Uncertainty Reduction in Modeling of Chemical Load in Streams

Erechtchoukova, M. G.[1] and P.A. Khaiter[1]

[1]Atkinson Faculty of Liberal and Professional Studies, York University, Canada,
E-Mail: marina@yorku.ca

Keywords: *Chemical load, Uncertainty analysis, Water quality, Monitoring design*

EXTENDED ABSTRACT

Stream chemical load is an important indicator of water quality. It is employed extensively in assessment and prediction of the state of natural waterbodies as well as in watershed monitoring and management activities. Stream chemical load is the total mass of a chemical ingredient passing through the cross section over a given period of time. Chemical load estimators usually use values of water discharge and concentration of an ingredient determined from samples collected at a given cross section of a waterbody. Water discharge is observed often and normally values obtained with at least daily frequency are available. Series of concentrations of water ingredients are significantly shorter due to financial and technical constraints.

Algorithms for chemical load calculation are inferred from mathematical assumptions regarding the relationships between water discharge and concentrations or from statistical properties of data sets. It is important to find an estimator which describes the chemical load with lesser uncertainty than other known estimators.

The uncertainty of chemical load can be formulated based on the concept of "model uncertainty". Two main sources of load uncertainty are considered. The first one is associated with the uncertainty of observation data. The resulting uncertainty from this source is of the same order as the uncertainty of observation data. Another type of uncertainty which must also be taken into account relates to the fact that continuous values of water discharge and concentrations of water ingredients are described by discrete data. Statistically, uncertainty can be defined as the variance of the estimator. The uncertainty of the estimation depends on the mathematical properties of formulae employed and available data sets.

The investigation of different formulae suggests the construction of a load estimator based on ratio estimates and stratification of observation data. The efficiency of the stratified estimator certainly depends on a stratification scheme. The scheme which reduces resulting uncertainty can be achieved for the majority of natural streams with distinct hydrological seasons. The suggested formula can significantly reduce the uncertainty of annual ingredient loads in streams even with limited number of concentrations at hand, but its practical application requires data sets collected according to a stratified monitoring design.

The number of observations required to achieve a desirable level of uncertainty in the result can be obtained from the problem of mathematical programming and refined using a presented iterative algorithm. The algorithm was applied to calculation of annual load of chloride ions in the Vyatka River. Two types of stratification were considered: one with strata corresponding to main hydrological seasons when stratum boundaries vary from year to year and another one with fixed calendar dates for strata. While the former type of stratification requires smaller number of observations to keep the load uncertainty under a specified level, the latter one is more convenient in terms of practical implementation.

For an ingredient which is not measured automatically, the reduction of uncertainty in the estimation of its load requires a monitoring design consistent with the formula selected for the calculation. Under considered stratification schemes, suggested numbers of observations for estimation of chloride ions load with 5% uncertainty remains high. These numbers significantly exceed the number of observations supplied by many monitoring systems.

The monitoring design common for all observed water quality indicators and supporting evaluation of loads with the given uncertainty is not always attainable. For tiered monitoring systems, different allowable levels of uncertainty can be recommended for water quality indicators according to their importance for a particular site.

1. INTRODUCTION

Stream chemical load, also referred to as mass discharge or mass flux, is the total mass of a chemical ingredient passing through the stream cross section for a given period of time. Chemical load is an important hydrochemical indicator of a stream and is employed extensively in the assessment and prediction of natural waterbody conditions and watershed monitoring and management. Chemical load is a resulting measure of all processes within a watershed that affect concentrations of an investigated ingredient. Therefore, it can be used to evaluate self-purification ability of an aquatic ecosystem, accumulation of chemical compounds in natural waterbodies, to serve as an objective for a monitoring network (Hooper *et al.* 2001) or to calculate the total maximum daily amount of a pollutant that a waterbody can receive without exceeding water quality standards. Being an important indicator of the quality of natural waters, the chemical mass discharge must be evaluated as accurate as possible.

Load amounts are calculated according to different approaches. The overview of these approaches and their classification can be found, for example, in Aulenbach and Hooper (2006). Different formulae for load calculation are presented in Preston *et al.* (1989). Chemical load estimators usually use values of water discharge and concentrations of an ingredient determined from the samples collected at a particular cross section of a waterbody over a period of time. The data may come from various sources (Khaiter *et al*. 2000): systematic, such as routine monitoring systems, or occasional like pilot projects. In many cases data are not collected specifically for load estimation. While water discharge is observed often and normally values obtained with at least daily frequency are available, series of concentrations of water ingredients are significantly shorter due to financial and technical constraints.

Formulae or algorithms for mass discharge calculation are inferred from mathematical assumptions regarding the relationships between water discharge and concentrations or statistical properties of data sets. The accuracy of the results obviously depends on the number of observations and formulae utilized including such aspect as conformity of the formulae with available data. Selection of an appropriate approach and a formula to calculate mass discharge is an important step in chemical load estimation process. The aim of the selection is straightforward, namely, to find an estimator which describes chemical load in the best possible way. Usually it means that the estimator has minimal or at least lesser uncertainty than other known estimators.

The paper examines the way to reduce uncertainty in chemical load estimations based on the data collected by a routine monitoring system.

2. UNCERTAINTY OF CHEMICAL LOAD

Uncertainty of a load estimate is derived from the definition of model uncertainty (Campolongo *et al.* 2000). It denotes possible deviation of values calculated based on available data sets and selected approach from actual value.

Calculated values of an ingredient load always contain uncertainty due to several reasons. First of all, raw data used for calculation cannot be determined precisely. Usually, analytical methods for sample processing return results with up to 5% error. This uncertainty in observation data propagates to the load estimate. First-order uncertainty analysis allows for a conclusion that resulted load values contain the uncertainty of the same order as the uncertainty of observation data. Another type of uncertainty which must also be taken into account relates to the fact that continuous values of water discharge and concentrations of water ingredients are described based on discrete, and sometimes scarce, data. This type of uncertainty can be evaluated based on simple statistical methods.

Statistically, uncertainty can be defined as the variance of the estimator. In order to evaluate the uncertainty, it is necessary to calculate the mean value and then the variance of the estimator using an available set of observation data. The calculated variance obviously depends on the formula applied to the load calculation. Therefore, the estimator that dominates others must be selected.

The uncertainty of an estimator depends on the variability of an investigated constituent and the size of a data set. The larger the set, the lesser the uncertainty of values calculated based on the set. However, for the majority of important water quality indicators and many sampling sites, extensive observations are not possible due to the constraints mentioned above.

Selection of an estimator can be conducted based on a comparison of variance values calculated for different estimators using the same set of available observation data. Relatively low values of variance can be obtained only if data collection was implemented under the same assumptions that were used to build the estimator. Thus, an estimator and an observation data set are

interdependent. Robertson and Roerish (1999) demonstrated how sampling strategies affect load estimates for small rivers. Practical implication of understanding the sources of uncertainty consists in the necessity to select formula or method for chemical load calculation which dominates others on an available data set.

3. SELECTION OF AN ESTIMATOR

A formula for chemical load estimation can be easily derived from the load definition and for a particular ingredient it takes the form:

$$L = \int_0^T c(t)Q(t)dt\,, \qquad (1)$$

where L is the ingredient load, $c(t)$ is the concentration of the ingredient averaged over a given cross section, $Q(t)$ is water discharge at this cross section, T is the investigated period of time. The main obstacle to the application of formula (1) is that functions $c(t)$ and $Q(t)$ are unknown and the load must be approximated based on observation data. The task actually consists in constructing an approximation of the integral in formula (1) which delivers lesser uncertainty than the other known approximations on a given data set.

The product of average concentration and water discharge over the period of time gives instantaneous ingredient load which, being multiplied by the duration of the period, is an approximation of L. Such an approximation is mathematically correct if accurate values of average concentration of the ingredient and water discharge are at hand. It can be achieved if the total number of samples is large enough to describe dynamics of concentrations and water discharge. At the same time, resulting uncertainty of the approximation implemented using available data is usually very high due to high seasonal variations of concentrations and water discharge.

The reduction of uncertainty is possible by selecting an estimator which employs additional available data. As a rule, water discharge is observed daily or even with a twice-a-day frequency, while concentrations of water ingredients are measured relatively seldom. It implies application of either regression analysis or ratio estimates.

Regression models are probably more popular. They are used independently or as a part of the composite method (Aulenbach and Hooper 2006). Regression model describes the relationships between water discharge and concentrations of a water ingredient and is actually used to restore missing or predicted values of concentrations. It is assumed that the relationship is steady and the model is suitable for the forecasting.

The main restriction in application of regression analysis is the number of available samples. If the relationship between concentrations and water discharge is linear and concentrations are normally distributed, the rule of thumb requires at least 50 samples to evaluate the reliability of regression coefficients. This number increases significantly for other types of regression equations and water quality indicators with high variability.

The same hydrological seasons at different calendar years may exhibit different hydrological properties. Therefore, underlying hydrological and hydrochemical processes controlling concentrations of an ingredient can make different contributions from year to year. This fact also warns of applying the regression model beyond the period when water samples were collected.

Another commonly accepted statistical approach to load calculation uses ratio estimates. It allows for calculation of average water discharge based on all available data and instantaneous load values only when concentrations are measured. Strictly speaking, the ratio estimator has a bias and must be corrected. Although ratio estimators have minimal uncertainty if the relationship between two investigated variables is linear, there is no need to make such a strict assumption. Ratio estimators dominate estimators obtained by simple expansion if the following condition holds (Cochran 1963):

$$\rho(l,q) > \frac{1}{2}\frac{cv_q}{cv_l}\,, \qquad (2)$$

where $\rho(l, q)$ is the correlation coefficient and cv_q and cv_l are coefficients of variation of water discharge and instantaneous load, respectively, calculated using data when the ingredient concentrations are available. Erechtchoukova and Tsirkunov (1989) showed that condition (2) holds for many large and small streams and water quality indicators. In many cases, ratio estimators require smaller number of observations than regression methods.

A further reduction of resulting uncertainty can be achieved by stratifying available data sets according to main hydrological seasons (Bodo and Unny 1983). The stratified estimator of an ingredient load dominates non-stratified one, if the variance of the instantaneous load within the strata

is less than its variance between the strata. A more precise criterion can be found in Cochran (1963). The efficiency of a stratified estimator certainly depends on a stratification scheme applied. It is worth to note that stratification reducing load uncertainty can be achieved for the majority of natural streams with distinct hydrological seasons. Then the following formula provides a good estimation of the load:

$$L = T\sum_{j=1}^{k}\frac{N_j}{N}Q_j\left(\frac{l_j}{q}+B\right), \qquad (3)$$

where Q_j is the average water discharge in j-th stratum calculated based on the most frequent measurements, q_j is the average water discharge in j-th stratum calculated using values corresponding to the observed concentrations, l_j is the average instantaneous ingredient load calculated based on the observed concentrations, N is the total number of water discharge observations over the period, N_j is the number of water discharge observations in j-th stratum, T is duration of the investigated period, B is the bias of the ratio estimate. The bias can be calculated according to the following formula (Cochran 1963):

$$B = \frac{l_j}{n_j q_j}(cv(q_j) - c_{ql}) \qquad (4)$$

where $cv(q_j)$ is the coefficient of variation of water discharge in j-th stratum and c_{ql} is relative covariance of water discharge and instantaneous load in j-th stratum. Formula (4) can be rewritten in different ways to simplify calculations.

Formula (3) can significantly reduce the uncertainty of annual ingredient load in streams, but its practical application requires data sets that are collected according to a stratified monitoring design (Erechtchoukova and Tsirkunov 1989).

4. DATA SET REDUCING THE UNCERTAINTY

Evaluation of uncertainty of a chemical load estimate along with its obtained value is strongly desirable. The statistical interpretation of a load estimate and its uncertainty implies the dependence of both on an observation data set used in the estimation. Strictly speaking, the task of evaluation of load uncertainty using the data set at hand is an inverse problem. This problem does not have a formal solution, but a heuristic method can be introduced. It seems reasonable to determine the required number of observations in order to keep the uncertainty of a chemical load estimate under a given level. If the estimation of a data set size is known, the available data set can be compared against it. The comparison gives investigators an insight about the quality of the load evaluation.

For many streams, a stratified ratio estimator of chemical load dominates other statistical estimators under the assumption that detailed series of chemical concentrations and water discharge at a given cross section are available. The main requirement for a series of concentrations of an ingredient is the data set to adequately reflect its variability.

The required number of concentrations which keeps the uncertainty of load estimation under a given level can be found from the problem of mathematical programming (Bodo and Unny 1983):

$$\min\sum_{j=1}^{k} n_j, \text{ when } D(L) \le V, \qquad (5)$$

where n_j is a required number of concentration values per j-th stratum, V is the given level of uncertainty and $D(L)$ is the ingredient load variance which can be obtained from the following formula:

$$D(L) = T^2\sum_{j=1}^{k}\frac{N_j^2}{N^2}S_j^2 \quad , \qquad (6)$$

where S_j^2 is the variance of average instantaneous load estimator in j-th stratum which depends on the number of observations, n_j, conducted during this stratum, k is the total number of strata. Such formulation is possible due to consistency of estimator (3). A given level of uncertainty for the load estimate is calculated using a desired precision and the calculated value of the ingredient load under the assumption of normal distribution of statistics (3).

Problem (5) can be solved analytically if additional assumptions regarding statistical properties of the investigated variables are made (Bodo and Unny 1983). Even though the approach gives analytical formulae for n_j, obtained values are only an approximation of a required sample size.

Another way to solve problem (5) is in applying a numerical algorithm which is designed based on Lagrange multiplier method. To improve the approximation of a sampling size and distribution of observations between the strata, an iterative procedure was suggested. It was assumed that observations are evenly distributed within each

stratum. The first iteration uses the most complete series of concentrations. Such a series can be obtained via pilot sampling or interpolation in time between observed values. After the first iteration, a sample of suggested size and structure is drawn from the complete series of concentrations in order to re-evaluate the required number of observations and their distribution among the strata. To improve robustness of the algorithm, at each iteration several random samples (e.g., one hundred samples) are made and the average number of observations per each stratum is calculated. The iteration process stops when a suggested number of observations on the next iteration is the same or higher than the one used on the current step.

5. CASE STUDY

The Vyatka River is a large Eastern-European river with a length of 1,370 km and a watershed area of 129,000 km^2 located in Kirov Oblast and the Republic of Tatarstan in the Russian Federation (Fig. 1). Its average water discharge is about 700m^3/s. The Vyatka rises in the foothills of the Central Urals and is a principle tributary of the Kama River.

Figure 1. The location of the Vyatka River.

The proposed approach was applied to estimation of annual load of chloride ions at the cross section near the town of Vyatskiye Polyany. This cross section is characterized by annual water discharge of about 22.6 km^3. Two years (1949 and 1954) were selected for calculation: with unimodal type of hydrograph (only spring-summer high flow events) in 1949 and bimodal type of hydrograph (high flow events took place in spring-summer and late fall) in 1954 (see Fig. 2). Both hydrographs exhibit distinct hydrological seasons with sharp rising and falling limbs for spring-summer high flow events. Peak discharge during late fall high flow event in 1954 is relatively moderate, but definitely affects physicochemical processes at this section of the river.

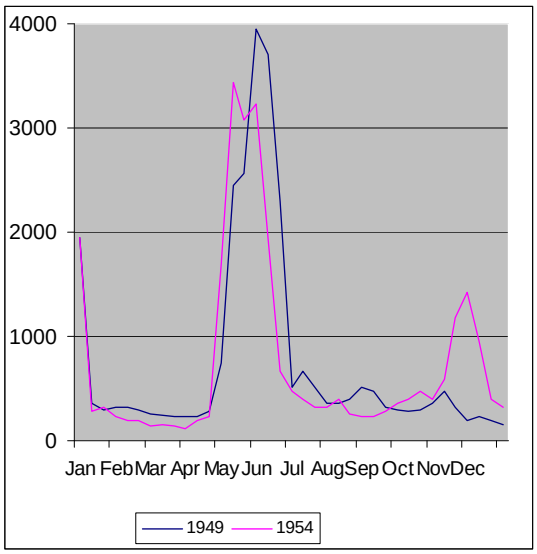

Figure 2. Water discharge near Vyatskiye Polyany, m^3/sec.

There were relatively detailed series of measurements of chloride ions implemented during all hydrological seasons at this location. Thus, the set of data can be considered as representative. Observed concentrations varied in different samples from 0.5 to almost 8.0 mg/l. In order to obtain the accurate value of annual load, concentrations measured at least once a day are required. Missing values of concentrations were surrogated by monotonic interpolation between two consequent observations (see Fig 3).

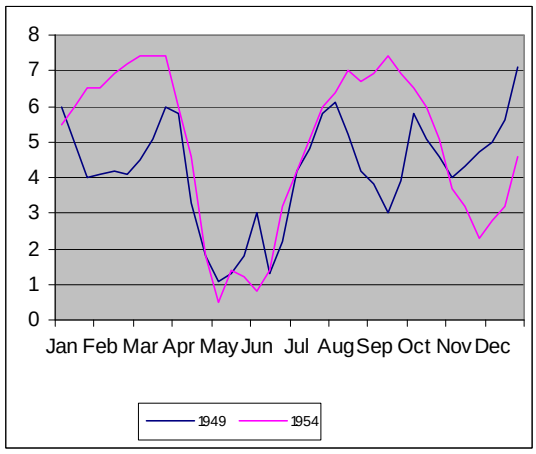

Figure 3. Concentrations of chloride ions near Vyatskiye Polyany, mg/l

Simulated values of concentrations were used to calculate instantaneous and annual loads of chloride ions. The latter was chosen as the most accurate estimation of the load for the algorithm validation. The accurate annual load of chloride ions at the investigated cross section was determined as 63,907 t in 1949 and 60,946 t in 1954. Calculation of basic statistical characteristics showed that inequality (2) holds for investigated set of instantaneous load values and observed water discharge which allowed for application of ratio estimate (3). Formula (3) on a data set with the structure suggested by the algorithm and selected from the complete series of concentrations gave the values of 61,500 t in 1949 and 60,570 t in 1954 which deviated from so-called accurate ones by less than 5%.

It is worth to point out the advantage of application of formula (3). In order to obtain the load value with the guaranteed 5% uncertainty using average values of concentrations and water discharge, more than 1000 samples must be collected over the year. In most of the cases such a detailed series is unrealistic. Application of the proposed algorithm (5) – (6) for required number of observations suggests from 86 to 177 observations depending on the year and the stratification of the investigated period (see Table 1).

Table 1. Distribution of required observations for estimation of the annual load of chloride ions with 5 % uncertainty.

Scheme	Year	Number of required observations per stratum				
		I	II	III	IV	V
A	1949	10	57	28	12	-
	1954	8	40	9	4	24
B	1949	16	134	25	2	-

The stratification was implemented based on available hydrographs. Scheme A reflected hydrograph features for each year in study. Stratum I represented winter low water events. Stratum II combined low parts of rising and falling limbs of the hydrograph. Stratum III included pick discharges and upper parts of rising and falling limbs of the hydrograph. Stratum IV corresponded to summer-fall low water events, and stratum V represented fall high water events. In 1949, fall high water events were not registered and, hence, the corresponding set of data was split into four strata only. Such stratification gives lower values of the water discharge variance, but dates for each stratum vary from year to year. Scheme B took into account average water discharges typical for each calendar month. It comprised of four strata. Stratum I corresponded to the lower water events in December-March, July and August. Stratum II included April and June when usually rising and falling limbs are observed. May (stratum III) presented pick discharges. The period from September to November (stratum IV) covered fall high water events. Fixed dates for stratification are preferable for practical applications, but result in high variance values within each stratum.

6. DISCUSSION

The investigated case study showed that in order to evaluate annual load of an ingredient when a detailed series of concentrations is available, selection of a formula for load calculations is not important. Having limited data of observations, it is important to select a formula which reduces resulting uncertainty. That is why ratio estimate along with stratification were employed. The selected formula can be efficient only if it is supported by an appropriate monitoring design providing sufficient data.

Stratification is important to reduce the uncertainty in the results. It is aimed at lessening the variance of instantaneous load values within each stratum without significant growth of the number of strata. Although formal approach such as clustering can be used for this purpose, stratification was implemented manually according to understanding of main hydrological events in streams. To simplify the application of the proposed approach to practical tasks, strata were defined based on the observed hydrograph instead of the set of instantaneous load values. Different strata required different frequency of observations for concentrations of an ingredient. Usually it is much easier to determine the beginning or the end of a hydrological event and switch the frequency of observations accordingly, because hydrological indicators are monitored with the frequency higher than frequencies for water ingredients.

The investigation of two years with different types of hydrograph showed the necessity to introduce different numbers of strata in each year. While it is relatively easy to implement a posteriori stratification, strata can be unknown for the year when measurements are being taken. Duration of each stratum varied from year to year, so did numbers of suggested observations. The comparison of the frequencies of required observations per stratum for strata with similar hydrological characteristics, however, demonstrated some similarities, which allow for a

generic recommendation. In order to provide reliable estimations of chloride ions loads, observations of this ingredient are to be conducted every other week during low flow events and every day during high flow events. The increase in variance and, therefore, the increase in numbers of suggested observations within a stratum with fixed calendar dates was expected. The required number of concentrations grew up on 60%. Stratification with fixed boundaries between the strata can still be preferable, because it simplifies significantly recommendations for monitoring design.

The proposed approach takes into account statistical characteristics of a set of observation data rather than physicochemical characteristics of a particular water ingredient. Understanding of the nature of variability of various chemical compounds in streams allows for the following remarks. Wide range of factors affects concentrations of chemical compounds in water column. Water ingredients exhibit dissimilarity in response to these factors resulting in significant variability of concentrations. Differences in the variances of concentrations of various water ingredients unavoidably call for different monitoring designs with significant fluctuations of frequencies of observations in order to provide reliable estimations of ingredient loads.

7. CONCLUSION

For an ingredient which is not measured automatically, the reduction of uncertainty in the estimation of chemical load requires a monitoring design corresponding to the formula selected for calculation. Under considered stratification schemes, the suggested numbers of observations for estimation of chloride ions load with 5% uncertainty remains high. These numbers significantly exceed the numbers of observations implemented at many sites of water quality monitoring systems.

A monitoring design which is common for all observed water ingredients at a given site and supporting evaluation of their annual load with a desired level of uncertainty is hardly attainable due to financial and technical constraints. For tiered monitoring systems different allowable levels of uncertainty can be recommended for water quality parameters according to their importance for a particular site.

8. ACKNOWLEDGMENTS

The authors are grateful to anonymous reviewers for their helpful suggestions and comments on the manuscript. The authors would like to express gratitude to Dr. V. Tsirkunov for his insightful suggestions on this topic. The work was implemented based on the data sets prepared in the Hydrochemical Institute, the Russian Federation.

9. REFERENCES

Aulenbach, B.T., and R.P. Hooper (2006), The composite method: An improved method for stream-water solute load estimation, *Hydrological Processes*, 20, 3029-3047.

Bodo B., and T.E. Unny (1983), Sampling strategies for mass-discharge estimation, *Journal of Environmental Engineering,* 109 (4), 812 – 829.

Campolongo, F., Saltelli, A., Sørensen, T., and S., Tarantola (2000), Hitchhiker's guide to sensitivity analysis, *Sensitivity Analysis*, ed. A. Saltelli, K. Chan and E.M. Scott., John Wiley, 15-47, Chichester.

Cochran, W.G. (1963), *Sampling Techniques*, 2nd ed., John Wiley, 413 pp., New York.

Erechtchoukova, M.G., and V.V. Tsirkunov (1989), "Determination of sampling strategy for calculation of mass-discharge in streams with a given accuracy". Proc. of the Conf. "Problems of Surface Hydrology", February, 1987, Gidrometeoizdat, 171-189, Leningrad.

Hooper, R.P, B.T. Aulenbach, and V.J. Kelly (2001), The National Stream Quality Accounting Network: A flux-based approach to monitoring the water quality of large rivers, *Hydrological Processes,* 15, 1089-1106.

Khaiter, P.A., A.M. Nikanorov, M.G. Yereschukova, K. Prach, A. Vadineanu, J. Oldfield, and G.E Petts (2000), River conservation in central and eastern Europe (incorporating the European parts of the Russian Federation), *Global perspectives on river conservation: science, policy and practice*, eds. P.J. Boon, B.R. Davies and G.E. Petts, John Wiley, 105-126, Chichester.

Preston, S.D., V.J. Bierman, Jr., and S.E. Silliman (1989), An Evaluation of Methods for the Estimation of Tributary Mass Loads., *Water Resources Research*, 25(6), 1379-1389.

Robertson, D.M., and E.D. Roerish (1999), Influence of Various Water Quality Sampling Strategies on Loads for Small Streams, *Water Resources Research,* 35(12), 3747-3759.