Dissolved Loads of Rivers and Surface Water Quantity/Quality Relationships
(Proceedings of the Hamburg Symposium, August 1983). IAHS Publ. no. 141.

Multivariate analysis of quality parameters to determine the chemical transport in rivers

JÁNOS FEHÉR
Research Centre for Water Resources Development, (VITUKI), H-1453 Budapest, PO Box 27, Hungary

ABSTRACT In Hungary at present, water quality analyses are performed on samples obtained at nearly 300 surface water quality sampling locations. The length and sampling frequency of the data series vary according to the importance of the sampling locations. The weekly or bi-weekly sampling frequency makes the determination of chemical massflux possible only with limited accuracy. Multi-regressive correlation functions were determined in order to calculate the chemical massflux more accurately. Multiple regression analyses were performed using a 10-year series of sampled and continuously monitored chemical and hydrological data at four selected sampling points. Characteristic statistical parameters are presented for a number of relationships which potentially provide a good fit to the observed data. The reliability of the calculated data obtained by these relations are examined in an example.

Fonctions à variables multiples des composantes de qualité de l'eau pour une évaluation du transport des matériaux chimique des rivières
RESUME Actuellement en Hongrie il y a un réseau de base pour prélévements d'échantillons d'eau de surface en vue d'étudier les composantes de la qualité des eaux. En effet, l'ancienneté des séries de mesures et la fréquence de prélévement de l'échantillon sont variées en fonction de l'importance des points de prélévement du réseau. Dans les cas où la fréquence de prélévement de l'échantillon de l'eau est de un par semaine, l'évaluation du transport continu des matériaux chimique n'est qu'une approximation. Dans notre étude nous avons établi les fonctions à variables multiples pour une évaluation plus précise du transport des matériaux chimiques. Nous avons choisi quatre points représentatifs du réseau où nous avons des séries de mesure dont l'ancienneté est de 10 années. A partir de ces séries nour avons réalisé des études concernant les composantes hydrochimiques et hydrologiques qui sont suivies et mesurés de façon continue. Nous avons représenté les types des fonctions dont les tests d'ajustements sont les plus favorables, ainsi que leurs paramètres statistiques. L'étude présente un exemple de calcul à titre d'exemple de la fiabilité de la série calculée a partir de nos fonctions à variables multiples.

NOTATION

$F = [R^2(N - v - 1)]/[v(1 - R^2)]$ = value of the computed F-test
N = number of elements in the data series
Q = flow rate ($m^3 s^{-1}$)
$\bar{Q}$ = mean flow rate ($m^3 s^{-1}$)
ΔQ = change in flow rate ($m^3 s^{-1}$)
R = multiple correlation coefficient associated with the multiple regression equation
$R_{S\%} = S_h/\bar{Y}_m$ = relative error statistic for the comparison of computed and measured data at calibration
S_Y = standard deviation of dependent variable
$S_h = S_y[N - 1)(N - v - 1)^{-1}(1 - R^2)]^{\frac{1}{2}}$ = estimation error statistic for the comparison of computed and measured data at calibration
t = water temperature (C°)
Y = dependent variable of regression equation
Y_{mi} = i-th element of the measured dependent variable
Y_{coi} = i-th element of the dependent variable computed by the regression equation
$\bar{Y}_m$ = mean value of measured dependent variable
v = number of independent variables
cond. = conductivity ($\mu S\ cm^{-1}$)
$\Delta\% = \Sigma(Y_{mi} - Y_{coi})/(N\ \bar{Y}_m)$ = relative error statistic for the residuals at calibration
τ = time (days)

INTRODUCTION

Determination of the quality conditions in surface waters of Hungary is carried out at the monitoring stations of the regular monitoring network. The sampling system, developed 10 years ago, is being updated and in the course of this exercise the following five basic fields of study are being investigated.

(a) number and location of water quality sampling points,
(b) frequency of sampling,
(c) quality components involved,
(d) storage of quality data,
(e) processing of quality data.

The present study is related to the processing of quality data which gives attention to three main topics:

(a) whether changes in water quality at a particular cross section can or cannot be represented, and what methods of computation are available for their description,

(b) whether monitoring of as many as 30 chemical parameters at each section in the regular network, as is practicised now, is necessary or whether some parameters can be neglected or decreased in frequency (the main groups of components are the parameters of oxygen demand and inorganic substances and special indices),

(c) whether or not certain chemical parameters can be computed instead of being measured in the field and what are the pre-conditions for such estimation.

The present paper is specifically concerned with the third of these

main topics. Processing of water quality data collected on a regular basis in Hungary is undertaken for water management purposes and for calculating the mass transport of each chemical component at particular cross sections. However, sampling on a weekly basis at the regular monitoring stations does not follow the water regime of the rivers so that the data collected are only valid for calculating weekly average values (Kovács, 1979). The present paper is based on the principle that a better estimate of the transport of any chemical component may be derived by using daily data for those chemical substances which can be measured continuously in conjunction with hydrological parameters determined on a weekly basis, providing a function can be found to relate the hydrological information with chemical data collected at a *weekly* frequency from surface waters. The aim of this study is therefore to determine the range of chemical components for which such correlation functions can be developed from the data available.

From some 300 sampling points, four have been chosen for investigation. The main consideration governing this selection was to produce sampling points with different flow rates (Fig.1) and

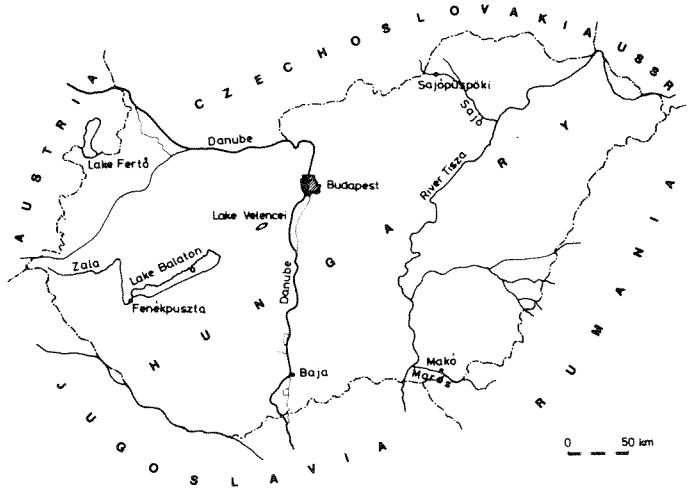

FIG.1 Location map.

quality conditions, and also with data series of weekly frequency over 10 years and for at least 30 quality components. The sampling stations with these requirements include the Danube at Baja, the Sajó at Sajópüspöki, the Maros at Makó and the Zala at Fenékpuszta.

DETERMINATION OF THE MULTIVARIATE FUNCTIONS

Both in Hungary and elsewhere the introduction of the regular water quality sampling has been quickly followed by efforts to establish correlations between water quality and hydrological data, and

especially to find a direct connection between quality components and flow rate. A comprehensive study of the relevant literature has been undertaken by Hock (1981) and from this work it can be concluded that the majority of authors assume a linear connection and consider a few parameters only in the quality/flow relationship. The first comprehensive analysis of quality data from the regular monitoring network in Hungary was carried out by Hock (1980, 1983) with special emphasis on determining the trends in time series of the components investigated. In the course of this analysis the interconnections between a small number of water quality parameters, the flow rate, time and water temperature was investigated employing data from several sites in the regular monitoring network, and in every case a multi-variable linear function was found to be appropriate. On the basis of these investigations, multi-variable linear functions have been adopted for the computations in the present study which are discussed below.

The parameters of linear regression functions investigated in the present study were determined in a preliminary calibration phase using nine years of the weekly data series, and then the functions were verified with data from the tenth year. The parameters which can be included as independent variables in the analysis are determined by the requirement for frequent or "continuous" observation rather than weekly sampling frequency. The components which are continuously measured include flow rate (Q), daily changes in flow rate (ΔQ), water temperature (t), conductivity and time (τ). Although it is possible to continuously monitor pH values, this parameter was disregarded as an independent variable because it emerges that pH does not significantly account for the variability in other components. Similarly, dissolved oxygen concentration may also be measured continuously, but has not been included as an independent variable since it does not provide any additional information to water temperature for the parameters involved as dependent variables. Previous studies (Manczak & Florczyk, 1971; Davis & Zobrist, 1978) have shown that changes in concentration induced by wastewater discharges are proportional to $1/Q$, while those originating from surface runoff are proportional to Q^2, and therefore both these parameters have been employed as independent variables in the present investigation.

Computations have been accomplished employing the method of least-squres regression for a linear combination of the independent variables and using the time series of 27 components at the four sampling stations as dependent variables. Of the possible combinations of the seven independent variables, only the eight most probable cases with five variables have been investigated. The larger the number of variables involved the greater will be the accuracy of the correlation, but there is a limit increasing the number of independent variables because the applicability of the functions in practice will decrease and at the same time the minimal number of data required will increase rapidly (Chatterjee & Price, 1977; Meyer, 1975). The eight types of functions used are:

$$c = f(Q, \Delta Q, Q^2, t, \text{conductivity}) \quad (1)$$

$$c = f(Q, 1/Q, Q^2, t, \text{conductivity}) \quad (2)$$

$$c = f(Q, \Delta Q, 1/Q, t, \text{conductivity}) \quad (3)$$

$$c = f(Q, \Delta Q, 1/Q, t, \tau) \quad (4)$$

$$c = f(\Delta Q, 1/Q, Q^2, t, \text{conductivity}) \quad (5)$$

$$c = f(\Delta Q, 1/Q, Q^2, t, \tau) \quad (6)$$

$$c = f(Q, \Delta Q, Q^2, t, \tau) \quad (7)$$

$$c = f(Q, 1/Q, Q^2, t, \tau) \quad (8)$$

Following the determination of function types the analyses listed below have been carried out for the data from each cross section:

(a) For each component involved, the regression function and its statistical parameters, i.e. multiple correlation coefficient (R), the error statistic from the residuals ($\Delta\%$), the value of F-test and the estimation error (S_h) were determined in the calibration phase.

(b) In order to select the function which provides the best approach for the majority of components at a particular station, significance testing was employed in conjunction with the regression analysis. Comparing the computed values of the test with the critical values belonging to the 5% probability level of the Student's t distribution, it can be decided if a particular independent factor has a significant contribution to the explanation of the dependent variable (Sváb, 1981).

(c) Taking into consideration the results gained from the first two steps, the type of function was selected for each of the four cross sections, and was expressed in a generalized form for the parameters which are to be computed. (Two main criteria have been considered when selecting the type of function: firstly the sum of squares of the residuals should be minimized, and secondly the independent variables of the function in the majority of cases should make a significant contribution to the explanation of the variation in the dependent variable).

(d) Stepwise methods (Mundroczó, 1981) were employed to determine which independent variables (not only those measured continuously) should be included in the generalized functions at each station, and the multiple correlation coefficient and the sum of the squares of residuals were employed as parameters of comparison.

(e) The predictive ability of the regression function for each component, based on the generalized function at each site, has been checked in the verification phase. (The ratio of measured and computed values is 1 if the estimation is correct). Introducing the variable $\xi = c_m/c_{co}$, and calculating the errors $H = 1 - \xi$, the standard deviation value can be calculated as: $\sigma_\Delta = [\Sigma H^2/(N - 1)]^{\frac{1}{2}}$, and $\sigma_{\Delta\%} = \sigma_\Delta \times 100$. The relative standard deviation of errors ($\sigma_{\Delta\%}$) has been introduced for characterizing the control computation, thus discriminating it from the relative standard deviation of errors which characterize the residuals of the regression function ($R_{S\%}$). In the case of correct fitting $\sigma_{\Delta\%} = 0$. No upper limit exists for this index, since theoretically the value of H may exceed any limit.

TABLE 1

1979	Total dissolved materials	Ca^{2+}	Na^{+}	Cl^{-}	K^{+}
RIVER MAROS AT MAKÓ					
$\sigma_{\Delta\%}$	8.3	15.3	16.9	18.9	42.6
R	0.986	0.930	0.890	0.909	0.728
$R_{S\%}$	5.6	12.6	20.2	22.1	22.7

1979	Alkalinity	Ca^{2+}	Total dissolved materials	Cl^{-}
RIVER SAJÓ AT SAJÓPÜSPÖKI				
$\sigma_{\Delta\%}$	11.6	15.9	24.1	25.1
R	0.809	0.794	0.724	0.710
$R_{S\%}$	10.4	13.1	18.1	22.5

1980	HCO_3^{-}	Ca^{2+}	Total dissolved materials	NO_3^{-}	NH_4^{+}
RIVER DANUBE AT BAJA					
$\sigma_{\Delta\%}$	4.6	5.4	11.1	24.7	75.4
R	0.760	0.702	0.757	0.700	0.746
$R_{S\%}$	8.8	10.7	10.3	26.8	47.3

1980	HCO_3^{-}	Total hardness	Alkalinity	Na^{+}
RIVER ZALA AT FENÉKPUSZTA				
$\sigma_{\Delta\%}$	7.9	8.5	10.4	25.2
R	0.783	0.722	0.773	0.748
$R_{S\%}$	9.0	10.1	9.2	19.8

RESULTS AND CONCLUSIONS

(a) It was concluded that good correlations exist between anions and cations of high mean concentration and natural origin and those parameters which can be measured continuously. For the sections under study, regression functions were obtained for Ca^{2+}, Na^{+}, K^{+}, Cl^{-} and total dissolved materials (TDM) in the case of the River Maros at Makó; for Ca, Cl^{-}, alkalinity and TDM in the case of the River Sajó at Sajópüspöki; for Ca^{2+}, HCO_3^{-}, TDM, NO_3^{-} and NH_4^{+} in the case of the River Danube at Baja; and for HCO_3^{-}, total hardness, alkalinity and Na^{+} in the case of the River Zala at Fenékpuszta. Of the 30 components investigated, about 15% are suitable for the development of regression functions.

(b) The generalized forms of the regression functions are:

$$c = f(Q, 1/Q, Q^2, t, \text{conductivity})$$

for the Maros at Makó, Sajó at Sajópüspöki and Zala at Fenékpuszta,

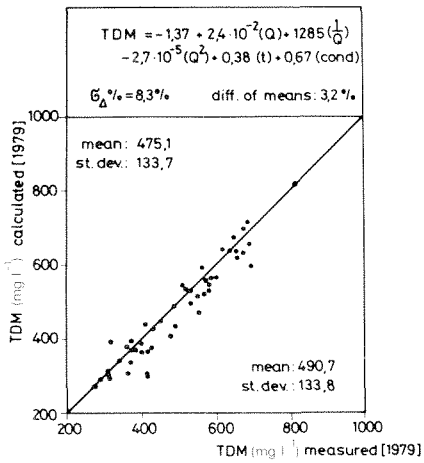

FIG.2 Comparison of measured and calculated TDM values for the River Maros at Makó (1979) employed in the validation procedure.

and

c = f(Q, ΔQ, 1/Q, t, conductivity)

for the Danube at Baja.

(c) The regression correlations for the components listed in (a), which have been derived from the function types listed in (b), when compared to the functions derived from five independent variables using stepwise methods, showed that no significant difference existed between the relative standard deviation of residuals and that of the square sum of deviations.

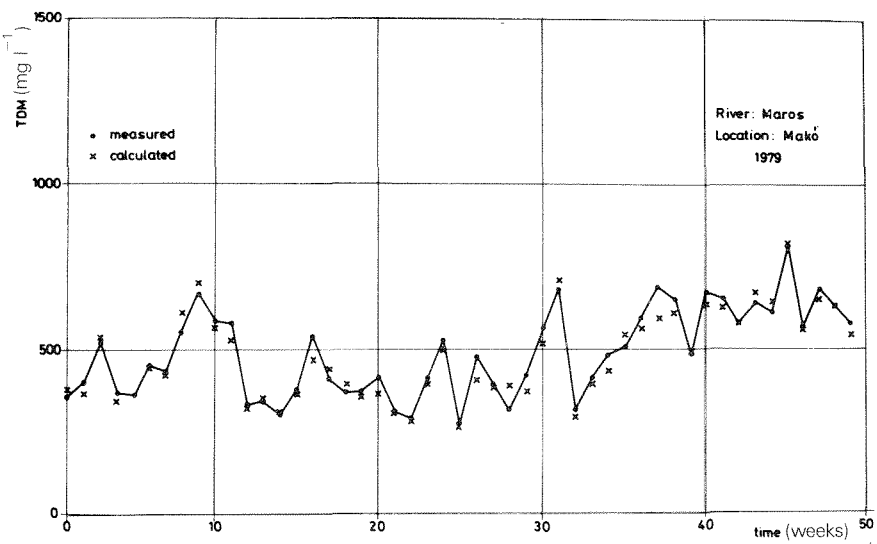

FIG.3 Measured and calculated TDM vs. time.

(d) The relative standard deviation of error ($\sigma_{\Delta\%}$) derived from the verification procedure, the multiple correlation coefficient of the regression function as well as the relative standard deviation of residuals are given in Table 1 for each parameter at each cross section.

The measured and calculated values have been plotted against each other for every component, and Fig.2 shows the distribution of points representing the total dissolved materials (TDM) values in the river Maros at Makó. Figure 3 presents the time trend of measured and computed values over a common period.

In summary, it can be concluded that a particular water quality component can be substituted by values computed from regression functions only if $R_{S\%} \leq 10\%$ for the function and $\sigma_{\Delta\%} \leq 10\%$ for the control.

REFERENCES

Chatterjee, S. & Price, B. (1977) *Regression Analysis by Example*. Wiley Series, New York.

Davis, J.A. & Zobrist, J. (1978) The interrelationships among chemical parameters in rivers analysing the effect of natural and anthropogenic sources. *Prog. Wat. Tech.* 10 (5/6).

Hock, B. (1980) Analysis of tendencies in water quality changes of rivers (in Hungarian). *VITUKI Report*.

Hock, B. (1981) Changes in water quality of free-flow rivers, taking into account the flow rate and water temperature (manuscript in Hungarian). VITUKI, Scientific dissertation.

Hock, B. (1983) Analysis of tendencies in water quality changes (in Hungarian). *Viz. Közl.* 1.

Kovács, G. (1979) Principles of modern hydrological activities, I-II (in Hungarian). *Viz. Közl.* 3-4.

Manczak, H. & Florczyk, H. (1971) Interpretation of results from the studies of pollution of surface flowing waters. *Wat. Res.* 5, 575-584.

Meyer, S.L. (1975) *Data Analysis for Scientists and Engineers*. Wiley Series, New York.

Mundroczó, G. (1981) Applied regression analysis (in Hungarian). *Akadémiai Kiadó, Budapest*.

Sváb, J. (1981) Biometrical methods in research (in Hungarian). *Mezőgazdasági Kiadó, Budapest*.