \\ \cdot & \\ b & \quad a \end{matrix}$$

where a is the number of samples with a subsample from the $i\underline{th}$ score and b is the number of samples with subsamples from the $i\underline{th}$ and $j\underline{th}$ sources. Additional restrictions are required for the use of the BIBD design. These are: (1) each sample must consist of the same number, k, of subsamples, (2) (s+1) = kn and (3) bs = a(k-1). Design matrices for different combinations a, S+1 and n are tabulated in many experimental design books (e.g., Cochran and Cox, 1957).

The estimated variance-covariance matrix for $\hat{\underline{\mu}}$ is given by $(X'X)^{-1}\hat{\sigma}^2$. In the BIBD cases, the inverse of (X'X) is

$$(X'X)^{-1} = \frac{1}{(a-b)(a+bs)} \begin{matrix} a+b(s-1) & -b & -b & -b \\ -b & a+b(s-1) & -b & -b \\ -b & -b & a+b(s-1) & -b \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ -b & -b & -b & a+b(s-1) \end{matrix}$$

This leads to estimating μ_i by:

$$\hat{\mu}_i = \frac{\sum_{j=1}^{n} x_{ji}\, y_i}{a - b} - \frac{b}{(a - b)(a + bs)} \sum_{i=1}^{s+1} \sum_{j=1}^{n} x_{ji}\, y_i$$

and the variance of $\hat{\mu}_i$ is

$$\mathrm{var}\,(\hat{\mu}_i) = \frac{\hat{\sigma}^2}{a - b}\left(1 - \frac{b}{a+bs}\right)$$

To illustrate the method, suppose that there are four major dischargers and let the remaining dischargers be regarded as a single significant discharger. Suppose that there are only 10 samples to be taken from these point sources. One possible sampling design is given in Table 6.

Table 6. BIBD Sampling Plan

Sample			
1	D_4	D_5	D_1
2	D_4	D_2	D_5
3	D_2	D_4	D_1
4	D_5	D_3	D_1
5	D_3	D_4	D_5
6	D_2	D_3	D_1
7	D_3	D_1	D_4
8	D_3	D_5	D_2
9	D_2	D_3	D_4
10	D_5	D_1	D_2

where in the Table, D_i = the i<u>th</u> discharger ($i = 1,2,\ldots,5$). Let the measurements that will be generated from the sample plan be $y_1,\ldots,y_{10}$.

The steps involved in estimating the concentration of each discharger are given in detail below for the purpose of illustrating the calculations involved. The design matrix X is

$$X = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

Hence,

$$(X'X) = \begin{matrix} 6 & 3 & 3 & 3 & 3 \\ 3 & 6 & 3 & 3 & 3 \\ 3 & 3 & 6 & 3 & 3 \\ 3 & 3 & 3 & 6 & 3 \\ 3 & 3 & 3 & 3 & 6 \end{matrix}$$

$$(X'X)^{-1} = \frac{1}{18} \begin{matrix} 5 & -1 & -1 & -1 & -1 \\ -1 & 5 & -1 & -1 & -1 \\ -1 & -1 & 5 & -1 & -1 \\ -1 & -1 & -1 & 5 & -1 \\ -1 & -1 & -1 & -1 & 5 \end{matrix}$$

and $Y'X = (Q_1\ Q_2\ Q_3\ Q_4\ Q_5)$,

where
$$Q_1 = y_1 + y_3 + y_4 + y_6 + y_7 + y_{10}$$
$$Q_2 = y_2 + y_3 + y_6 + y_8 + y_9 + y_{10}$$
$$Q_3 = y_4 + y_5 + y_6 + y_7 + y_8 + y_9$$
$$Q_4 = y_1 + y_2 + y_3 + y_5 + y_7 + y_9$$
$$Q_5 = y_1 + y_2 + y_4 + y_5 + y_8 + y_{10}$$

Hence the estimates $\hat{\mu}_i$ ($i = 1,2,\ldots,5$) are given by:

$$\hat{\mu}_i = \frac{Q_i}{6} - \sum_{i=1}^{5} Q_i/18$$

with variance

$$\text{var}\ (\hat{\mu}_i) = \frac{5}{18}\,\hat{\sigma}^2$$

and

$$\hat{\sigma}^2 = \frac{1}{5}\,\{\sum_{i=1}^{10} y_i^2 - \sum_{i=1}^{5} \hat{\mu}_i\, Q_i\}$$

CONCLUSIONS

Estimates of the loads discharged from point sources are important for determining compliance with jurisdictional control requirements. In addition, as is the case in the Niagara River, load estimates may also be required for the development of simple mass balance scenarios for aquatic systems. In either case, it is important that these estimates be reliable and have low bias and high precision. Both bias and precision are affected by sampling design and method of estimating the loads. This paper has demonstrated the significance of these factors using data from the Niagara River which is currently being employed to effect management decision. We began this paper initially suggesting that closer scrutiny of the data in two recently released point source reports (McMahon, 1987; MOE, 1987) raises some concern about the reliability of the reported load estimates. Briefly, we have shown the following based on our examination of the data:

1. There is significant variability in both concentrations and loads both during and between sampling periods associated with the point sources sampled which, if not considered in the sample design and estimate of the load, can significantly bias the load estimates. Given this variation, without some estimation of the uncertainty of the data, the load estimate is highly questionable, if not misleading. From a management perspective, this raises some serious questions about the adequacy of single point in time samples for determining "compliance" not to mention comparing loads between years.

2. Choice of method of estimating and expressing the load is also important. The current data show that mean values based on single point in time sampling are not comparable with median values based on monthly facility data. Furthermore, even within the monthly facility data the median tended to underestimate the mean of the individual facility data in the majority of cases. The <u>mean</u> loads based on the single point in time sampling done by DEC were used as the basis for the reported load reductions between 1981/82 and

1985/86. In the majority of cases these reported mean loads were significantly less than the facility median values. From a management perspective, given the above, one must question the reliability and significance of the reported load reductions between the two sampling periods.

We can appreciate that much of the problem associated with getting reliable loads from sources can be directly related to logistics and resources. In this paper, we have proposed a sampling design based on a novel application of the BIBD plans which for a reduced cost will:

1. estimate the mean concentrations for each significant point source;
2. estimate the mean concentrations for the nonsignificant point sources;
3. estimate the overall mean concentration of the input, and
4. test the significance of the difference between the mean concentrations of two point sources.

We emphasize that this design is based on the assumption that the flow information is reasonably well known or can be measured easily and cheaply, but the input concentrations are unknown. Our inspection of the data indicate the flow measurements are probably less accurate and imprecise than those of concentrations. We would like to end this paper by recommending that dischargers implement better programs to measure flows.

REFERENCES

Cochran, W.G. and Cox, G.M. 1975. '<u>Experimental Design</u>'. 2nd ed. New York: Wiley.

McMahon, J.C. 1987. '<u>Comparison of 1981-82 and 1985-86 Toxic Substance Discharges to the Niagara River</u>'. New York State/Department of Environmental Conservation publication.

Niagara River Toxics Committee, 1984. A joint publication of Environment Canada, United States Environmental Protection Agency, Ontario Ministry of the Environment and New York State Department of Environmental Conservation.

Ontario Ministry of the Environment , 1987. '<u>Priority Chemical Loadings from Ontario Point Sources Discharging to the Niagara River, 1981 to 1987</u>'. Niagara River Improvement Project (ISBN No. 0-7729-2981-5).

Table 1. Total Phosphorus (TP) Data from the Niagara Falls (Stanford) WWTP[1].

Sample Data	Flow L/s	TP Concentration mg/L	Load mg/s
29/07/86	597.030	2.234	1327.795
30/07/86	592.822	2.120	1256.783
31/07/86	604.291	2.040	1232.754
03/03/87	658.954	0.720	474.447
04/03/87	580.301	0.480	278.544
05/03/87	586.403	0.640	375.298
24/03/87	476.551	1.020	486.082
25/03/87	841.778	0.860	723.929
26/03/87	522.417	1.030	538.090
21/07/87	704.463	0.530	373.365
21/07/87	704.463	0.520	366.321
22/07/87	708.672	1.240	878.753
23/07/87	712.197	0.980	697.953

[1]data from MOE

Table 2. Variance Ratios for the Flows, TP Concentrations and TP Loads at the Niagara Falls (Stanford) WWTP.

Flow	0.841
TP Concentration	13.392***
TP Load	10.179***

***significant at the 1% level

Table 3. Monthly Average Total Priority Pollutant Loads (ℓb/d) for the Ten Major U.S. Dischargers to the Niagara River, 1985/86.

	Bethlehem Steel Corp.	Buffalo Sewer	Mohawk Power Corp.	Town of Tonawanda	Spaulding Fibre Co.	Town of Amherst	Occidental Chem. Corp. Burez Div.	Occidental Chem. Corp. Niagara Plant	Olin Corp.	City of Niagara Falls, N.Y.
April	57.0		3.6	48.1	29.0		7.2	48.3	10.4	
May	93.3		4.1	29.3	46.2		6.8	35.0	7.1	
June	59.6		8.6	10.7	21.4		6.9	33.0	6.1	
July	80.6		5.8	21.3	21.3		2.4	31.8	4.8	
August	317.2		6.0	20.3	25.9	35.1	2.3	16.8	3.2	92.7
September	127.6		1.8	32.2	28.1	15.3	2.1	19.5	3.3	92.3
October	84.1	136.9	2.0	19.4	22.6	18.5	3.2	17.1	3.8	72.9
November	51.6	268.3	0.1	74.8	23.4	2.8	2.7	18.7	2.7	130.2
December	73.1	426.3	0.1	12.5	30.0	3.3	2.8	40.3	2.3	93.1
January	96.9	299.8	3.1	101.9	41.2	10.2	5.6	37.5	0.8	94.5
February	81.0	250.4	0.1	49.7	71.1	7.6	5.7	47.9	2.0	80.4
March	168.6	615.4	0.1	60.4	85.2	11.5	6.3	36.2	1.4	77.1

blank indicates data are not available

Table 4. Variance Ratios for Seasonal Differences in Total Priority Loads of Major U.S. Dischargers with Complete Data.

Discharger	Variance Ratio
Bethlehem Steel Corp.	1.614
Mohawk Power Corp.	3.910*
Town of Tonawanda	2.259
Spaulding Fibre Co.	6.518**
Occidental Chem. Corp. (Burez Div.)	221.429***
Occidental Chem. Corp. (Niagara Plant)	3.012*
Olin Corp.	10.413***

*significant at the 10% level
**significant at the 5% level
***significant at the 1% level

Table 5. Comparison of DEC* Sample and Facility Data for 1985/86 Loadings of Total Priority Pollutants from the Most Significant Dischargers to the Niagara River.

	DEC Sample Loading lb/d	Median lb/d	Mean lb/d	S.D. lb/d	CV $\sigma/\bar{x}$
Buffalo Sewer Authority	160.8	363.0	366.2	167.0	0.46
City of Niagara Falls	120.2	92.5	91.7	17.7	0.19
Bethlehem Steel Corp.	24.9	82.5	107.6	73.6	0.68
Niagara Mohawk Power Corp.	7.1	2.5	3.0	2.8	0.93
Olin Corp.	1.3	3.2	4.0	2.7	0.68
Spaulding Fibre Co.	92.8	28.5	37.1	20.9	0.56
Town of Tonawanda	10.9	30.7	40.0	27.9	0.69
Town of Amherst WWTP	15.1	10.8	13.3	10.2	0.77

*The DEC data represent the results of one to three unannounced 24 h composite time or flow proportional samples of wastewater.