Nordic Hydrology 6, 1975, 28–42
Published by Munksgaard, Copenhagen, Denmark

ECONOMIC OPTIMUM RECORD LENGTH
Defined in the Context of a Statistical Decision Approach

SVEN JACOBI

Nielsen & Rauschenberger, Consulting Engineers, Ltd.
Copenhagen, Denmark

Uncertainty in hydrologic design parameters is reflected as an increase in expected project costs. The Bayesian or statistical decision approach produces a minimum-cost decision for a specific design, yielding the expected opportunity loss, EOL, a measure of the uncertainty inherent in the decision process. The uncertainty stems from the fact that the population value of the design parameters is unknown. Additional information can be obtained by collecting more data, and the value of these data is measured by the decrease in EOL. But there are costs involved in getting additional data: the cost of continued data collection and the cost of delaying the construction of the project (benefits foregone). The expected economic optimum record length is defined at the point where the marginal benefits of additional data are equal to the marginal costs of obtaining those data. The theory is applied to sediment load data used to design the sediment storage portion of a reservoir. Based on an economic efficiency criterion, the expected optimum record length is found to be 12 years, given that 5 years of data are available at the time of the analysis. Objectives like environmental quality and social benefits are disregarded in this analysis.

It has always been the case that the design engineer is faced with the problem of taking a particular action and making decisions under uncertainty. The decision-maker should never neglect to ask, "Is there enough information in the data I possess, or should I collect more data in order to reduce the uncertainty of the

decision?" Inherent in every decision process is an amount of uncertainty about the "true" state of nature; in most cases the population parameters of the data series used for design purpose must be treated as unknown.

The statistical decision approach is a method for choosing and evaluating design alternatives for a project, when the "true" state of nature or other factors are not known. Davis (1971) was the first to make a comprehensive study in the field of hydrology applying the Bayesian decision method. The effect of uncertainties is taken into consideration through the use of probability density functions. This type of decision theory focuses on the decision to be made and not on the hydrologic parameters as an end result; from that point of view the statistical decision analysis is appealing to the design engineer. Also, such analysis makes it possible to estimate the dollar value of the uncertainties considered in the problem. It is this latter characteristic which is the basis for the data valuation study presented in this paper.

A more economically efficient design might be obtained if the available data sample could be extended by postponing the proposed project and collecting additional data. But there is a price in connection with this increased information; namely, the cost of having a sampling station in operation and the cost of delaying construction of the project. This latter cost will show up as benefits foregone. The trade-off between the information gained as a result of the addition of extra data points, measured in monetary terms, and the cost associated with the data collection, leads to a definition of expected economic optimum record length.

THEORETICAL BACKGROUND

The statistical, or Bayesian, decision approach is simply a procedure for applying logical thinking and cannot be called a strict method. According to Howard (1966) and Davis (1971), the different steps in the decision procedure can be outlined as:

A. Define the decision to be made and identify the alternatives.

B. Form the goal function, which includes the variables describing the "state of nature".

C. Derive stochastic properties of state variables.

D. Select best alternative by:

   1) calculating the expected value of the goal function for each alternative, and

2) choosing the alternative which minimizes the expected value of the goal function.

The design taken up for investigation is the size of storage allocated for sediment deposition in a reservoir. Thus, the alternatives in the decision process involve the variable, V, the volume of a sediment pool big enough to accumulate sediment throughout the lifetime of the reservoir. The hydrologic parameters treated as being uncertain, the state variables in the problem, are the mean and variance of the annual sediment load series used for designing. Sediment load data recorded at the U. S. Geological Survey Otowi Bridge sampling station on the Rio Grande are used in connection with the Corps of Engineers Cochiti Lake project just downstream of the sampling station. The goal function, $G(V|\mu,\sigma^2)$, is a penalty function. It indicates the excess cost that has to be paid because of either a realized overdesign or an underdesign of the sediment storage part of the reservoir.

Inherent in every decision process is an amount of uncertainty about the "true" state of nature, in this case the population parameters, mean, μ, and variance, σ^2. This uncertainty is reflected by assigning a probability density function to the assumed values of the population parameters. The function is called $f(\mu,\sigma^2|n)$. This distribution depends on the given observed sample of length, n, as will be shown later. A decision is made by choosing the alternative, V^*, that minimizes the expected value of the goal function:

$$R(V^*) \equiv \underset{V}{\text{Min}} \iint G(V|\mu,\sigma^2) \cdot f(\mu,\sigma^2|n) \cdot d\mu \cdot d\sigma^2, \tag{1}$$

In the literature $R(V^*)$ is often called Bayesian Risk and V^* the Bayesian solution. After the decision is made, the uncertainty analysis can be pursued.

If the "true" values of the state variables were known, this information would yield V_t, the design alternative that gives the minimum future cost. Having used V^* instead of V_t, an opportunity loss has resulted. The expected opportunity loss is calculated as

$$EOL(n) \equiv \iint [G(V^*|\mu,\sigma^2) - G(V_t|\mu,\sigma^2)] \cdot f(\mu,\sigma^2|n) \cdot d\mu \cdot d\sigma^2. \tag{2}$$

It is seen that EOL is a function of the available data sample.

The last steps in the outline are then:

E. Evaluate uncertainties and find the worth of additional data by:
   1) determining the expected opportunity loss, EOL, and

2) finding the reduction in EOL by collecting more data. The calculations outlined in steps C to E.1 are performed with the new data included, and the reduction in EOL is the difference between the EOL values obtained before and after the addition of extra data. Increased information concerning the unknown population parameters is measured as the reduction in EOL.

F. Find the net worth of additional information in monetary terms as the decrease in EOL minus the cost of obtaining those extra data.

The economic optimum record length is defined as the size of the data sample which provides the analysis with a maximum net worth of the additional data. In economic terminology that point is reached where the marginal benefit curve intersects the marginal cost curve. In the paper by Moss (1972), the concepts of expected optimum record length are taken up for a thorough discussion. Fig. 1 shows that data should be collected as long as their marginal worth exceeds its expected marginal cost.

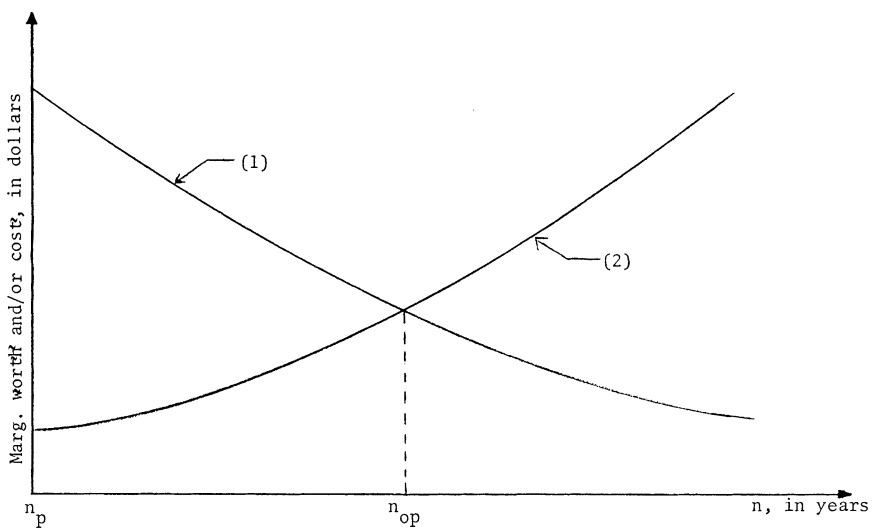

Fig. 1.

Marginal worth and marginal cost versus record length: (1) worth of additional data; (2) cost of additional data, includes both cost of collecting data and of project postponement; n_p = present time; and n_{op} = optimum record length.

With regard to a recommendation of project postponement in order to collect more data, it should be kept in mind that decision-making in water resources planning is often a more complex process than the present investigation might indicate, where only part of an economic efficiency criterion is applied. According to the Water Resources Council (1973) the following main objectives should be considered by federal, state, and local governments for planning the use of a nation's water resources and related land:

1) Economic efficiency,
2) Quality of environment, and
3) Social benefits.

In a realistic decision process in connection with allocation of reservoir storage for different purposes, all three objectives have to be considered. In the present study only a portion of the first objective about an economically efficient project has been used as the determining factor in the different findings reported. For example, the economic consequences of taking the flood control purpose of the Cochiti Reservoir into consideration are omitted.

It is recalled that the probability density function, $f(\mu,\sigma^2)$, and accordingly EOL are both dependent on the given data sample. Therefore, the question arises how to find the values of EOL when future "unknown" data are used to extend the available record.

In this paper a method recommended by Moss and Dawdy (1973) is employed. This approach combines a Monte Carlo simulation technique and an expected-value criterion. The combination method can be applied in many different ways depending on the investigator's particular case; here the purpose is to achieve a measure of the expected value of the expected opportunity loss, EVEOL, when future data are used to extend an available sample. The Monte Carlo approach has the serious restriction that the hydrologic parameters must be known or assumed prior to the analysis. This deficiency can be avoided, as will be explained in the following.

For a particular set of assigned values of the population parameters μ and σ^2, it is possible to generate synthetic annual sediment loads, Q_i, using the model recommended by Matalas (1967):

$$Q_{i+1} = \mu + \varrho(1) \cdot [Q_i - \mu] + \sqrt{1-\varrho(1)^2} \cdot \sigma \cdot \varepsilon_{i+1}, \tag{3}$$

which is the general equation for simulation of data which are normally distributed and possess autocorrelative dependency represented by the lag-one serial correlation coefficient, $\varrho(1)$. This model is valid under the condition of

stationarity, i.e., the distribution of Q_i is identical to Q_{i+k} for all integer values of k. ε_{i+1} is a random normal component with zero mean and unit variance, and independent of Q_i.

Through the application of Eq. (3), a sequence of n_2 future events can be generated. This sample of data pooled with the observed sample of a length n_1, yields a new sample mean and variance, which are used to develop a revised distribution of μ and σ^2, $f(\mu,\sigma^2|n_1+n_2)$. Bayesian Risk and EOL calculations can now be carried out in the regular manner, as described earlier. What has been obtained is a value of EOL given the synthetic record, or $(EOL(n_1+n_2)|\mu,\sigma^2)$. By repeating the above procedure a sufficient number of times for randomly selected values of μ and σ^2 (covering the "possible" range of these parameters in accordance with the *a priori* distribution), a set of EOL's can be defined. The average of this set is an estimate of the expected value of expected opportunity loss (EVEOL). The average value is found as a weighted average of the EOL values using the original distribution of μ and σ^2, derived from the observed sample, to weight the different synthesized values. It is seen that this method is a blend of a Monte Carlo simulation technique and an expected-value criterion. That means,

$$EVEOL(n_1+n_2) \equiv \sum_{i=1}^{n_\mu} \sum_{j=1}^{n_{\sigma^2}} (EOL(n_1+n_2)\,|\mu_i,\sigma_j^2) \cdot f(\mu_i,\sigma_j^2) \cdot \triangle_i \cdot \triangle_j, \qquad (4)$$

where n_μ and n_{σ^2} are the number of intervals into which the ranges of the mean and variance respectively are divided, and $\triangle_i$ and $\triangle_j$ are, respectively, the sizes of the ith interval of the mean and the jth interval of the variance.

The above analysis can now be repeated for each consecutive year of data or groups of data added beyond the given observed sediment load series (for example $n_2 \equiv 1$, $n_2 \equiv 5$, $n_2 = 10$, $n_2 \equiv 25$, etc.). The worth-of-data curve is hereby defined, and the economic optimum record length can be found if the cost of collecting the data and cost of project postponement can be specified, as will be shown in the next section.

APPLICATION OF THE THEORY

One of the best known sediment-oriented problem areas in the world is the Rio Grande Basin, especially in the New Mexico region. Harmful deposition in major reaches along the rivers in the system has been experienced. The U. S.

Army Corps of Engineers was authorized to investigate and control the situation, and as a part of this it was decided to construct the Cochiti Reservoir on the Rio Grande with a two-fold purpose: sediment trapping and flood control. The Cochiti Dam is one of the world's largest earthfill dams, and the construction is just about to be completed. For this study, the emphasis is on the sediment deposition part of the reservoir.

The problem is to determine the design alternative (the size of the sediment storage), which minimizes the cost due to building the reservoir either too large or too small. An overdesign naturally results in an increased cost measured in dollars per acre-foot of excess storage; an underdesign requires a removal of sediment to restore the situation or a loss of reservoir storage allocated for other purposes (flood control, water supply, etc.). It is the trade-off between these two types of costs which are the basis for the economic minimization problem. A detailed explanation of the goal function, $G(V|\mu,\sigma^2)$, and its explicit functional form is given in the study by Jacobi (1974, Ch. 4.2).

The U. S. Geological Survey station at Otowi Bridge near San Ildefonso, New

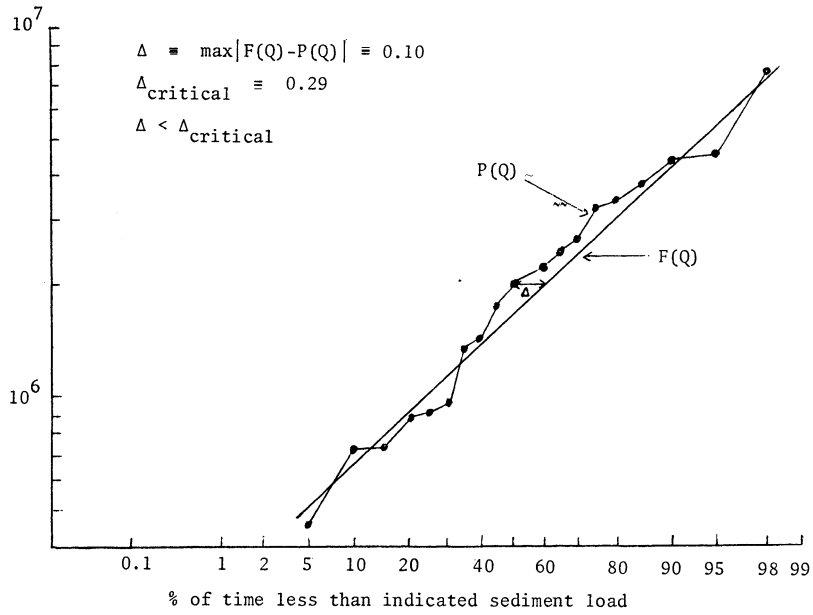

Fig. 2.
Kolmogorov-Smirnov test for normal distribution of annual sediment load.

Mexico, has been recording sediment loads in the Rio Grande for more than 20 consecutive years. In Fig. 2 the data from this station is plotted on logarithmic-probability graph paper. A Kolmogorov-Smirnov test is performed to justify the assumption that the logarithm of the annual sediment load follows a normal distribution.

The parameters, mean μ and variance σ^2, of the normal distribution constitute the state variables, which describe the "state of nature" and are not known with certainty.

Raiffa and Schlaifer (1961, p. 300) and Benjamin and Cornell (1970, p. 628) derive the joint distribution of the random variables μ and σ^2. It is shown that the distribution, given a set of sample statistics, $\bar{x}$ and s^2, takes the form of a so-called normal-Chi square density function,

$$f(\mu,\sigma^2|n,\bar{x},s^2) \equiv$$

$$\sqrt{\frac{n}{2\pi\sigma^2}} \cdot e^{-\frac{n(\mu-\bar{x})^2}{2\sigma^2}} \cdot \frac{\left(\frac{n}{2}\right)^{\frac{n-1}{2}}}{\Gamma\frac{n-1}{2}} \cdot \frac{1}{\sigma^2} \cdot \left(\frac{s^2}{\sigma^2}\right)^{\frac{n-1}{2}} \cdot e^{-\frac{ns^2}{2\sigma^2}} \quad . \tag{5}$$

With the goal function and the probability density function, $f(\mu,\sigma^2|n)$, defined, EOL calculations can be carried out for an increasing number of data points included in the underlying sample. When the size of the sample exceeds the 20 years of observed data, the EVEOL concept – as explained earlier in this paper – is used to define the EOL curve. Fig. 3 gives the result of these computations.

The EOL curve can be assumed to follow a decreasing power function, and the functional form is found to be,

$$EOL(n) \equiv \frac{50 \times 10^6}{n} \tag{6}$$

where EOL is expressed in dollars, and n in years.

Assume now that the given data base consists of 5 years of observed sediment loads. How many additional years of data – if any – would it be economically worthwhile to include in the given data base with respect to the Cochiti sediment storage project?

The expected benefits as a result of extra data are measured as decrease in EOL. That means, the total benefit function is

$$B(n) \equiv EOL(5) - EVEOL(n) \tag{7}$$

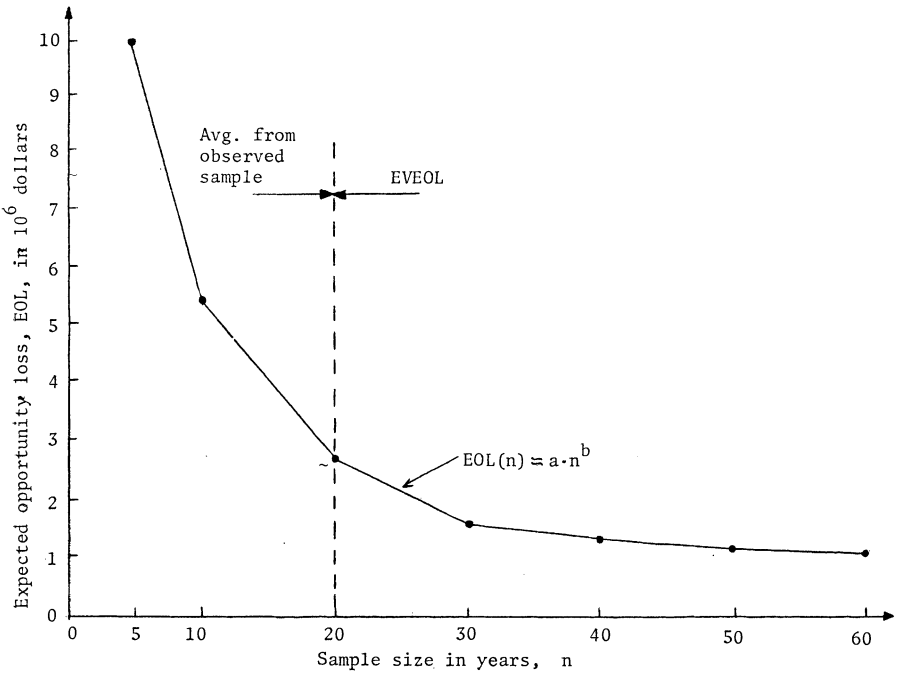

Fig. 3.
The relationship of EOL(n) to sample size, n.

or substituting from equation (6) yields,

$$B(n) \equiv 10^7 - \frac{5.0 \times 10^7}{n} \tag{8}$$

The functional form of the total cost curve is a sum of two terms: the annual expenditures of operating the sediment load sampling facilities, c_1, and the benefits foregone, c_2, by not having the project in operation. Jacobi and Richardson (1974) show that the total cost function takes the form:

$$C(n) = c_1 \cdot \frac{\left(\frac{1}{1+r_2}\right)^{n-5} - 1}{\left(\frac{1}{1+r_2}\right) - 1} + c_2 \cdot \frac{\left(\frac{1+r_1}{1+r_2}\right)^{n-5} - 1}{\left(\frac{1+r_1}{1+r_2}\right) - 1} \tag{9}$$

where

$n \equiv$ number of data points in the sample; expressed in years ($n \geqq 5$); i.e., (n-5) = number of years of construction delay.

$c_1 \equiv$ average annual cost of having a fully equipped sediment sampling station in operation. Estimated to be $ 12,000.

$c_2 \equiv$ benefits foregone parameter. An estimated value of $ 275,000 anually is found through written communication wih the U. S. Army Corps of Engineers, Albuquerque District, New Mexico.

$r_1 \equiv$ a percentage which indicates an increasing trend of the benefits foregone parameters because of continued development of the middle Rio Grande flood plain. r_1 is estimated to be 4 %. The figure is reached by assuming that the increase in benefits foregone is proportional to the population growth in the region. U. S. Census Reports show approximately a four percent annual increase in the population of the Rio Grande Valley taken as an average of rural, farm, and urban areas.

$r_2 =$ discount factor, which is used to bring future costs back to present time; this factor should include an inflation allowance. r_2 is assumed to be 2 %.

For the given values of the cost factors and r_1 and r_2, Eq. (9) reduces to,

$$C(n) \equiv 0.6 \cdot 10^6 \cdot (1-0.98^{n-5}) + 13.75 \cdot 10^6 \cdot (1.02^{n-5} - 1) \tag{10}$$

In Fig. 4 the total benefit function, Eq. (8), and the total cost function, Eq. (10), are plotted versus number of data points included in the sample.

The economic optimal point is defined as that point where the difference between total benefits and total costs (net worth of data) is as large as possible. That means, the function $D(n) \equiv B(n) - C(n)$ is subject to maximization with respect to n. D(n) is drawn in Fig. 4. In order to find the point, the derivative $\frac{dD(n)}{dn}$ is set equal to zero, or

$$\frac{dB(n)}{dn} = \frac{dC(n)}{dn} \tag{11}$$

Substituting Eqs. (8) and (10) for B(n) and C(n) in Eq. (11) and taking the derivatives yields,

$$\frac{5.0 \times 10^7}{n^2} = (-0.6 \times 10^6) \cdot (0.98^{n-5}) \cdot \ln(0.98) + (13.73 \times 10^6) \cdot (1.02)^{n-5} \cdot \ln(1.02) \tag{12}$$

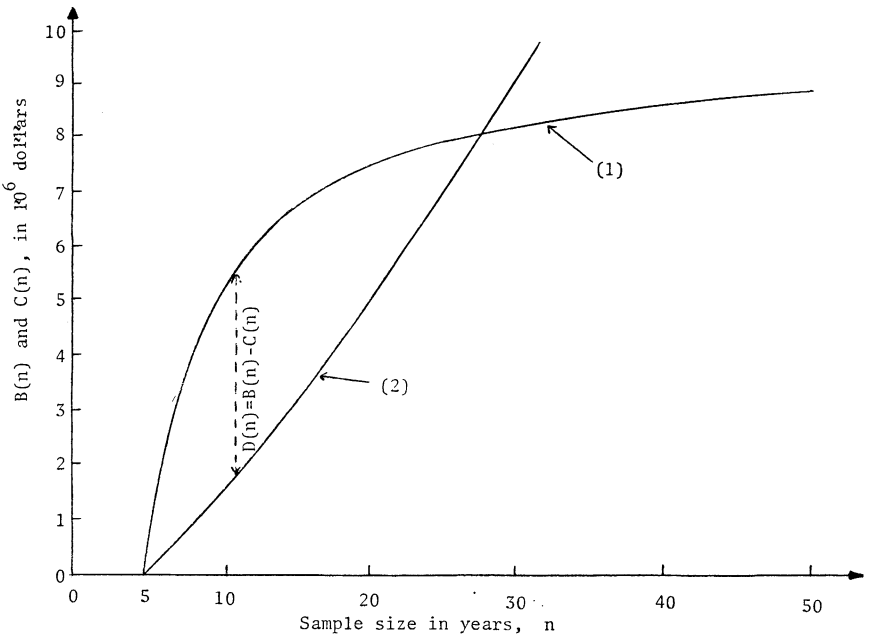

Fig. 4.
Total benefits and total costs curves plotted against length of sample: (1) total benefits, Eq. 8; and (2) total costs, Eq. 10.

Rearranging the terms results in the equation,

$$2.72 \cdot (1.02)^{n-5} + 0.12 \cdot (0.98)^{n-5} - 500/n^2 = 0 \qquad (13)$$

Eq. (13) cannot be solved explicitly, but by applying the "trial and error" method the equation is found to be satisfied for $n = 12.4$. The left side of Eq. (13) is equal to -0.24 for $n = 12$, and equal to 0.33 for $n = 13$.

In an economic analysis this optimum point is also obtained where the marginal benefit curve intersects the marginal cost curve as explained in connection with Fig. 1. Fig. 5 shows those curves in this particular case study and gives the graphical solution to the problem.

From the above analysis the conclusion can be drawn that the expected economic optimum size of the sample used in the decision process is approximately 12 years

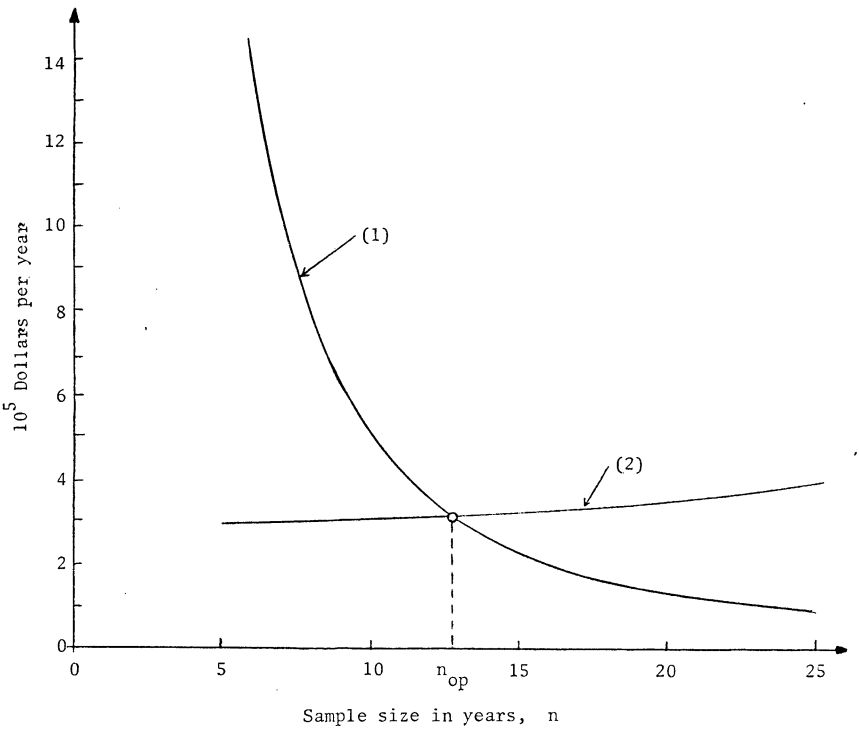

Fig. 5.
Marginal benefits and marginal costs curves plotted against length of sample: (1) marg. benefits; (2) marg. costs; and n_{op} = economic optimum recor l length.

RESULTS AND DISCUSSION

It is recalled from Eq. 6 that the EOL curve is continuously decreasing as a function of the sample size used for designing, which is the same as saying that information is gained about the unknown state parameters for every new additional data point. This implies that a fundamental difference of interest between the data gatherer and the design engineer or builder is very often apparent. While the data expert wishes to improve the reliability and accuracy of his product, thus requiring more years of data collection, the engineer generally wants to or has to get on with his work.

In order to satisfy both interest groups, the concept of expected economic

optimum record length was introduced. It is found at the point in time where the marginal cost of data (cost of sampling plus cost of project postponement) is equal to the marginal benefits provided by the additional data. For annual sediment load data in connection with allocation of deposition storage in the Cochiti Reservoir, it is shown that the expected economic optimal sample has a length of 12 years, providing that 5 years of data already exist. That is, a project delay of 7 years is recommended so the remaining data can be obtained. However, it should be kept in mind that this record length is what the decision-maker anticipates to be the economic optimal, given the information he already has in the observed 5 year sample. This further implies that there is no guarantee that, after additional data has been collected, the recommended economic optimum record length will not differ from the first one. The result of such an analysis is conditioned, reasonably enough, on the available information at the time of decision.

The present study shows how the design engineer, after a political decision has been made to build a project, might be forced to suggest a postponement of the construction for economic reasons. If the political decision-makers, in spite of this, want the project initiated, it is possibly because they consider its intangible benefits (for instance, regional development, recreation, environmental quality) more important than the extra costs created by an economic sub-optimal design.

NOTATION

Symbol	*Meaning*
$B(n)$	Total benefits in economic analysis as a function of sample size.
$C(n)$	Total costs in economic analysis as a function of sample size.
c_1	Annual cost of having a sediment sampling station in operation.
c_2	Benefits foregone parameter (in dollars per year).
$D(n)$	Difference between total benefits and total costs as a function of sample size (net worth of data).
$EOL(n)$	Expected opportunity loss as a function of sample size (in dollars).
$EVEOL(n_1 + n_2)$	Expected value of expected opportunity loss given n_1 observed data plus n_2 future data (in dollars).
$F(Q)$	Theoretical distribution used in Kolmogorov-Smirnov test.
$f(\mu,\sigma^2\|n)$	Probability density function for the state parameters given a sample of size n.

$G(V\|\mu,\sigma^2)$	Goal function dependent on the design alternative and the state parameters.
n	Size of observed data sample.
P(Q)	Sample distribution used in Kolmogorov-Smirnov test.
Q	Annual sediment load data in logarithmic form.
R(V*)	Minimum Bayesian Risk (in dollars).
r_1	Percentage which indicates annual population growth in the Rio Grande Valley.
r_2	Discount factor including inflation allowance.
s^2	Sample variance of observed data set in logarithmic transformation.
V	Design alternative for the size of the sediment storage expressed in tons per year.
V*	Minimum Bayesian Risk solution.
V_t	Design alternative which yields minimum value of the goal function when the true values of the state variables are known.
$\bar{x}$	Sample mean of observed data set in logarithmic transformation.
ε_i	Normal random component with zero mean and unit variance.
μ	Population mean of annual sediment load in logarithmic form (state variable).
σ^2	Population variance of annual sediment load in logarithmic form (state variable).
$\varrho(1)$	First order autocorrelation coefficient.

REFERENCES

Benjamin, J. R. & Cornell, C. A. (1970) Probability, statistics and decision for civil engineers. McGraw Hill Book Company, Inc., New York.

Davis, D. R. (1971) Decision making under uncertainty in systems hydrology. Tech. report no. 2, Hydrology and water resources interdisciplinary program, University of Arizona, Tucson, Arizona.

Howard, R. (1966) Decision analysis: applied decision theory. 4th Intl. conference of operations research, John Wiley, pp. 55–71, New York.

Jacobi, S. (1974) Economic worth of sediment load data in a statistical decision framework. Ph. D. dissertation, Civil Engineering Department, Colorado State University, Fort Collins, Colorado.

Jacobi, S. & Richardson, E. V. (1974) Economic value of sediment discharge data. Hydrology Paper No. 67, Colorado State University, Fort Collins, Colorado.

Matalas, N. C. (1967) Mathematical assessment of synthetic hydrology. *Water Resources Research,* v. 3, n. 4, pp. 937–945.

Moss, M. E. (1972) Expected optimum record length as a basis for hydrologic network design. Proc. of Int. Symp. on water resources planning, Mexico City, Dec. 4–8, 1972.

Moss, M. E. & Dawdy, D. R. (1973) The worth of data in hydrologic design. *Journal of the Highway Research Board.* In preparation.

Raiffa, H. & Schlaifer, R. (1961) Applied statistical decision theory. MIT Press, Boston.

Water Resources Council (1973) Principles and standards for planning water and related land resources. Federal Register, Washington, D. C., v. 38, no. 174.

Received April 1974.

Address:
"Nielsen & Rauschenberger"
Consulting Engineers, Ltd.,
Lundtoftevej 7,
DK-2800, Lyngby,
Denmark.