OPTIMIZATION OF HYDROLOGICAL MEASUREMENTS

By

I. V. NAGY

Department of Water Management, Institute of Water Management and Hydraulic Engineering, Technical University, Budapest

Received: September 15, 1979

In the recent decade, both in this country and abroad, the approach to interpretation and methods of investigation of hydrological phenomena has significantly changed as a result of the generalization of up-to-date mathematical statistics and computer technique. Actually, the hydrologic time-series are characterized by applying different mathematical models of the stochastic processes for making use of the information involved. The practically available series of measurements form one of the finite realizations (i.e., models) of the theoretically infinite time-series, wherefore the basic question remains: what is the *information* content of the available model for describing (reproducing) the process at an acceptable accuracy. Assuming an appropriately selected mathematical model, a definite answer might only be given by the practice, i.e., by the measurements to be carried out in the future, but still, there are means for drawing preliminary conclusions.

The second question (of theoretical and practical significance) is, *what is the advisable frequency* of measurements, thus what a (time or space) interval is to be chosen between the consecutive measurements. As a rough approximation, theoretically, the interval should be zero, i.e., the most of information is given by *continuous* measurements. But, at a closer look, it appears that:

- for the acceptable reproduction of an actual natural (hydraulic, hydrologic, water quality, etc.) process it is quite enough to measure the actual value of a given process only at predetermined intervals τ;
- in case of reduction of the interval τ (increasing the number of measurements), the expenses of the storage and processing of data will increase;
- upon increasing the interval τ, the information content in the model decreases, and so do accuracy and reliability of the parameters calculated from the model data;
- the optimum value of the interval τ depends on the type of the process, thus, different values are obtained in the case of nearly deterministic, stochastic or random processes;

— the optimum value of τ depends on what kind of statistical parameter (e.g. expected value, standard deviation etc.) has to be calculated from the model;
— finally, the optimum measurement interval depends on the intended application, on accuracy requirements of the actual or assumed future planning problem, and on the total planned costs of the project.

1. Previous publications

Several optimum sampling procedures have been presented in the literature on hydrologic statistics known to the author, the problem seems, however, not at all solved from several, significant aspects. In analysing the problem of measuring the rainfall and runoff, EAGLESON and SHACK [3] drew the conclusion that the optimum rainfall measurement interval should be chosen from the viewpoint of the desired accuracy of the runoff description. The procedure is based upon the method of spectral analysis, where the highest frequency of the rainfall hydrograph is arbitrarily chosen, at a serious detriment, however, to the range of applications of the procedure.

VAN DE NES and HENDRIKS [5] investigated linearly distributed model of surface runoff and adopted essentially the procedure above deriving the sampling interval from the spectrum function of the runoff time series.

QUIMPO and YANG [7] deal with the problems of sampling water discharge and temperature and use the quotient of the variances of the correlated by uncorrelated time series as index of the data set, so that the data set increases with decreasing correlation. This conclusion is not new in itself, also its practical utility is significantly restricted by the omission of the effect of averaging.

A conclusion of fundamental importance by DYHR-NIELSEN [2] is the significant loss of information upon taking the averaged measurement results, at definite intervals, rather than results of discrete measurements at these intervals.

YEVJEVICH [11] tackled the problem of the optimum sampling interval by considering the time series as a MARKOV's chain, and describing the variation of FISCHER's information quantity by the relationship between the variance of the random component and the interval τ.

Investigating the problem of sampling at uniform intervals in discrete points, OGINK [6] found the mean value, the variance and the second spectral moment to differently depend on the interval τ. The optimum value of τ is found, as a rule, near the NYQUIST's interval.

In examining the optimum distance between verticals in surveying the river cross sections, RUPPERT [10] found a relation between the error due to the increase of measurement intervals Δl and the coefficient of variation,

assuming that the consecutive cross-sectional depths are independent of each other.

DIMAKSYAN [1] calculates, just as OGINK, the optimum value on the basis of maximum frequency, thus, the number of measurements in a period is twice the number of harmonics.

From the above short survey it is evident that the significance of measurement optimization has been recognized in many countries, and even the partial results are significant, but no generalizable solutions have been found to now.

Recently, the *Department for Water Management* has been concerned with the optimum sampling intervals of different type processes and calculation of the amount of information needed for statistical conclusion of given accuracy and reliability, with the conclusion that pure random, stochastic, or nearly deterministic processes are advisably studied on different mathematical models.

This is a report on the first part of a research work of three parts, making use of a usual, comparatively simple but practically acceptable way of description of stochastic processes. In two subsequent papers, the problem of deterministic, and of purely random processes will be presented.

2. Information theory of solution to the problem

The optimum sampling interval τ is to be determined for the discrete stochastic process:

$$[\xi(k\tau)]_{k=0}^{\infty}$$

or, more exactly, for one of its realizations, in the considered case a discharge time-series measured in a given cross section of a water course, as statistical sample. For convenience, denote the stochastic process $\xi(k\tau)$ by ξ_k, $(k = 0,1, 2, \ldots, n)$.

The sampling problem will be solved for the special case where the discharge time-series is described by the autoregressive model often used in hydrology; in our case:

$$\xi_{k+n} = \mu + p(\xi_k - \mu) = z_{k+1}\sigma\sqrt{1 - p^2} \tag{1}$$

where

ξ_{k+n}, ξ_k — values of discharge at times $(k + n)$, (k);
μ — expected value;
p — process parameter, in our case the first coefficient of autocorrelation ($|p| \leq 1$);
z_{k+1} — independent, standard variate of normal distribution;
$\sigma > 0$ — standard deviation of time series.

In our case, if
$\xi_0 \in N(\mu, \sigma)$ then, according to A. RÉNYI [9], also the condition $\xi_k \in N(\mu, \sigma)$ will be satisfied in all cases where $k > 0$. The case of arbitrary distribution of ξ_0 will be considered.

For convenience let $\xi_k^* = \xi_k - \mu$ whereby (1) reads as follows:

$$\xi_{k+1}^* = p\xi_k^* + z_{k+1}\sigma\sqrt{1-p^2}\,. \tag{1'}$$

Serializing this deduction, one obtains:

$$\xi_n^* = p^n\xi_0^* + \sigma\sqrt{1-p^2}\,(p^{n-1}z_1 + p^{n-2}z_2 + \ldots + p\,z_{n-1} + z_1)$$

or, in another form:

$$\xi_n^* = p^n\xi_0^* + \sigma\, z_n^*\sqrt{1-p^{2n}}\,;\; z_n^* \in N(0,1)\,. \tag{2}$$

Let us denote the characteristic functions ξ_n^*, $(n = 0, 1, 2, \ldots)$; z_n^* by terms $\varphi_n(t)$ and $\psi_n(t)$; ξ_0^* and z_n^* being independent of each other,

$$\varphi_n(t) = \varphi_0(p^n t)\,\psi(\sigma t\sqrt{1-p^{2n}})\,. \tag{3}$$

The characteristic function of the standard normal distribution is known to be:

$$\varphi(t) = e^{-t^2/2}$$

consequently

$$\varphi_n(t) = \varphi_0(p^n t)\exp\left\{-\frac{1}{2}\left[\sigma\sqrt{1-p^{2n}}\,\frac{t}{2}\right]^2\right\}.$$

Since an arbitrary characteristic function $\varphi(t)$ is continuous on the number scale, and $\varphi(0) = 1$, as well as $|p| \leq 1$, thus

$$\lim_{n\to\infty}\varphi_n(t) = e^{\frac{-\sigma^2 t^2}{2}}\,; \qquad -\infty < t < \infty$$

where the right-hand side term is the characteristic function $\varphi(t)$ of the normal distribution of standard deviation σ and zero expected value.

Since according to A. RÉNYI [9] the distribution functions $F_n(x)$, $(n = 1, 2, \ldots)$ converge to a distribution function $F(x)$ at each point of continuity iff the characteristic functions $\varphi_n(t)$ of the distribution functions $F_n(x)$ converge, in case of $n \to \infty$, to a function $\varphi(t)$ continuous at $t = 0$. Now, $\varphi(t)$ is the characteristic function of $F(x)$, and the convergence of the characteristic functions $\varphi_n(t)$ is uniform in every finite interval.

However, from this statement directly follows that the distribution of ξ_n (provided the value of n is high enough) tends to the normal distribution of standard deviation σ and expected value μ.

Reducing the number of measurements by carrying out only every n-th measurement leads to the series $[\xi_{nk}]$ which also may be expressed by an autoregressive model:

$$\xi_{n(k+1)} = \mu + p'(\xi_{nk} - \mu) + z'_{k+1}\sigma\sqrt{1 - p^{12}} \tag{4}$$

$$p' = p^n$$

$$z'_{k+1} = \frac{p^{n-1}z_{nk+1} + p^{n-2}z_{nk+2} + \ldots + pz_{nk+n-1} + z_{nk+n}}{[(1 - p^{2n})/(1 - p^2)]^{1/2}}. \tag{5}$$

Remind that $z'_k \in N(0, 1)$ for any k value, and the z'_k are absolutely independent of each other.

Let us now determine how much information is contained in a series of n measurements as understood by SHANNON.

Since the series of discharge measurements has been assumed as approximately stationary, the wanted information content is independent of where the n samples have been taken from, thus, for simplifying the notations let the examined variables be $\xi_1, \xi_2, \ldots, \xi_n$. Thus $\underline{\xi} = (\xi_1, \xi_2, \ldots, \xi_n)$ is a random variable of n dimensions and of normal distribution, leading to the expected value:

$$E\xi = \underline{\mu}(\mu, \mu, \ldots, \mu)$$

and be D the covariance matrix of the variable ξ.

No difficulty arises from the assumption $\mu = 0$ since the expected value is not significant for the entropy:

$$H_\xi \cong H_{\xi + c}.$$

Concerning the covariance matrix D:

$$D = (d_{ij})_{nxn}; \qquad d_{ij} = E\xi_i\xi_j.$$

Making use of the expression

$$\xi_i = p^{j-i}\xi_j + \sigma\sqrt{1 - p^2}(p^{j-i-1}z_{j+1} + \ldots + pz_{i-1} + z_i); \qquad j > 1$$

obviously:

$$d_{ij} = \sigma^2 p^{|i-j|}$$

that is:

$$\mathbf{D} = \sigma^2 \begin{vmatrix} 1 & p & p^2 & p^3 & \ldots & p^{n-1} \\ p & 1 & p & p^2 & \ldots & p^{n-2} \\ p^2 & p & 1 & p & \ldots & . \\ . & . & . & & & . \\ . & . & . & & & . \\ . & . & . & & & . \\ p^{n-1} & . & . & & & 1 \end{vmatrix}$$

and the determinant of matrix **D**:

$$|D| = \sigma^{2n}(1 - p^2)^{n-1}.$$

Be the density function of the random vector-variable ξ:

$$f(\underline{x}) = f(x_1, x_2, \ldots, x_n)$$

and its entropy:

$$H_\xi = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} f(x)^2 \log f(x)\, dx_1 \ldots dx_n$$

leading to the density function of ξ:

$$f(\underline{x}) = \frac{1}{(2\pi)^{n/2}\,|D|^{1/2}} \exp\left[-\frac{1}{2}\, x^* D^{-1} x\right].$$

(**D** being a symmetrical, positive definite matrix, there is an orthogonal matrix **C** ($C^{-1} = C^*$) such that $D = C^*SC$ and **S** are diagonal and $|S| = |D|$).

Substitution $C\underline{x} = y$ transforms the expression $x^*D^{-1}\underline{x}$ into the square sum

$$y^*(D^{-1}C^*y) = y^*S^{-1}y\,.$$

Jacobi's determinant of the substitution:

$$\Delta = |C| = \pm 1$$

thus, the entropy of the random vector-variable is:

$$H_\xi = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} \frac{1}{(2\pi)^{n/2}|S|^{1/2}} \exp\left[-\frac{1}{2}y^*S^{-1}y\right] \cdot$$

$$\cdot \left[{}^2\log \frac{1}{(2\pi)^{n/2}|S|^{1/2}} - \frac{1}{2}y^*S^{-1}y \log e\right] dy_1 \ldots dy_n =$$

$$= -\log \frac{1}{(2\pi)^{n/2}|S|^{1/2}} + \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} \frac{\log e}{(2\pi)^{n/2}|S|^{1/2}} \frac{1}{2} y^*S^{-1}y \cdot$$

$$\cdot \exp\left[-\frac{1}{2}y^*S^{-1}y\right] dy_1 \ldots dy_1 .$$

The integral involved in the sum is of a value $n/2 \log e$, hence,

$$H_\xi = {}^2\log (2\pi)^{n/2}|S|^{1/2} + \frac{n}{2}\log e = n \log \sqrt{2\pi e} + \frac{1}{2}\lg |S|$$

$$|S| = |D|$$

$$H_\xi = nC + \frac{1}{2}\log |D| ; \qquad C = \log\sqrt{2\pi e}. \tag{7}$$

After these preliminary considerations the *optimum density* of the measurements may be examined.

3. Determination of the optimum interval between measurements

For a reliable description of hydrologic processes, theoretically, as long a time-series as possible is needed. Due to the limited length of period, the parameters of the population can only be determined with a certain *error*. The error could be reduced by *increasing the measurement period*, however, such an increase has both practical and economical limits.

On the other hand, by *discretizing* the essentially continuous process, the accuracy of parameter calculation *decreases*, by making, however, the measurements less frequent, the expenses of measurements, data processing and storage *decrease*.

Another aspect is the intended utilization of the *data existing and to be stored* — for example, in a data bank — or *still to be taken*. For less significant and less expensive projects it is quite unnecessary to strive to data of high all-around quality, long measuring period, with dense, or even continuous readings, etc., neither is worth while to integrally store them.

In planning measurements for new engineering projects, beside the expenses of sampling, it is advisable to take also the costs of designing the project into account, as pointed out by OGINK [6]. In his opinion, if the expenses of the measurements of e.g. water discharge do not significantly increase the overall design costs (Fig. 1/a), N_1 measurements during a relatively longer period might be planned, the more accurate parameters permitting to design a more economical engineering structure. On the other hand, the higher ex-

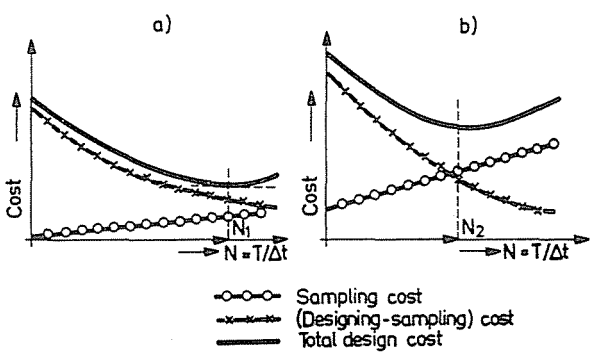

Fig. 1

penses of data acquisition (measurement) would justify to specify a lower number of data N_2 (see Fig. 1/b), counteracted by overdimensioning in the a priori strive to increased safety motivated by inaccurate hydrologic parameters. OGINK [6] did not try to resolve this obvious contradiction, so that his diagrams are rather illustrative for the complexity of the problem.

Thus, it can be stated that, beyond a certain limit, the *purely hydrologic statistical analyses* hit difficulties which can only be overcome by *hydro-economical considerations.*

Therefore, in the following, merely the starting point of *one of the possible* forms of a subsequent economical analysis will be presented, without striving to completeness.

In the following, the period τ between two consecutive measurements will be considered as *unit.* T will be the planned or specified measurement period of k units of time, thus, with notations in Fig. 1, $\Delta t = k\tau$. Hence, the number of measurements during the period T is $N = <T/k>$.

Let us introduce an arbitrary "benefit" function

$$h(T, k) = aH_{\xi_n, \xi_{n+k}, \ldots, \xi_{n-Nk}} - NK \tag{8}$$

which is the difference between the benefit proportional to the information obtained during the period T, in fact, the benefit from a more economical design of the engineering structure due to the higher accuracy of the design hydrological parameters and the costs of measurements, storage, processing, etc. emerging during the period T. K is the cost of *one single* measurement, and with previous notations:

$$<x> = \min [n : n \geq x]$$

i.e., the minimum integer, not lower in value than x.

According to the above, the total of N random variables

$$\xi_n, \xi_{n+k}, \ldots, \xi_{n+Nk}$$

are merely a sector of the autoregressive process, subject, according to relationship (4), to the equality $p' = p^k$.

Making use of relationships (6) and (7), one obtains expression (8) in the form:

$$h(T, k) = N\{a[c + \log \sigma] - K\} + \frac{N-1}{2} \log (1 - p^{2k}) .$$

Herein, the quotient $h(T, k)/T$ is the mean benefit per unit time if measurements are carried out at each k time units, while the limiting value

$$\lim_{T \to \infty} \frac{h(T,k)}{T} = h(k)$$

permits to calculate the maximum benefit per unit time.

It may easily be understood that

$$(h)k = \frac{1}{k}\{a(c + \log \sigma) - K + \log \sqrt{1 - p^{2k}}\}\,. \tag{9}$$

For convenience of computerization, function $h(k)$ will be replaced by a function $h(x)$ wherein $0 < x < \infty$ is an arbitrary real number.

It is familiar that the function $h(x)$ interpreted in the interval $(0, \infty)$ has a positive maximum iff

$$a(c + \log \sigma) - K > 0\,; \qquad c = \log \sqrt{2\pi e}\,.$$

The above quantity is the benefit from a single measurement of the process which obviously must be positive. But this statement involves also that if it is worth while to measure — which is, however, evident — then there exists a procedure which, in the above sense, is the optimum, i.e., it results in the *maximum benefit.*

Let us introduce the notation

$$a(C + \log \sigma) - K = A\,; \qquad p^2 = B\,.$$

In this case

$$h(x) = \frac{1}{x}(A + \log \sqrt{1 - B^x})\,. \tag{10}$$

Since $h(x)$ is continuous in the interval $(0, \infty)$,

$$\lim_{x \to 0} h(x) = -\infty \tag{11}$$

and

$$\lim_{x \to \infty} h(x) = 0\,. \tag{12}$$

The function $\log (1 - B^x)^{1/2}$ takes on all negative values in the interval $(0, \infty)$, therefore it exists $x_0 > 0$, i.e., $A + \log (1 - B^{x_0})^{1/2} = 0$ thus $h(x_0) = 0$, further, $\log (1 - B^x)^{1/2}$ being monotonous increasing,

$$h(x) > 0\,, \qquad \text{if} \qquad x > x_0\,.$$

On the other hand, from the foregoing, considering the relationship (12), it follows that the function $h(x)$ has indeed a *positive maximum.*

4. Conclusions, tasks for the future

For the case of a hydrologic problem describable by the mathematical model of a simple autoregressive process, a possible way of thought for optimizing the measurements has been outlined above. The problem of *auto-*

regressive processes of higher order may be solved in a quite similar way, but at significantly more calculation work.

From among the factors entering in the benefit function $h(k)$ to be considered as the final result of the present statistical analysis, the determination of p and σ is not difficult, p being the first autocorrelation coefficient and σ the standard deviation of the process, since they are directly obtained from the data.

Another problem to be investigated is the determination of factors of cost $\underline{K}$ and of proportionality $\underline{a}$. It may be assumed, for example, that the factor $\underline{a}$ may be brought into hydrologic statistical relation with NYQUIST's interval, in case of certain hydrologic processes, by pondering the different information needed for determining the expected value, variance, covariance, etc. at a given accuracy. Besides, however, also economy parameters have to be considered.

The cost $\underline{K}$, like factor $\underline{a}$, may be considered as a multivariable function of a type depending on the actual problem.

Thus, if data are available (the scope of data bank), function $\underline{K}$ evidently will not include the measurement costs, but costs of storage, data processing etc. depending on the utilization. Therefore the determination of the maximum of function $h(x)$ is a problem to which only definite (non-general) solutions exist.

Summary

After a survey of publications on the determination of optimum measurement intervals and information set of the hydrologic data series, the solution is shown to depend on the type of the process, hence different model types have to be applied for pure random, stochastic or deterministic processes. A possible solution is presented for the case of a simple autoregressive process, and the benefit function suiting optimization is derived. Finally, some aspects of the determination of the function are presented.

References

1. DIMAKSYAN, A. M.: Hydrological Devices. Publ. Gidrometeoizdat, Leningrad, 1972. (in Russian).
2. DYHR-NIELSEN, M.: Loss of Information by Discretizing Hydrologic Series. Hydr. Papers No. 54. Colorado State Univ. 1972.
3. EAGLESON, P. S.—SHACK, W. J.: Some Criteria for the Measurement of Rainfall and Runoff. Water Resources Research, Vol. 2. No. 3. 1966.
4. KOTELNIKOV, V. A.: Calculations in Electrical Engineering. RKKA Publ. 1933, Moscow (in Russian).
5. VAN DE NES, TH. J.—HENDRICKS, M. H.: Analysis of a Linear Distributed Model of Surface Runoff. Agr. Univ. Wageningen, Report No. 1. 1971.
6. OGINK, H. J. M.: Determination of an Effective Sampling Interval for Hydrologic Time-Series. VITUKI publ. No. 10. Budapest, 1974.
7. QUIMPO, R. G.—YANG, J.: Sampling Consideration in Stream Discharge and Temperature Measurements. Water Res. Research Vol. 6. No. 6. 1970.

8. REIMANN, J.: Determination of the Optimum Number of Measurements in Case of Random Processes.* Manuscript. Budapest, 1976.
9. RÉNYI, A.: Probability Calculus.* Budapest, 1962.
10. RUPPERT, M. L.: Statistical Method for Determining the Optimal Number of Measurements in River Cross Section. GGI publ. No. 150, 1968, Leningrad (in Russian)
11. YEVJEVICH, V.: Stochastic Processes in Hydrology. Water Res. Publ. Fort Collins, Colorado, 1972.
12. SZÖLLŐSI-NAGY, A.: The Effect of Sampling Interval on the Information Content of Hydrographs.* Water Resources Publ. Budapest, 1972.

Prof. Dr. Imre V. NAGY, H-1521, Budapest

* In Hungarian