Impact of non-detects in water quality data on estimation of constituent mass loading

M. Kayhanian*, A. Singh and S. Meyer****

* Center for Environmental and Water Resources Engineering, Department of Civil and Environmental Engineering, University of California, Davis, CA 95616, USA

** Office of Water Program, Department of Civil Engineering, California State University, Sacramento, CA 95819, USA

Abstract Often, fractions of stormwater constituents are not detected above laboratory reporting limits and are reported as non-detect (ND), or censored data. Analysts and stormwater modelers represent these NDs in stormwater data sets using a variety of methods. Application of these different methods results in different estimates of constituent mean concentrations that will, in turn, affect mass loading computations. In this paper, different methods of data analysis were introduced to determine constituent mean concentrations from water quality datasets that include ND values. Depending on the number of NDs and the method of data analysis, differences ranging from 1 to 70 percent have been observed in mean values. Differences in mean values were, as shown by simulation, found to have significant impacts on estimations of constituent mass loading.

Keywords Mass loading; mean; non-detect; pollutant constituent; statistical method; water quality

Introduction

With the passage of the Clean Water Act (CWA) in 1972, and an amendment in 1987, the federal government placed a priority on restoring and maintaining the physical, chemical, and biological integrity of the Nation's receiving waters. The CWA prohibits point or non-point source discharge pollutants into waters of the U.S. without a National Pollutant Discharge Elimination System (NPDES) permit. Many municipalities and state organizations are now regularly monitoring storm water to comply with NPDES requirements. The California Department of Transportation (Caltrans) has begun a comprehensive research and monitoring program to evaluate the environmental effects of storm water runoff from their facilities. As part of this ongoing effort, numerous water quality constituents have been analyzed and a large volume of data has been collected for modeling and mass loading estimation.

Due to the variability of the monitoring sites and other environmental conditions a fraction of the water quality constituents fell below the detection limits (DL) and are therefore reported as non-detect (ND). Such a data set is said to be censored. Censored data are actually missing values. Analytical data that contain large numbers of non-detect data cannot simply be ignored. If a few values are missing, we could probably find some way to fill in the missing values without distorting the pattern of the series. The difficulty with censored data is that they are not missing in a random pattern, but they are all missing at one end of the distribution range. We cannot proceed as if they never existed, or pretend that they are zero.

Various methods of analysis can be used to deal with these types of data (Helsel and Gilliom, 1986; Newman *et al.*, 1989; Mac Berthouex and Brown, 1994; She, 1997; Clarke, 1998; Shumway *et al.*, 2001). Applying these methods will result in different estimates of constituent mean concentrations that will, in turn, affect mass loading computations. The focus of this paper is to simulate the impact of mean value, computed from censored data, on constituent mass loading prediction.

Water Science and Technology **Vol 45 No 9 pp 219–225** © IWA Publishing 2002

Methods

Water quality data

Water quality data presented in this paper are from the Caltrans highway stormwater runoff characterization study. Over 50 highway sites were monitored during the past three years (Kayhanian *et al*., 2001). In general, composite samples were collected by mixing some number of individual sample aliquots based on flow rate throughout the storm event. Therefore, the pollutant concentrations for each storm event are considered to be event mean concentrations (EMC). Automated composite sampling methods were used, except for oil and grease, petroleum hydrocarbons, and bacteria. Flow rates were measured with automated flow meters using area velocity, bubbler, pressure transducer, or ultrasonic sensor measurements. Precipitation was measured using electronic "tipping bucket" rain gauges.

Composite samples were analyzed according to standards specified by the U.S. Environmental Protection Agency (USEPA). Analytical data were reviewed to ensure quality assurance (QA) and quality control (QC). QA/QC parameters that were reviewed include: reporting limits, holding times, contaminations check results, precision, and accuracy analysis results. The validated data were reported in Excel spreadsheets based on specifications established in Caltrans Data Reporting Protocol (Caltrans Environmental Program, 2000). Once the Excel spreadsheets were reported to Caltrans, data were imported into an Access database. The water quality database was queried to extract files that contained analytical results for further analysis and evaluation.

Data analysis methodology

The following five statistical methods were used to estimate mean concentrations of censored water quality data: (1) conventional, (2) Cohen's maximum likelihood estimation (MLE), (3) maximum likelihood estimation by delta and bootstrap methods, (4) regression on order statistics (ROS), and (5) EPA delta lognormal statistics method.

Conventional. In this method all non-detected (ND) values are substituted with arbitrary values. Two simple substitution methods that are commonly used include: (1) substituting ND with reporting limits, and (2) substituting ND with one-half of the reporting limit. Conventional statistical analysis proceeded using these modified values.

Cohen's maximum likelihood estimation. This method provides adjusted estimates of the sample mean and standard deviation that account for data below the detection limit. The adjusted estimates are based on the statistical technique of maximum likelihood estimation of the mean and variance (Cohen, 1959; USEPA, 1996). Under this method, crude estimates of the mean and variance are first estimated using the non-censored data. These estimates are then adjusted using the factor λ. Cohen (1961) provided tables to estimate the value of λ. Estimation of the value of λ usually requires data interpolation. Hass and Scheff (1990) developed an empirical equation that estimates the value of λ to within 6 percent relative error of the tabulated values (Berthouex and Brown, 1994). For the purpose of this paper, computation of λ for the adjustment of the mean was estimated using the Hass and Scheff (1990) empirical equation.

Maximum likelihood estimation by delta and bootstrap methods. A procedure was developed for computing confidence limits for the mean using large-sample theory for maximum likelihood estimators by the delta method and the non-parametric bootstrap method (Shumway et al., 1989; Shumway et al., 2001). Under these methods, a more general class of power transformation due to Box and Cox (1964) is applied to the data, which includes the logarithm, no transformation, and various other power laws as special cases. This procedure leads to MLEs for the means in the original scale, but it does not immediately produce an estimate for variance or a confidence interval. With a confidence interval, one can make an assessment of the probable range within which the true mean can be expected. The confidence interval is predicted by the delta method for large sample sizes or by the bootstrap method for smaller sample sizes.

Regression on order statistics (ROS). Regression on order statistics (ROS) is based on the modified probability plotting that was developed by Helsel (1990) and Helsel and Cohn (1988). This method is also known as Helsel's robust method. Helsel's approach fits a regression line to the log transformed observation values above the RL and their corresponding z scores. Next, the regression is used to predict "fill in" values for the RL observations. All values, including the predicted "fill in" values, are then back transformed to arithmetic units. The mean and standard deviations are estimated using the data set that now includes "fill in" values for the censored observation.

EPA delta lognormal statistics method. The USEPA Delta lognormal statistical technique was mainly developed for regulatory use (USEPA, 1991). This method assumes a lognormal distribution and the mean and standard deviations are computed using modified equations accounting for non-detects.

Results

Storm water runoff characteristics

Over 300 constituents in broad categories of conventionals, nutrients, microbiological, metals, major ions, volatile organic compounds (VOCs), semi-volatile organic compounds (SVOCs) including polycyclic aromatic hydrocarbons (PAHs), pesticides and petroleum hydrocarbons have been analyzed for the past three years. These data were collected from different Caltrans facilities such as freeways, maintenance stations, park and ride lots, and construction sites. Results obtained to date revealed that most of the VOCs, SVOCs, PAHs, pesticides, and metals such as silver, mercury, selenium and titanium were reported below detection limits. For constituents (based on broad categories) that were reported above the detection limits, on average percent detected values ranged from 19 to 92 percent (see Figure 1).

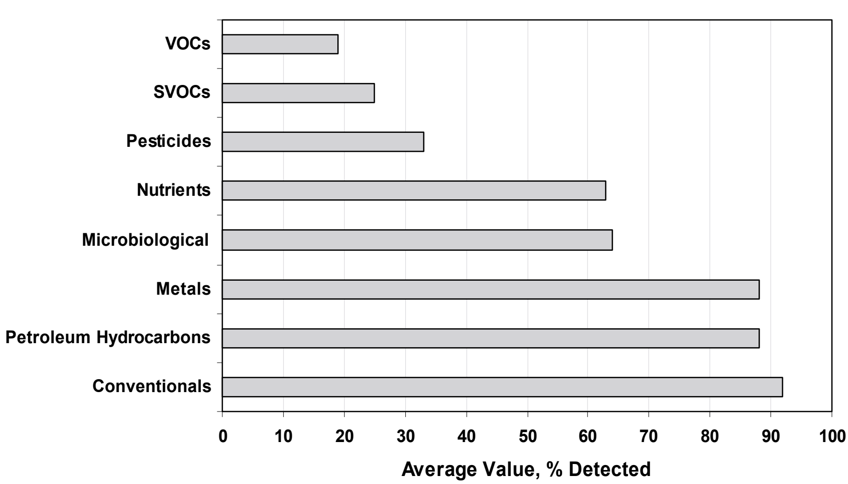

Figure 1 Average percent detect values for different constituent categories

Data for some of the constituents found in highway runoff were analyzed using five different statistical methods. Mean concentrations of selected constituents with percent non-detect ranging from about 2 to 88 percent are summarized in Table 1.

Discussion

Influence of detection limit

The number of non-detects can impact the estimation of mean value as illustrated in this section. Consider the water quality data for total Ni shown in Table 2.

As can be seen, the number of detected values in this data set is 100 percent, since the detection limit was set at an extremely low detection limit of 0.5 µg/L. However, if analytical tests were conducted at detection limits of 1, 2, 3, 4, or 5 µg/L, there would have been significant number of non-detects. For example, if the detection limit was set at 5 µg/L, all numbers equal to or less than 5 µg/L would have been reported as non-detect, which would amount to about 54 percent of total data points. Applying these non-detects as part of the data set, some difference in mean estimation would be expected. Considering conventional, Cohen's, ROS, MLE and EPA Delta Log methods for analysis of these data sets will result in mean concentrations as illustrated in Figure 2. The impact of these different mean values on load estimation is simulated below.

Table1 Constituent mean concentrations based on different methods of data analysis

				Statistical analysis				EPA	
Constituent	**Unit**	**n**	**% ND**	**ND=DL**	**ND=DL/2**	**ROS**	**MLE**	**Cohen's**	**delta log**
Chemical Oxygen Demand (COD)	mg/L	65	1.5	121.45	121.37	121.6	122.1	120.86	123.52
Aluminium-Dissolved	µg/L	25	20	191.16	188.7	187.4	138.8	105.9	140.7
Arsenic-Dissolved	µg/L	42	38.1	2.04	2.03	1.94	1.94	1.38	2.04
Arsenic-Total	µg/L	46	28.3	4.25	4.25	4.18	4.08	3.27	4.29
Cadmium-Dissolved	µg/L	450	87.8	0.58	0.36	N/A[a]	0.30	N/A[b]	0.58
Cadmium-Total	µg/L	373	30.8	1.30	1.23	1.25	1.26	0.94	1.28
Chromium-Dissolved	µg/L	462	45.7	2.62	2.57	2.15	2.49	2.77	1.15
Chromium-Total	µg/L	383	5.2	11.36	11.35	11.34	11.32	11.03	11.14
Nickel-Dissolved	µg/L	461	38.6	4.24	3.85	4.01	4.12	2.58	4.14
Nickel-Total	µg/L	383	4.2	13.72	13.68	13.69	12.79	13.06	13.06
Lead-Dissolved	µg/L	523	31.2	6.04	5.89	5.90	5.91	1.75	5.25
Silver-Total	µg/L	36	72.2	3.88	3.70	3.56	4.01	N/A[b]	4.18
Nitrite-N	mg/L	383	46.2	0.29	0.27	0.26	0.24	N/A[b]	0.25
Ortho-P	mg/L	199	39.7	0.16	0.16	0.16	0.19	0.17	0.10
Phosphorus-Total	mg/L	602	40.7	0.30	0.29	0.30	0.29	N/A[b]	0.28
Oil & Grease	mg/L	428	54.9	11.29	9.92	10.00	10.11	N/A[b]	10.99

[a] N/A = not analyzed. This method will not be used for data sets that contain more than 80 percent non-detect
[b] N/A = not analyzed. This method will not be used for data sets that contain more than 40 percent non-detect

Table 2 Event mean concentration of total Nickel with detection limit of 0.5 µg/L

Constituent	**Detection limit**	**Event mean concentration (µg/L)**									
		47	0.97	13	6.8	6.5	0.51	2.5	2.5	35	14
		1.8	1.6	12	7.0	6.0	0.52	2.6	2.3	2.9	3.2
		1.8	1.4	9.9	7.1	5.8	0.55	2.7	2.3	3.0	3.2
		1.7	1.4	9.7	7.2	5.6	0.63	38	2.2	2.1	3.3
Nickel, Ni	0.5 (µg/L)	1.6	8.4	9.4	7.2	5.4	26	45	23	2.1	3.4
		11	8.5	8.8	7.4	5.2	27	0.98	22	2.1	3.5
		11	2.7	8.6	7.8	5.0	30	1.1	19	16	3.6
		10	2.8	1.3	8.2	4.8	32	2.2	17	15	3.9
		0.96	2.8	1.3	2.9	4.6	0.74	4.2	2.2	14	4.0
		4.0	1.3	34	2.9	4.4	0.85	24	2.1	13	50

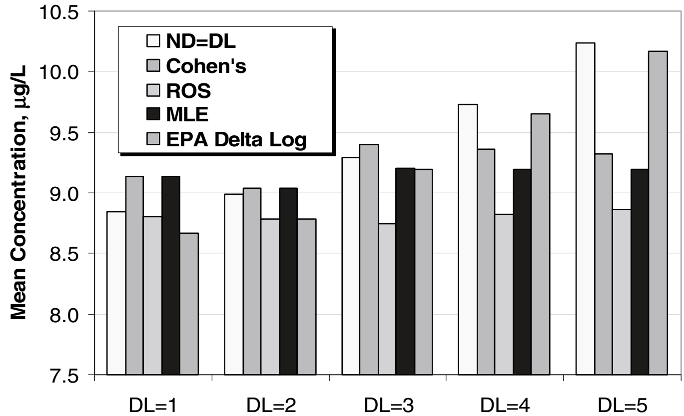

Figure 2 Influence of detection limit on Nickel mean concentration using different method of analysis

Constituent mass loading simulation

To illustrate the impact of NDs on pollutant mass loading estimation, we considered the San Diego watershed (see Figure 3) as an example. Mass loading of those constituents listed in Table 1 is estimated for highways within the San Diego River watershed as summarized in Table 3. The mass loading was estimated by a "Rational Method" model (RMM) using the following mathematical relationship:

$$L_{i,k} = \Sigma \, V_{j,k} \, C_{i,j}$$

where:

$L_{i,k}$ = annual storm water load from constituent i in area k

$V_{j,k}$ = annual runoff volume from land use j in area k, m^3

$C_{i,j}$ = mean EMC of constituent i in runoff from land use j, mg/L

Analyses of loads presented in Table 3 suggest that annual loads can vary significantly depending on the method of analysis used. Various methods may result in values that have differences between less than one percent to over 70 percent.

Implication on total maximum daily load

To further display the effects of NDs on pollutant mass loadings four hypothetical total maximum daily loads (TMDLs) were assumed for the San Diego River; arsenic 25 g/day,

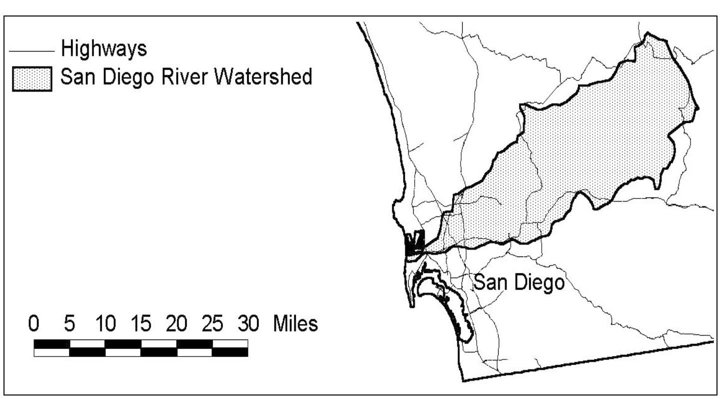

Figure 3 San Diego River Watershed

Table 3 Annual Storm Water Loads based on different methods of data analysis

Constituent	Unit	Statistical analysis				
		ND=DL	ROS	MLE	Cohen's	EPA delta log
Chemical Oxygen Demand	kg/yr	745739	746660	749730	742116	758449
Aluminium-Dissolved	kg/yr	1174	1151	852	650	864
Arsenic-Dissolved	g/yr	12526	11912	11912	8474	12526
Arsenic-Total	g/yr	26096	25666	25052	20079	26342
Cadmium-Dissolved	g/yr	3561	N/A	1842	N/A	3561
Cadmium-Total	g/yr	7982	7675	7737	5772	7860
Chromium-Dissolved	g/yr	16088	13202	15289	17009	7061
Chromium-Total	g/yr	69754	69631	69508	67727	68403
Nickel-Dissolved	g/yr	26035	24623	25298	15842	25421
Nickel-Total	g/yr	84245	84061	78534	80192	80192
Lead-Dissolved	g/yr	37087	36228	36289	10746	32237
Silver-Total	g/yr	23824	21859	24623	N/A	25666
Nitrite-N	kg/yr	1781	1596	1474	N/A	1535
Ortho-P	kg/yr	982	982	1167	1044	614
Phosphorus-Total	kg/yr	1842	1842	1781	N/A	1719
Oil & Grease	kg/yr	69324	61403	62078	N/A	67482

chromium 40 g/day, nickel 70 g/day, and lead 90 g/day. These TMDLs were compared to the data in Table 3 to show how the four methods produce different results (see Figure 4).

As shown, the choice of statistical analysis may have a large impact on the results of a study. These variances in the results may be critical, especially in a regulatory environment.

Conclusions

The following conclusions can be drawn from the present study.

1. A significant number of constituents monitored as part of the Caltrans storm water runoff characterization can contain large numbers of non-detects.
2. Detection limits set by analytical laboratories can affect the number of non-detects in water quality data.
3. Many different statistical methods are available to estimate constituent mean values. Depending on data distribution and the numbers of non-detects in a data set, large variations in mean value can be observed. These variations in mean values have shown to significantly affect the constituent mass loading estimations.

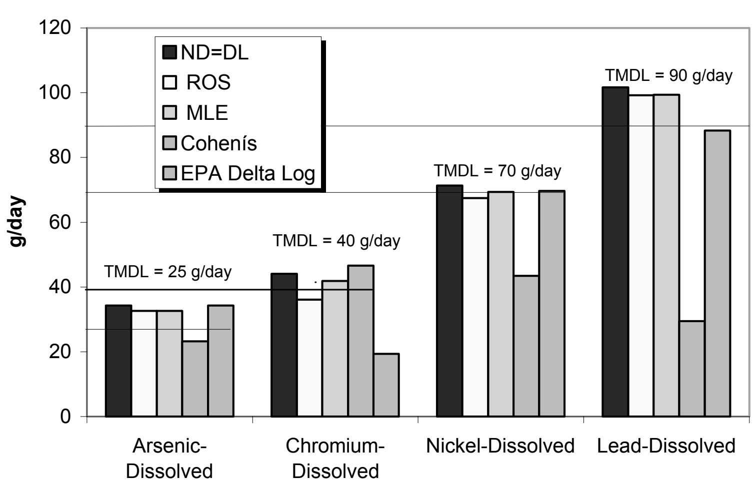

Figure 4 Impact of data analysis on hypothetical TMDLs in the San Diego River

4. The statistical approach in analyzing the water quality data with non-detects may affect the outcome of TMDLs or other regulatory requirements.

Acknowledgements

The authors greatly acknowledge the support and encouragement provided by Mr. Steve Borroum, director of the Caltrans Stormwater Management Program for the preparation of this manuscript.

References

Box, G.E.P. and Cox, D.R. (1964). An Analysis of Transformations (with discussion). *J. Royal Statist. Soc*. **B** (39), 211–252.

Caltrans Environmental Program (2000). *Caltrans Data Reporting Protocol*, California Department of Transportation, Report No. CTSW-TM-00-002, Sacramento, CA.

Clarke, U.J. (1998). Evaluation of Censored Data Methods to Allow Statistical Comparisons among Very Small Samples with Below Detection Limit Observations, *Enviro. Sci. Technol.* **32**(1), 177–183.

Cohen, A.J., Jr. (1959). Simplified estimators for the normal distribution when samples are singly censored or truncated, *Technometrics* **1**(3), 217–237.

El-Shaarawi, H.A. (1989). Inferences About the Mean from Censored Water quality Data, *Water Resource Research* **25**(4), 685–690.

Gilliom, J.R. and Helsel, D.R. (1986). Estimation of Distributional Parameter for censored Trace Level Water Quality Data 1. Estimation Techniques, *Water Resources Research* **22**(2), 135–146.

Hass, N.C. and Scheff, A.P. (1990). Estimation of Averages in Truncated Samples, *Environ. Sci. Technol.* **24**(6), 912–919.

Helsel, D.R. and Gilliom, J.R. (1986). Estimation of Distributional Parameter for censored Trace Level Water Quality Data 2. Verification and Applications, *Water Resources Research* **22**(2), 147–155.

Helsel, D.R. and Cohn, T.A. (1988). Estimation of Descriptive Statistics for Multiply Censored Water Quality Data. *Water Resources Research* **24**(12), 1997–2004.

Helsel, D.R. (1990). Less than Obvious, *Enviro. Sci. Technol.* **24**(12), 1767–1774.

Kayhanian, M., Johnson, J., Yamaguchi, H. and Borroum, S. (2001). Caltrans Storm Water Management Program. *Stormwater*, Vol. 2, No. 2.

Mac Berthouex, P. and Brown, C.L. (1994). *Statistics for Environmental Engineers*, Lewis Publishers, Boca Raton, Florida.

Newman, M.C., Dixon, P.M., Looney, D.B. and Pinder, J.E. III (1989). Estimating mean and variance for environmental samples with below detection limit observations. *Water Res. Bull*. **22**, 904–916.

She, N. (1997). Analyzing Censored Water Quality Data using a Non-parametric Approach, *Journal of American Water Resources Association* **33**(3), 615–624.

Shumway, H.R., Azari, S.A. and Johnson, P. (1989). Estimating mean concentrations under transformation for environmental data with detection limits, *Technometrics* **31**(3), 347–356.

Shumway, H.R., Azari, S.A. and Kayhanian, M. (2001). Statistical approaches to estimating mean water quality concentrations with detection limits, *Environmental Science & Technology (Submitted).*

US EPA (1991). *Estimating the mean of data sets that include measurement below the limit of detection*. NCASI Technical Bulletin No. 621.

US EPA (1996). *Guidance for Data Quality Assessment, Practical methods for Data Analysis*. EPA QA/G-9, QA96 Version. Report No. 600/R-96/084.