MATHEMATICAL METHODS IN WATER QUALITY CONTROL

B. Barabás, J. Császár and J. Reimann

Department of Civil Engineering Mathematics,
Technical University, H-1521 Budapest

Received June 20, 1987

Abstract

In this paper we intend to summarize the most important mathematical statistical methods which play a basic role in water quality control. In the first section we will use a random variable model and it will be shown how to observe the changing of water quality by testing homogeneity.

In the second section we will compare the means of the random variables. A few new idea will be introduced in the third and fourth sections based on stochastic process models.

In water quality control a fundamental problem is to observe changes in the concentration of characteristic components. Suppose we have data from observations at the same place and at different dates. For example we have the weekly means of ammonium concentrations for some years and we would like to know if the average concentration increased, decreased or remained unchanged in the second year.

To observe the change one can use a common test for homogeneity if the sample elements are independents. This condition does not hold in most cases. We cannot tell the data form a statistical sample, i.e. they are independent, identically distributed random variables (i.i.d.r.v.). They form a series of observations in a stochastic process that we will call time series.

If there is only a short time between two observations we can find a strong connection between the data. Increasing the time the dependence will decrease. To construct an empirical distribution function based on the data would be correct if the data were independent. Inspite of this fact the examination of water quality is frequently based on these empirical distribution functions. However, in a lot of cases we cannot judge that as wrong, because we can get correct conclusions if the dependence is weak. We suggest to check in every case.

Denote by X the concentration observed on a randomly chosen day. Given an x, let$Fn(x)$ be the relative frequency of the event $\{X < x\}$. Then $Fn(x)$ is a good approximation for the probability $P(X < x) = F(x)$ even if the data are just shightly dependent.

If the concentration of some component strongly depends on something (for example on the temperature) this method is not suitable. We are going to suggest different methods which may have practical inportance.

One of these methods is based on the monthly means of concentrations. Dependence between the monthly means is much lower than on the daily or weekly means. For example if we want to know whether the oxigen concentration remained unchanged or not in the last two years, let us compare a the previous three years and we obtain

I $X1, X2, \ldots X36$, data for the first 3 years and
II $Y1, Y2, \ldots Y24$, data for the next 2 years.

We suggest to test homogeneity based on combinatorial considerations which is a generalised version of the Gnedenko—Koroljuk test. The efficiency of this test is almost the same as of the Kolmogorov—Smirnov test (see for example [10]).

As an other method, we suggest the comparison of the monthly or annual means. Due to the central limit theorem, we can suppose a Gaussian distribution for the arithmetic means and the well known two-sample test of Student can be used for homogeneity.

The third suggested method will be very different from the previous two. We will consider the data as a time series. Sometimes these time series are very difficult and then we will associate a certain random variable to them and observe the change of the distributions of these random variables to conclude the change in water quality.

1. Combinatorial method to test homogeneity

Suppose we have data about a certain component of the water quality for the last 5 years. Say we have m data for the first 3 years and n for the last 2 years. Let us denote these data by

I $X1, X2, \ldots Xm$ and II $Y1, Y2, \ldots Yn$

For example let X and Y be the monthly means of the chemical oxygen demand (COD) in the first 3 and the last 2 years, respectively. Furthermore, let I and II be the corresponding samples. Let $F(x)$ be the distribution function of X and $G(y)$ the distribution function of Y. We want to test whether the function $F(x)$ is identical with $G(y)$?

Let the hypothesis Ho be: $F(x) = G(y)$. For the sake of simplicity let us suppose $m > n$ and $m - n$ as not too large, e.g. $m - n = c\sqrt{m + n}$. Denote the ordered sample arises from I and II by

I′ $X1' \; X2' \ldots Xm'$
II′ $Y1 \; Y2' \ldots Yn'$

and the ordered sample of the union of I' and II' be

$$III' \; Z1' \; Z2' \; \ldots \; Zn + m'$$

We also introduce the notation

$$Vi = \begin{cases} +1 \text{ if } Zi' \in I' \\ -1 \text{ if } Zi' \in II' \end{cases}$$

and

$$Si = V1 + V2 + \ldots . + Vi \quad i = 1, 2, \ldots \ldots m + n$$

The partial sum of Vi is illustrated in Fig. 1.

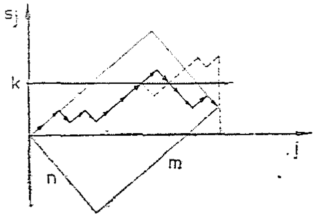

Fig. 1

Every possible graph, we will call them trajectories, runs in the rectangle drawn by dotted line and there are exactly

$$\binom{m+n}{n} = \binom{m+n}{m} \text{ trajectories}$$

Fixing an arbitrary level, K, let us consider the trajectories which reach this level K at a certain point, say t. Reflect that part of the trajectory where $i > t$ to the straight line $y = K$. The endpoint of the trajectory is then at $2K-(m - n)$ after $m + n$ steps. Let us calculate how many trajectories we have which reach level K. For the sake of simplicity we suppose that both m and n are even or odd, say $m + n = 2N$ and $m - n = 2L$. (In the worst case we omit one observation.) We want to find out how many trajectories have the endpoint at $2K - 2L$ after $2n$ steps. Let the number of upward steps be A and the downward ones be B. Then

$$\begin{array}{l} A + B = 2N \\ A - B = 2K - 2L \\ \hline A = N + K - L \\ B = N - K + L \end{array}$$

Thus the number of trajectories which reach the level K is:

$$\binom{2N}{N+K-L}$$

It implies that

$$P(\max_j S_j > K) = \frac{\binom{2N}{N+K-L}}{\binom{2N}{n}} = \frac{\binom{2N}{N+K-L}}{\binom{2N}{N-L}} \tag{1}$$

To calculate the exact value of the probability (1) one can use the table of binomial distribution (if $N < 25$) the following way:

In columne $P = 0.5$ we find

$$P_{N+K-L} = \binom{2N}{N+K-L}\left(\frac{1}{2}\right)^{N+K-L}\left(\frac{1}{2}\right)^{N-K+L} = \binom{2N}{N+K-L}\left(\frac{1}{2}\right)^{2N} \tag{2}$$

$$P_{N-L} = \binom{2N}{N-L}\left(\frac{1}{2}\right)^{N-L}\left(\frac{1}{2}\right)^{N+L} = \binom{2N}{N-L}\left(\frac{1}{2}\right)^{2N} \tag{3}$$

Hence

$$\frac{P_{N+K-L}}{P_{N-L}} = \frac{\binom{2N}{N+K-L}}{\binom{2N}{N-L}} \tag{4}$$

The ratio (4) can be calculated for $K = 1, 2, \ldots N$ and that value of K can be found, when this ratio becomes less than a properly chosen $\varepsilon > 0$, if we test the hypothesis Ho at level $1 - \varepsilon$. When N is large enough we can estimate K as follows. Using the well-know approximation:

$$\lim_{N\to\infty} \frac{\binom{2N}{N+L}}{\binom{2N}{N}} \approx e^{-\frac{L^2}{N}} \tag{5}$$

and rewriting the expression (4) to the form:

$$\frac{\binom{2N}{N+K-L}}{\binom{2N}{N-L}} = \frac{\binom{2N}{N+K-L}}{\binom{2N}{N}} \cdot \frac{\binom{2N}{N}}{\binom{2N}{N-L}} \tag{6}$$

Using the limit (5) we have:

$$\lim_{N\to\infty}\frac{\binom{2\,N}{N+K-L}}{\binom{2\,N}{N-L}}=\lim_{N\to\infty}\frac{\binom{2\,N}{N+K-L}}{\binom{2\,N}{N}}\cdot\lim_{N\to\infty}\frac{\binom{2\,N}{N}}{\binom{2\,N}{N-L}}=$$
$$=\exp\left[-\frac{(K-L)^2}{N}\right]\cdot\exp\left[\frac{L^2}{N}\right]=\exp\left[-\frac{K^2-2\,KL}{N}\right] \tag{7}$$

Hence

$$P(\max_i S_i\geq K)\approx\exp\left[-\frac{K^2-2\,KL}{N}\right] \tag{8}$$

From this expression we obtain K by the equation

$$\exp\left[-\frac{K^2-2\,KL}{N}\right)=\varepsilon \tag{9}$$

We get by simple calculation: $K^2-2KL+N\ln\varepsilon=0$ and

$$K=L+\sqrt{L^2-N\ln\varepsilon} \tag{10}$$

In practice $\varepsilon=0.05$ is frequently used. Then

$$K=L+\sqrt{L^2+3N} \tag{11}$$

(because $\ln 0.05=-2.99\approx-3$)

As an illustration let us consider the case $m=17$, $n=13$. Then $2N=30$ and $L=2$. From the formula (11) we obtain:

$$K=2+\sqrt{49}=9 \tag{12}$$

That is

$$\frac{\binom{2\,N}{N+K-L}}{\binom{2\,N}{N-L}}=\frac{\binom{30}{22}}{\binom{30}{13}}\leq 0.05 \tag{13}$$

On the other hand from the table of binomial distribution we have

$$\frac{P_{22}}{P_{13}}=\frac{\binom{30}{22}}{\binom{30}{13}}=\frac{0.00545}{0.11152}=0.048\approx 0.05 \tag{14}$$

It shows the estimation (10) to be fairly good even if N is not too large.

Notice, that especially if $m = n$ (i.e. $L = 0$) we get the result of Gnedenko and Koroljuk from Equ. (1) which is

$$P(\sup n\,[F_n(x) - G_n(x)] > K) = \frac{\binom{2\,N}{N+K}}{\binom{2\,N}{N}} \approx e^{-\frac{K^2}{N}} \qquad (15)$$

In this case the table of binomial distribution is unnecessary, because the critical value of K is

$$K = \sqrt{-N \ln \varepsilon} \qquad (16)$$

For example if $m = n$ and the $\varepsilon = 0.05$ we obtain the critical value of K:

$$K = \sqrt{3N} \qquad (17)$$

From expression (8) we can conclude an interesting relation.

Introducing notation $Bm,\, n = \max\limits_{x} [mF_m(x) - nG_n(x)]$ we have

$$P(\max_{x} S_x < K) = P(B^{+}_{m,n} < K) = 1 - \exp\left[-\frac{K^2 - 2\,KL}{N}\right] \qquad (18)$$

If $K = z\sqrt{2N}$ and $L = m - n = c\sqrt{2N}$ then from expression (18) we get

$$P\left(\frac{B^{+}_{m,n}}{\sqrt{m+n}} < Z\right) = 1 - e^{-2Z^2-4cZ} \qquad (19)$$

This is a Kolmogorov—Smirnov type distribution.

It can be proved to be an assymptotically consistent test to the counter-hypothesis H1: $F(x) > G(y)$. The Gnedenko—Koroljuk test can be generalized in a very similar way for statistics

$$Bm,\, n = \max_{x}\left[nF_n(x) - nG_n(x) + \frac{m-n}{2}\right] - \frac{m-n}{2} \qquad (20)$$

As an example let us consider the monthly means of COD for the years 1976, 1977, 1978 relative to the monthly means of COD for the years 1979, 1980 for the Danube river at Baja. The data for the first 3 years are:

C(1) = 19.87
C(2) = 19.87
C(3) = 24.77
C(4) = 23.55
C(5) = 22.88
C(6) = 15
C(7) = 18.33

C(8) = 19.66
C(9) = 16.88
C(10) = 16.25
C(11) = 20.55
C(12) = 23
C(13) = 29.22
C(14) = 24
C(15) = 17.55
C(16) = 23.87
C(17) = 20.66
C(18) = 21.66
C(19) = 22
C(20) = 21.33
C(21) = 22.66
C(22) = 29.5
C(23) = 25.66
C(24) = 22.66
C(25) = 24.77
C(26) = 29.25
C(27) = 23
C(28) = 20.87
C(29) = 19.22
C(30) = 16.55
C(31) = 17.33
C(32) = 21.88
C(33) = 23.25
C(34) = 17.66
C(35) = 19.88
C(36) = 27
average: $m = 21.72$
$\sigma = 3.61$

autocorrelation
$r(1) = 0.3801$
$r(2) = 0.0109$

The data for the second 2 years are:

C(1) = 26.33
C(2) = 28.5
C(3) = 24.33
C(4) = 22.62
C(5) = 20.2
C(6) = 19.55
C(7) = 13.66

$C(8) = 22.44$
$C(9) = 26$
$C(10) = 19.88$
$C(11) = 21.66$
$C(12) = 18.14$
$C(13) = 20.11$
$C(14) = 19.66$
$C(15) = 22.44$
$C(16) = 20.87$
$C(17) = 19.62$
$C(18) = 15.55$
$C(19) = 14.33$
$C(20) = 13.37$
$C(21) = 20.88$
$C(22) = 17.33$
$C(23) = 21.12$
$C(24) = 21.66$
$m = 20.42$
$\sigma = 3.74$
$r(1) = 0.4604$
$r(2) = 0.1737$

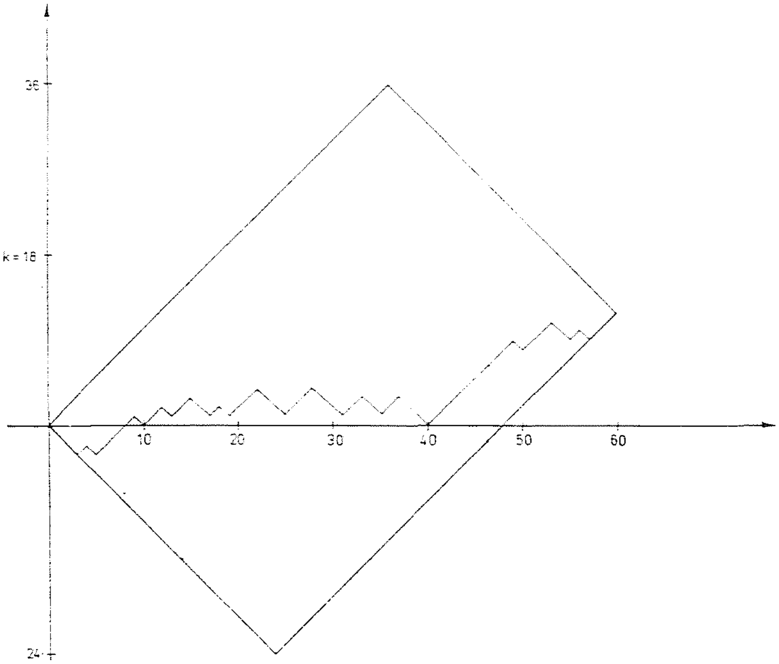

Fig. 2

2. Comparison of two arithmetic means

As we mentioned the change in the concentration of some components of water quality can be observed by the comparison of the monthly or annual means. Denote by X and Y the random variables which describe the concentration in the first and the second period, respectively (for example the first 3 years and the following 2 years of the last 5 years).

The statistical sample (say monthly means) will be

I $X1, X2, \ldots Xn$ and
II $Y1, Y2, \ldots Ym$

We suppose X and Y to have Gaussian distribution or at least near to that. Then, to test the hypothesis Ho: $X = Y$ we can use the two-sample test to Student if the variances are equal. If this is not the case then we have of use the Welch-test. Now, we have the test:

$$t_{n+m-2} = \frac{\overline{X} - \overline{Y}}{\sqrt{(n-1)\, S_x^{*2} + (m-1)\, S_y^{*2}}} \sqrt{\frac{nm(n+m-2)}{n+m}} . \qquad (21)$$

This statistics has a Student distribution with parameter $n + m - 2$ and the following inequality holds with a probability of 95% if $n + m - 2 \approx \approx 60$.

$$-2 \leq t_{n+m-2} \leq 2 \qquad (22)$$

Now we can answer one of the most important question from a practical point of view: How big a pollution implies a significant increase of the means? The answer is based on the t-test.

Increasing the value of the random variable X by a constant c, the expected value of the random variable $Y = X + c$ will be $E(Y) = E(X) + c$. Estimating the expected values by the arithmetic means $Y = X + c$ we get

$$2 \approx t_{n+m-2} = \frac{c}{\sqrt{(n-1)\, S_x^{*2} + (m-1)\, S_y^{*2}}} \sqrt{\frac{nm(n+m-2)}{n+m}} . \qquad (23)$$

Thus if

$$C \geq \frac{2\sqrt{(n-1)\, S_x^{*2} + (m-1)\, S_y^{*2}}\, \sqrt{n+m}}{\sqrt{nm\,(n+m-2)}} \qquad (24)$$

we get a significant difference in the means by the t-test.

For example:

The annual mean of the ammonium concentration of the Danube at Baja in 1977, was:

$X = 0.50$ mg/l and $S_n^* = 0.4$ $n = 104$ (two data/week).

We have, by formula (24)

$$C = \frac{2\sqrt{2}\sqrt{n-1}\,S_n^*\sqrt{2n}}{n\sqrt{2}\sqrt{n-1}} = \frac{2\sqrt{2}\,S_n^*}{\sqrt{n}} = \frac{2.1{,}41.0{,}4}{104} = 0.112 \text{ mg/l}.$$

At the same place, the average concentration was $Y = 0.64$ mg/l with the same variance in 1979. Thus we can conclude the water quality was worse than in 1977, with probability of 95%.

3. Excess of a certain level

It is of particular interest to examine the excess of a high level of a certain pollution and duration of this excess. Let level c be for example when the water becomes third class quality if the concentration of a certain component exceeds this level.

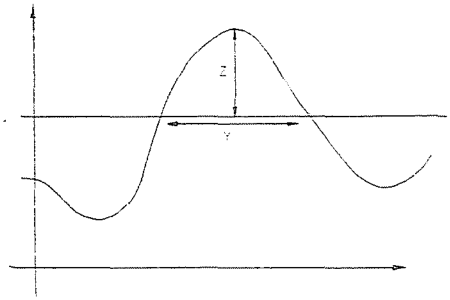

Fig. 3

Let us consider the excess as a random variable denoted by Z in respect to level c, and Y the duration above level c. The distribution of Z and Y can be determined by the sample. If c is high enough both random variables have an exponential distribution, at least theoretically.

In the following example we give the corresponding data of the excess and duration of COD, where level $c = 25$ mg/l is the second class quality level.

The data are for the years 1971—1980 for the Danube at Baja and we have 104 data for every year.

Here we cannot expect the empirical distribution function to fits an exponential function very well, because the data are integers. But a good approximation seems possible for Figs 4 and 5.

In many cases the higher value of Z is the higher one also of Y (i.e. there can be a strong positive correlation between X and Y).

If Z and Y belong to the same family of distributions (for example both are exponential) then the regression is linear. This fact helps to determine the entire quantity of pollution. The joint distribution function $H(X, Y)$ of random

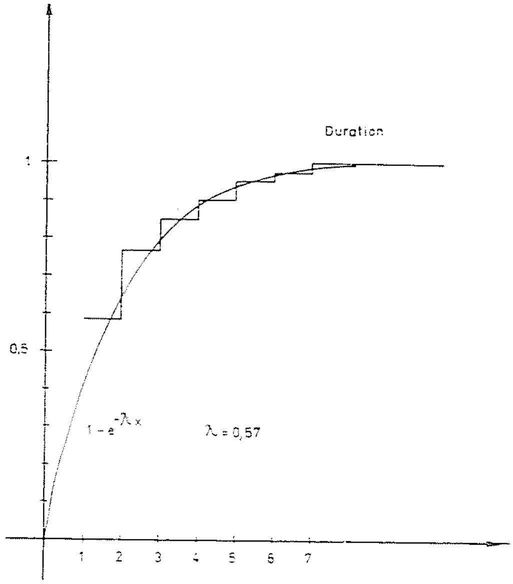

Fig. 4

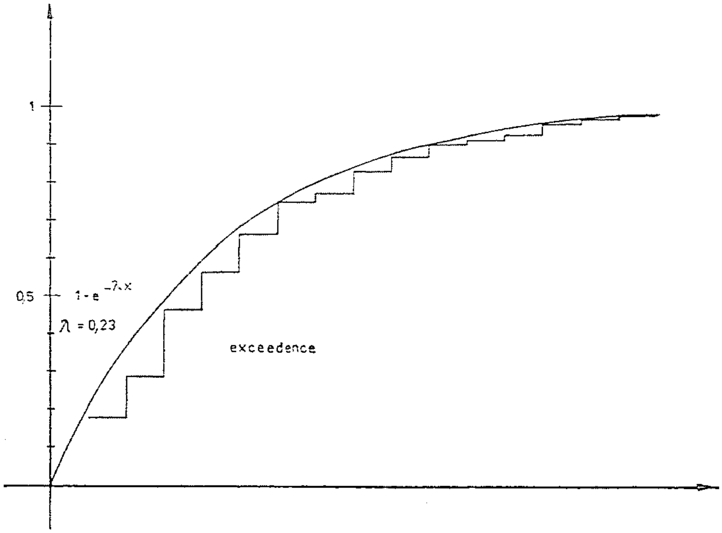

Fig. 5

variables Z and Y also plays an important role in judging water quality. We found the joint distribution function $H(Z, Y)$ to fit the following type of bivariate distribution functions:

$$H(Z, Y) = \min [F(Z), G(Y)] + (1 - \alpha) F(Z) G(Y) \qquad (25)$$

where $0 \leq \alpha \leq 1$.

The parameter α can be estimated using the so-called median correlation:

$$q = 4\,H\,(\tilde{Z}_{1/2}; \tilde{Y}_{1/2}) - 1 \tag{26}$$

For the distribution function $H(Z, Y)$ of (25)

$$q = 4\,H\,(\tilde{Z}_{1/2};\ \tilde{Y}_{1/2}) - 1 = 4\left(\alpha\frac{1}{2} + (1-\alpha)\frac{1}{4}\right) - 1 = 4\frac{1+\alpha}{4} - 1. \tag{27}$$

Hence

$$q = \alpha \tag{28}$$

From the joint distribution function $H(Z, Y)$ we get the probability

$$P(Z \geq z;\ Y \geq y)\ = 1 - F(Z) - G(y) + H(Z, y) \tag{29}$$

To detect the change in water quality it is enough to find the change of this probability.

4. Time series model

In the first two sections we used a random variable model to describe the behaviour of water quality. In the third section we considered a stochastic process model and we made two random variables associate to the process and studied them. In the present section we are going to use a time series model.

Usually we have daily data or two data week for the components of the water quality. These data are not independent. To measure the dependence we always use the autocorrelation function. Here we are going to give a short review of the autoregressive — moving average processes which can be used very well to describe the water quality. We must stress that there is not only a single true model.

Let Xt be a discrete time series. We call it an autoregressive process of order P ($AR(P)$ in short) if

$$X_t = r_1 X_{t-1} + r_2 X_{t-2} + \ldots + r_p X_{t-p} + a_t \tag{30}$$

where $r_1, r_2, \ldots, r_p$ are constants and a_t is an independent random 'shock'. A sequence of random variables $a_t a_{t-1} \ldots$ is called a white noise process. These are random drawings from a fixed distribution, usually assumed to be Gaussian and having a zero mean and variance σ^2.

Another kind of model, of great practical importance when representing observed time series, is the so-called finite moving average process:

$$X_t = a_t - S_1\, a_{t-1} - S_2\, a_{t-2} \ldots S_q\, a_{t-q} \tag{31}$$

It is, more exactly, a moving average process of order q (MA (q)) for short where $S_1 S_2 \ldots S_q$ are constants and a_t is a white noise process.

To achieve greater flexibility in fitting actual time series, it is sometimes advantageous to include both autoregressive and moving average terms in the model. This leads to the mixed autoregressive — moving average model:

$$X_t = r_1 X_{t-1} + \ldots + r_p X_{t-p} + a_t - S_1 a_{t-1} - \ldots S_q a_{t-q}. \tag{32}$$

This is an autoregressive — moving average process or ARMA (p, q) process for short. Due to the high variety of first and second order process realisations we pay special interest to them.

The AR(1) first order autoregressive process is a Gauss—Markov process, too.

$$X_t = r X_{t-1} + a_t \tag{33}$$

If we want to apply this model then the first question is how to determine parameters r and $Var\ (a_t)$. The second question is how to check the fitting of the model, and a third question: What to do with the model?

To answer the first question we determine the theoretical autocorrelation function. The autocorrelation function is a function of the positive integers and has the value the correlation between X, and X_{t+K} separated by K intervals of time, that is autocorrelation at lag K is

$$\varrho_k = \frac{E[(X_t - \mu)(X_{t+K} - \mu)]}{\sqrt{E(X_t - \mu)^2 E(X_{t+K} - \mu)^2}} \tag{34}$$

where

$$\mu = E(X_t) \tag{35}$$

The theoretical autocorrelation function of AR(1) process is

$$\varrho_K = r^K \tag{36}$$

We can obtain a simple estimation for the parameter r by estamating $r = \varrho_1$.

How can we check if our model fits or not? Let us transform the process to the form:

$$X_t - r X_{t-1} = a_t \tag{37}$$

This is a white noise process. We therefore replace r by the estimated value $\tilde{r}$ and make a goodness of fit test concerning normality. It may not be the best solution but we found it satisfactory.

As a matter of fact the third question is the most interesting. Possessing the mathematical model we can forecast, for example by computer simulation. This is important in itself but even more so is that we can determine a band where the realisations of the process must run with a high probability. If we find the real process step from this band we can conclude to a sudden change in water quality. According to our experience this method is effective especially in short term forecasting.

As an example let us consider the data drawn from COD for the Danube at Baja in 1971.

The autocorrelation function has the values:

$$C_1 = 0.5492$$
$$C_2 = 0.4626$$
$$C_3 = 0.3381$$
$$C_4 = 0.2431$$
$$C_5 = 0.1430$$
$$C_6 = 0.0901$$
$$C_7 = 0.0324$$

If someone decided to use ARMA (p, q) model, parameters p and q have to be found first. This procedure is called identification and is usually based on the autocorrelation function. For example we know the autocorrelation function of an AR(1) process to decrease exponentially. In our example the empirical autocorrelation function $c_i (i = 1, 2, \ldots, 7)$ is also decreasing. Hence it does not seem to be wrong to use AR(1) model. Of course it is not the only possible case. We also tried to fit the ARMA(2.0), ARMA(1.1), ARMA(1.2) models.

The next step is the estimation of parameters and the third is diagnostic checking.

By diagnostic checking we mean checking the fitted model as to its relation to the data with the intent to reveal model inadequacies. We found in our example that ARMA(1.1) is inadequate, and one can use the following models:

$$\text{ARMA(1.0)}: X_t = 0.55\ X_{t-1} + 3.21\ a_t$$
$$\text{ARMA(2.0)}: X_t = 0.42\ X_{t-1} + 0.23\ X_{t-2} + 3.12\ a_t$$
$$\text{ARMA(1.2)}: X_t = 0.78\ X_{t-1} + 2.81\ a_t - 0.86\ a_{t-1} + 0.33\ a_{t-2}$$

References

1. BLOMQVIST, N.: On a measure of dependence between two random variables. Am. Math. Statist. *21* (1950).
2. BOX, G. E. P.—JENKINS, G. M.: Time series analysis, forecasting and control. Holden-Day Inc. 1976.
3. CRAMER, H.—LEADBETTER, M. R.: Stationary and related stochastic processes. John Wiley and Sons 1976.
4. FELLER, W.: An introduction to probability theory and its application. John Wiley and Sons 1957.
5. REIMANN, J.—NAGY, V.: Hidrológiai statisztika. Tankönyvkiadó 1984.
6. REIMANN, J.: Dependence analysis. Period. Politechn. Civil Eng. *27*, 1 (1983).
7. TAKÁCS, L.: Tartózkodási idő problémákról. MTA III. oszt. Közl. VII. *3—4*. 1957.
8. TODOROV, I.—Zelenhasic, E.: A stochastic model for flood-analysis. Water Research Publ. 1970.

9. TUSNÁDY G.—ZIERMANN M.: Idősorok analízise. Műszaki Könyvkiadó. Budapest 1986· In Hungarian.
10. VINCZE I.: Matematikai statisztika ipari alkalmazásokkal. Műszaki Könyvkiadó. Budapest 1968. In Hungarian.
11. YEVJEVICH, V.: Stochastic processes in Hydrology. Water Research Publ. Colorado USA 1972.

Dr. Béla BARABÁS
Prof. Dr. József REIMANN } H-1521 Budapest
Dr. József CSÁSZÁR OVH H-1394 Budapest POB 351