RRB-88-94

INFERENCES ABOUT THE VARIABILITY OF MEANS FROM CENSORED DATA

by

A.H. El-Shaarawi*, P.B. Kauss+,
M.K. Kirby+, and M. Walsh+

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

+Ontario Ministry of
the Environment
Water Resources Branch
Toronto, Ontario M4V 1K6

December 1988
NWRI Contribution #89-119

ABSTRACT

In recent years, intensive biological monitoring studies have been carried out on the Niagara River by the Ontario Ministry of the Environment. The basic objective was to determine the relative bioavailability of trace contaminants at various locations in the river, and to identify sources. A recurring difficulty encountered with the generated data is that substantial portions of sample concentrations of many toxic pollutants are below the limits of detection established by analytical laboratories. Under the assumption that the distribution of the data is log normal, the likelihood ratio test for testing the equality of several means for type I censored data is derived and its use for evaluating the spatial variability of trace contaminants in the river is illustrated.

RÉSUMÉ

Au cours des dernières années, le ministère de l'Environnement de l'Ontario a fait d'importantes études de surveillance biologique dans la rivièrer Niagara avec comme principal objectif l'étude de la biodisponibilité relative des contaminants à l'état de traces à divers endroits dans la rivière, et l'identification des sources. Une difficulté qui revient fréquemment avec les données produites est qu'une grande partie des concentrations des échantillons de nombreux polluants toxiques sont en-deçà des limites de détection établies par les laboratoires analytiques. En supposant que la distribution des données est lognormale, le test du rapport des vraisemblances dans le but de tester l'égalité de plusieurs moyennes de données censurées de type I est dérivé, et son utilisation en vue de l'évaluation de la variabilité spatiale des contaminants à l'état de traces dans la rivière est illustrée.

MANAGEMENT PERSPECTIVE

In the presence of censored water quality data, a test statistic has been developed for testing the spatial and temporal heterogeneity of the concentration data for trace contaminants. The test does not require extensive computation and is fairly robust to distributional assumptions. The basic objective for developing the test was to determine the relative bioavailability of trace contaminants at various locations in a river and to identify sources. The application of the test is illustrated using an example from the Niagara River biomonitoring data.

PERSPECTIVE-GESTION

En présence de données censurées sur la qualité de l'eau, un test statistique a été mis au point pour tester l'hétérogénéité spatio-temporelle des donneés sur la concentration des contaminants à l'état de traces. Le test ne nécessite pas d'importants calculs et se prête assez bien aux hypothèses de distribution. Le principal objectif de la mise au point de ce test était d'établir la biodisponibilité relative des contaminants à l'état de traces à divers endroits dans la rivière et d'en identifier les sources. L'application du test est illustrée à l'aide d'un exemple faisant appel aux données de biosurveillance de la rivière Niagara.

INTRODUCTION

Both ambient water sampling and effluent monitoring are important tools in the identification and quantification of point and non-point contaminant inputs and their consequent effects on water quality. However, the importance of biological monitoring, either with natural or introduced species, should not be neglected in such monitoring studies, since it offers several distinct advantages. This stems from the continuous contact of the aquatic organisms with their environment and the tendency of persistent contaminants to accumulate in their tissues or lipids to concentrations which are more readily detectable than in ambient water samples. Therefore, contaminant levels in biota can provide information on the presence of nearby sources of contaminants. Furthermore, biota will tend to reflect inputs of contaminants which may be too sporadic to be detected by routine water quality monitoring programs.

In recent years, extensive biological monitoring programs have been conducted in the Niagara River by the Ontario Ministry of the Environment's (MOE) Water Resources Branch. These studies, using both natural and introduced species from various trophic levels, were included in the report of the Niagara River Toxics Committee (NRTC) released in 1984. In its report, the NRTC recommended a minimum "Long Term Monitoring Program" for a number of "Chemicals of Concern" identified in the river. This list of chemicals included those group I contaminants requiring: "immediate attention to determine

their origin and spatial and temporal trends", as well as those requiring additional data to make an adequate assessment of their significance.

Since then, annual biomonitoring using filamentous algae and young forage fish has been conducted at many of the sites recommended by the NRTC report. In addition, an extensive study using introduced (caged) mussels was conducted in the river by the MOE during 1983. One of the main objectives of the latter study was to determine the relative bioavailability of organochlorine contaminants in different areas of the Niagara River and to identify sources.

A recurring difficulty encountered during the course of the study was that, in a substantial portion of the samples, concentrations of many toxic pollutants were below the limits of detection established by analytical laboratories. The problem of estimating the mean and standard deviation from censored water quality data has been addressed by Gilliom and Helsel (1986), Helsel (1986), El-Shaarawi (1989), and El-Shaarawi and Dolan (1989).

In this paper we extend the work of El-Shaarawi and Dolan (1989) from making inferences about the log normal mean to that of testing the differences between several means. Specifically, the likelihood ratio test (LRT) for testing the equality of several means for type I censored data is derived. The application of the LRT test is illustrated using an example from the Niagara River biomonitoring data.

FIELD STUDY

The 1983 Niagara River biomonitoring study was conducted at 27 nearshore stations (Figure 1) and consisted of six surveys. At each station, clean mussels were exposed for 15 weeks (August 3 to November 16) in galvanized wire cages (Survey 6) and then recovered and processed as described by Kauss and Hamdy (1985). Mussels were also exposed for five consecutive 3 week periods (Surveys 1 to 5) during the above period at 20 of the stations to determine the temporal variability in contaminants input/bioavailability. These "intensive" stations were concentrated near suspected contaminant sources along the New York mainland shore (Figure 1). For each station and survey, the complete soft tissues of three replicates were each analyzed for 25 organochlorine contaminants including pesticides and industrial organics.

STATISTICAL METHODS

1. One Way Analysis of Variance for Censored Data

Let X_{ij} be the concentration of the contaminant in the jth sample at the ith sampling location where $j = 1,2,\ldots,n_i$; $i = 1,2,\ldots,k$ and let X_0 be the limit of detection established by the analytical laboratory. Suppose that d_i out of n_i samples have concentrations below X_0 and the remaining $m_i = n_i - d_i$ samples have the measured concentrations $X_{i1},X_{i2},\ldots,X_{im_i}$. Assuming

the X_{ij} is an observation from the log normal distribution with mean $\eta_i = \text{Exp}\ \{\mu_i + \sigma^2/2\}$ and variance $\beta_i = \eta_i^2\ \{\text{Exp}(\sigma^2) - 1\}$, then $y_{ij} = \ln X_{ij}$ is normal with mean μ_i and variance σ^2. The aim is to test the equality of the means at the different stations, i.e., the null hypothesis is

$$H_0: \ \eta_i = \eta \qquad \text{for all } i = 1,2,\ldots k.$$

This is equivalent to testing the equality of $\mu_1,\ldots,\mu_k$.

The likelihood function for $\mu_1, \mu_2,\ldots,\mu_k$ and σ^2 is given by

$$L_1 \alpha \prod_{i=1}^{k} \Phi^{d_i}(\xi_i)\ \sigma^{-m_i}\ \text{Exp}\ \{-\frac{m_i\ (s_i^2 + (\bar{y}_i - \mu_i)^2)}{2\sigma^2}\}, \qquad (1)$$

where

$$s_i^2 = \frac{1}{m_i} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2,\ \bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij},$$

$$\xi_i = (y_0 - \mu_i)/\sigma,\ \Phi\ (\xi_i) = \int_{-\infty}^{\xi_i} \phi\ (\xi_i)\ d\xi_i,$$

and

$$\phi(\xi_i) = \frac{1}{\sqrt{2\Pi}}\ \text{Exp}\ (-\xi_i^2/2).$$

The maximum likelihood estimates (MLE) $\hat{\mu}_i$ and $\hat{\sigma}^2$ satisfy the equations

$$\hat{\sigma}^2 = \sum_{i=1}^{k} m_i\{s_i^2 + (\bar{y}_i - \hat{\mu}_i)\ (\bar{y}_i - y_0)\}/M, \tag{2}$$

and $$\hat{\mu}_i = \bar{y}_i - \frac{d_i\hat{\sigma}}{m_i}\, g\ (\hat{\xi}_i)\ (i = 1,2,\ldots,k), \tag{3}$$

where

$$M = \sum_{i=1}^{k} m_i, \qquad g\ (\xi_i) = \phi\ (\xi_i)/\Phi\ (\xi_i),$$

$$\hat{\xi}_i = (y_0 - \hat{\mu}_i)/\hat{\sigma}, \text{ and } y_0 = \ln X_0.$$

The solution of the above system of equations for $\hat{\sigma}$ and $\hat{\mu}_i$ is not very simple due to the non-linearity of $g(\hat{\xi}_i)$. Tiku's (1967) approximation for $g(\hat{\xi}_i)$ results in obtaining explicit estimates for $\hat{\mu}_i$ and $\hat{\sigma}^2$ which have the same asymptotic efficiencies as those estimates obtained by the method of maximum likelihood. The approximate expression for $g(\hat{\xi}_i)$ is $a_i + b_i\ \hat{\xi}_i$, where

$$b_i = \{g\ (\hat{\xi}_{iu}) - g(\hat{\xi}_{il})\}/(\hat{\xi}_{iu} - \hat{\xi}_{il}),$$

$$a_i = g\ (\hat{\xi}_{il}) - \hat{\xi}_{il}\ b_i,$$

$$\hat{\xi}_{iu} = \Phi^{-1}\ (P_i + \sqrt{\frac{P_i(1-P_i)}{n_i}}\ )$$

$$\hat{\xi}_{il} = \Phi^{-1}\ (P_i - \sqrt{\frac{P_i(1-P_i)}{n_i}}\ )$$

and $P_i = d_i/n_i$.

Using this approximation results in estimating μ_i by

$$\hat{\mu}_i = (\bar{y}_i - h_i b_i y_0 - h_i a_i \hat{\sigma})/(1 - h_i b_i), \tag{4}$$

where $h_i = d_i/m_i$. Substituting (4) into (2) results in a quadratic equation for $\hat{\sigma}$. The positive root of the equation gives the estimate for σ which is

$$\hat{\sigma} = -\frac{1}{2} B + \frac{1}{2} \sqrt{B^2 + 4C} \tag{5}$$

where $B = \sum \{m_i (y_0 - \bar{y}_i) h_i a_i / M (1 - h_i b_i)\}$,

and $$C = \sum \frac{m_i}{M} \{s_i^2 - \frac{h_i b_i (y_0 - \bar{y}_i)^2}{1 - h_i b_i}\}$$

The same procedure can be used to estimate μ and σ under the null hypothesis H_0. Specifically, these estimates $\tilde{\mu}$ and $\tilde{\sigma}$ are given by

$$\tilde{\mu} = (\bar{y} - hb\, y_0 - h\, a\, \tilde{\sigma})/(1-hb) \tag{6}$$

and $$\tilde{\sigma} = -\frac{1}{2} \tilde{B} + \frac{1}{2} \sqrt{\tilde{B}^2 + 4\, \tilde{C}}, \tag{7}$$

where $\bar{y} = \sum m_i y_i/M$, $s^2 = \sum m_i s_i^2/M$

$$\tilde{B} = (y_0 - \bar{y})\, h\, a/(1-hb),$$

$$\tilde{C} = s^2 - \frac{hb(y_0 - \bar{y})^2}{1-hb},$$

$h = \sum d_i/M$, $b = (g(\tilde{\xi}_u) - g(\tilde{\xi}_1))/(\tilde{\xi}_u - \tilde{\xi}_1)$,

$a = g(\tilde{\xi}_1) - b\,\tilde{\xi}_1$, $\tilde{\xi}_u = \Phi^{-1}\left(P - \sqrt{\frac{P(1-P)}{n}}\right)$,

$\xi_1 = \Phi^{-1}\left(P + \sqrt{\frac{P(1-P)}{n}}\right)$ and $P = \sum d_i/n$.

The likelihood ratio test for testing H_0 is given by

$$\Lambda = 2\left[\sum d_i \ln\{\Phi(\hat{\xi}_i)/\Phi(\tilde{\xi}_i)\} + M \ln(\tilde{\sigma}/\hat{\sigma}) - \sum m_i\left\{\frac{(\bar{y}_i - \hat{\mu}_i)(y_0 - \hat{\mu}_i)}{2\hat{\sigma}^2}\right\}\right.$$

$$\left. + M\,\frac{(\bar{y} - \tilde{\mu})(y_0 - \tilde{\mu})}{2\tilde{\sigma}^2}\right]. \qquad (8)$$

The distribution of Λ is approximately Chi-square (χ^2) with k-1 degrees of freedom.

2. Inferences About the Lognormal Parameters

The inverse of the observed Fisher's information matrix I provides a general method for deriving the variances and covariances of the estimated parameters. The variance-covariance matrix V is given by

$$V = I^{-1} = -\begin{bmatrix} \frac{\partial^2 \log L}{\partial \mu^2} & \frac{\partial^2 \log L}{\partial \mu \partial \sigma} \\ \frac{\partial^2 \log L}{\partial \mu \partial \sigma} & \frac{\partial^2 \log L}{\partial \mu^2} \end{bmatrix}^{-1}_{\hat{\mu},\ \hat{\sigma}}$$

$$= \begin{bmatrix} \mathrm{Var}(\hat{\mu}) & \mathrm{Cov}(\hat{\mu},\hat{\sigma}) \\ \mathrm{Cov}(\hat{\mu},\hat{\sigma}) & \mathrm{Var}(\hat{\sigma}) \end{bmatrix} \qquad (9)$$

where $\frac{\partial^2 \log L}{\partial \mu^2}$ is the second derivative of log L with respect to μ the other elements of the matrix I are defined similarly. Also $\mathrm{Var}(\hat{\mu})$, $\mathrm{Var}(\hat{\sigma})$ and $\mathrm{Cov}(\hat{\mu},\hat{\sigma})$ represent the variances and covariances of the estimates. Hence confidence limits for $\hat{\mu}$ and $\hat{\sigma}$ either jointly or individually can be obtained in a standard way. The expressions for the elements of I are given as follows:

$$\frac{\partial^2 \log L}{\partial \mu^2} = -\frac{M}{\sigma^2}(1-hb);\ \frac{\partial^2 \log L}{\partial \mu \partial \sigma} = -\frac{2M}{\sigma^2}\left\{\frac{(\bar{y}-\mu)-hb(y_0-\mu)}{\sigma} - \frac{ha}{2}\right\}$$

and

$$\frac{\partial^2 \log L}{\partial \sigma^2} = \frac{M}{\sigma^2}\left\{1 - 3\frac{(s^2 + (\bar{y}-\mu)^2)}{\sigma^2} + \frac{3hb(y_0-\mu)^2}{\sigma^2} + 2ha\frac{(y_0-\mu)}{\sigma}\right\}$$

Furthermore, El-Shaarawi (1989) used the above results to derive an approximate confidence limits for the lognormal mean $\eta = \mathrm{Exp}\{\mu + \frac{1}{2}\sigma^2\}$ which is given as

$$\hat{\eta}\ \text{Exp}\ \{-Q\} \leq \eta \leq \hat{\eta}\ \text{Exp}\{Q\}, \qquad (10)$$

where $Q = Z_{1-\alpha/2} \sqrt{\text{Var}(\hat{\mu}) + 2\ \text{Cov}(\hat{\mu},\hat{\sigma})\hat{\sigma} + \text{Var}(\hat{\sigma})\ \hat{\sigma}^2}$

$Z_{1-\alpha/2}$ is the endpoint of the normal distribution corresponding to the appropriate confidence coefficient and $\hat{\eta} = \text{Exp}\ \{\hat{\mu} + \frac{1}{2}\ \hat{\sigma}^2\}$

It should be clear that in order to carry out the above analysis, at least one observation should exceed the level of detection in the sample.

APPLICATIONS

Figure 1 shows a map of the Niagara River and the locations of the biomonitoring sampling stations. Total polychlorinated biphenyls (PCBs) concentrations are used to illustrate the applications of the methods presented in this paper to study the spatial and temporal variability of contaminants observed in mussels. PCBs constitute a major organochlorine input to Lake Ontario from the Niagara River and several PCB Aroclors were included in the NRTC's group I Chemicals of Concern list.

Table 1 suggests that: (1) the background station has concentrations below the level of detection ($X_0 = 100$) for all surveys; (2) the concentrations vary temporally, with variabilities for the first two surveys indicating high levels of contamination while the other surveys indicate lower levels. In fact, all the data of the 5th survey are below the level of detection; (3) the degree of spatial variability depends on the survey, but generally the concentrations are lower at the lower and upper sections of the river and higher at the middle section; and (4) there are very high levels of variability among the replicates.

Table 2 gives the likelihood ratio test statistics for evaluating the differences between stations for each survey, with the exception of the 5th survey where all the values were below X_0. For each survey, stations with all values below X_0 were not included in computing the test statistic. Only the first two surveys showed highly significant ($P<0.01$) spatial variability among the stations included in the test.

Table 3A gives the estimates of the means at each station included in the computation. Station 7 shows a consistent, high level of PCBs contamination, while stations with erratic patterns such as station 17 may be indicative of an erratic contamination source. Table 3B gives the estimates of the means under H_0 and the estimates of σ under both H and H_0. It is clear that a major proportion of variability is due to the spatial variabilities. For example, the estimate of σ from the first survey has decreased from 0.878 to 0.456

when H_0 is correct and false respectively, which indicates that about half the variability can be attributed to spatial differences.

Finally, Table 3C gives the estimates of the lognormal means and their associated confidence limits. Testing the equality of the two lognormal means from survey 1 and 2 using El-Shaarawi's (1989) method indicates no significant difference between the two surveys. Furthermore, when the detection limit X_0 is not included within the confidence limits, then it is likely that a station with all the values below X_0 is significantly different from the stations used in generating the limits.

CONCLUSIONS

This paper extends the work of El-Shaarawi (1989) and El-Shaarawi and Dolan (1989) to the case of comparing the means of several populations. For illustration, the methods presented were applied to compare the spatial differences of total PCBs in a biomonitor between sampling locations and surveys along the Niagara River. The methods can be applied when the data are assumed to be generated from the normal and lognormal models.

REFERENCES

El-Shaarawi, A.H.: 1989, 'Inferences about the mean from censored water quality data'. Water Resources Research, V25, N4, 685-690.

El-Shaarawi, A.H. and Dolan, D.M: 1989, 'Maximum likelihood estimation of water quality concentrations from censored data'. Can. J. Fish. Aquat. (in press).

Gilliom, R.J. and Helsel, D.R.: 1986, 'Estimation of distributional parameters for censored trace level water quality data 1. Estimation techniques'. Water Resources Research, 22(2), 135-146.

Helsel, D.R.: 1986, 'Estimation of distributional parameters for censored water quality data'. Developments in Water Science (ed. A.H. El-Shaarawi and R.E. Kwiatkowski), v. 27.

Kauss, P.B. and Hamdy, Y.S.: 1985, 'Biological monitoring of organochlorine contaminants in the St. Clair and Detroit Rivers using introduced clams, Elliptio complanatus'. J. Great Lakes Res., 11, 247.

Tiku, M.L.: 1967, 'Estimating the mean and standard deviation from a censored normal sample'. Biometrika, 54(1), 155-158.

Table 1. Total PCBs, Concentrations ($\mu g \cdot kg^{-1}$, wet weight) at Different Sampling Stations and During the Six Surveys

Station	Survey #1			Survey #2			Survey #3			Survey #4			Survey #5			Survey #6		
	Replications			Replications			Replications			Replications			Replications			Replications		
0	N*	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
1																		
2	N	N	N	N	N	N	N	N	N	N	N	N						
3	N	N	N	N	115.0	380.0	N	N	555.2	228.0	181.7	N	N	N	N	N	N	N
4	N	N	N	N	N	N	N	N	N									
5				N	N	N												
6				500.0	280.0	375.0												
7	981.0	2136.0	814.3	1380.0	1800.0	2500.0	N	5185.0	4215.0	493.0	193.8	590.5	N	N	N	N	N	1198.0
8	N	N	N				N	N	N	N	N	N	N	N	N	N	N	N
9	201.6	195.2	N	130.0	150.0	150.0	N	N	N	266.7	N	N	N	N	N	N	N	N
10	584.1	500.0	473.0	550.0	530.0	940.0	465.1	N	N	N	N	330.0	N	N	N	N	N	N
11	166.7	339.7	285.6	455.0	545.0	525.0	N	N	N	350.0	668.6	455.6				589.2	445.7	551.0
12	415.0	923.2	817.5	435.0	435.0	N	349.2	N	N	455.6	404.9	205.9	N	N	N	782.3	197.6	599.2
13	302.5	465.0	411.9	110.0	130.0	125.0	N	N	N	223.3	452.9	N						
14	N	286.8	301.3	140.0	115.0	210.0	N	N	N	N	N	N	N	N	N	N	N	N
15				115.0	100.0	170.0												
16				N	120.0	115.0												
17	283.2	N	252.2	N	N	N	N	N	N	N	N	N	N	N	N	N	423.7	N
18				N	N	N										N	N	625.1
19	N	294.0	287.3	N	N	N	N	372.5	259.4	283.8	N	N	N	N	N	N	1191.0	N
20	N	174.6	285.2	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
21	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
22																N	N	N
27																N	N	N

N* indicates value below detection limit
Blank indicates no data available

Table 2. Likelihood Ratio Tests for all Surveys Except the Fifth.

Survey Number	Degrees of Freedom	Likelihood Ratio Test Statistic
1	9	30.76**
2	10	62.52**
3	4	5.42
4	7	8.28
6	5	4.62

**significant at the 1% level

Table 3. Maximum Likelihood Estimates of the Parameters

Table 3A. Means under H

Station	Survey #1	Survey #2	Survey #3	Survey #4	Survey #6
0					
1					
2					
3		5.041	4.192	4.971	
4					
5					
6		5.925			
7	7.086	7.516	6.962	5.949	5.198
8					
9	4.966	4.963		4.669	
10	6.248	6.476	4.101	5.799	
11	5.533	6.228		6.161	6.263
12	6.521	5.576	3.952	5.818	6.115
13	5.958	4.799		5.206	
14	5.254	5.011			
15		4.829			
16		4.600			
17	5.184				6.049
18					6.438
19	5.245		4.984	5.648	7.083
20	5.052				
21					
22					
27					

Table 3B. Standard deviations under H and H_0 and the mean under H_0

Parameter	Survey #1	Survey #2	Survey #3	Survey #4	Survey #6
Mean (H_0)	5.674 (0.164)	5.511 (0.172)	4.658 (0.802)	5.222 (0.250)	5.095 (0.440)
S.D. (H_0)	0.878 (0.131)	0.980 (0.129)	2.137 (0.731)	0.933 (0.183)	1.589 (0.406)
S.D. (H)	0.456	0.338	1.477	0.504	0.786

S.D. = standard deviation
Values in brackets are standard errors of the estimates

Table 3C. Means and confidence limits (measurement units)

	Survey #1	Survey #2	Survey #3	Survey #4	Survey #6
Mean	428.150	399.924	1636.409	286.352	576.811
Confidence limits (95%)	294.725-621.978	265.733-601.88	81.331-32924.960	181.663-451.371	169.588-1961.886

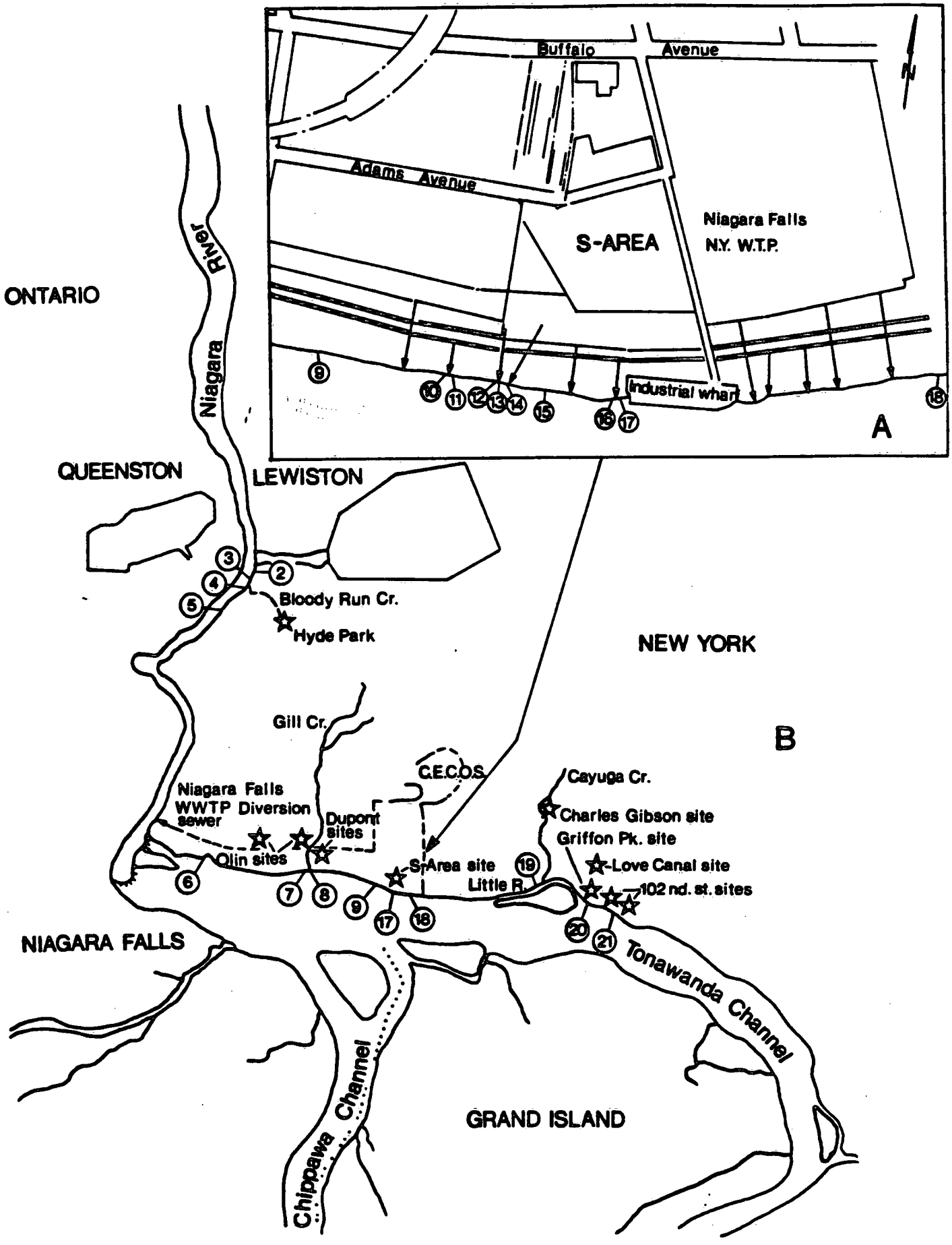

Figure 1. Locations of the biomonitoring sampling stations.