LOSS OF INFORMATION BY DISCRETIZING HYDROLOGIC SERIES

by

MOGENS DYHR-NIELSEN

October 1972

HYDROLOGY PAPERS
COLORADO STATE UNIVERSITY
Fort Collins, Colorado

54

LOSS OF INFORMATION BY DISCRETIZING HYDROLOGIC SERIES

by

Mogens Dyhr-Nielsen*

HYDROLOGY PAPERS
COLORADO STATE UNIVERSITY
FORT COLLINS, COLORADO 80521

OCTOBER 1972 No. 54

*At the date of writing of this paper: Graduate Research Assistant, Department of Civil Engineering, Colorado State University, Fort Collins, Colorado

TABLE OF CONTENTS

TABLE OF CONTENTS — (continued)

ACKNOWLEDGEMENTS

This paper is based on the Ph.D. dissertation submitted by M. Dyhr-Nielsen. Dr. V. Yevjevich served as chairman of the graduate committee. Valuable suggestions were offered by Dr. C. F. Nordin, Dr. D. C. Boes and Dr. M. M. Siddiqui.

Financial support obtained from the U. S. National Science Foundation Grant No. GK11564, OWRR Project B-030-COLO, and OWRR Project A-009-COLO is gratefully acknowledged.

Mogens Dyhr-Nielsen

ABSTRACT

LOSS OF INFORMATION BY DISCRETIZATION

A procedure is presented for quantitative evaluation of the loss of information when parameters of continuous stochastic processes are estimated on the basis of discrete sampled data. Three discretization procedures are considered:

(1) Discrete point sampling, where the process is sampled at periodic time intervals as a series of instantaneous values.

(2) Average sampling, where the process is sampled as a series of average values.

(3) Quantization of the variable, where its values are pooled into class intervals.

The decision theoretical concept of "expected information loss", based on a linear loss function, is used as the measure of information content.

For each discretization procedure general expressions for the expected information loss in estimating mean, variance and autocovariance are found as functions of the discretization interval, the length of the sampling period, and the mean, variance and autocovariance of the continuous process. For normal processes the expected information loss is determined for estimation of probabilities of extremes, the mean number of runs, the mean run length and the mean run sum.

A stream flow series is analyzed to show the applicability and potential of the approach. The loss due to quantization is found to be negligible in most practical cases. A small sampling interval is essential to prevent a large information loss about extremes and runs, but is of less importance for estimation of mean and variance.

Average sampling introduces significant losses of information due to the biasing effect inherent in the sampling procedure. With the exception of the mean, a sample of instantaneous values contain more information about the parameters investigated than a sample of average values, taken over the same sampling interval.

Mogens Dyhr-Nielsen
Civil Engineering Department
Colorado State University
Fort Collins, Colorado 80521
May, 1972

PREFACE

The majority of most important variables in water resources conservation, development and control are hydrologic continuous random variables (any value in a given range occurs in nature), and their space and time series are usually continuous stochastic or periodic-stochastic processes. Because simple approaches have been developed in probability theory, mathematical statistics and stochastic processes for treating discrete variables and discrete processes, and because most calculations are oriented to the digital computer, a trend has developed to transform continuous random variables and series into discrete series of continuous variables or into discrete series of discrete variables. An example of this present trend is the development of optimization techniques in water resources. A range of reservoir level or storage fluctuations between its two boundaries of empty and full reservoir is divided into "states", as layer states and boundary states, and considered as discrete random events. The continuous reservoir input or output time series and the time series of reservoir levels or storage volumes are basically treated as discrete series. There is bound to be a discrepancy between mathematical and computer expediencies of using the discrete instead of the continuous approach and the true conditions, while conserving, controlling and developing water resources. Therefore, a loss of information and various uncertainties are associated with this contemporaneous trend of discretization of hydrologic variables and series.

In the research project on applications of stochastic processes in hydrology and water resources, sponsored by the U. S. National Science Foundation, Grant No. GK-31512X, the problem of the loss of information by discretization is considered as a subject worthwhile of investigation, particularly in assessing how good the results of this discretization may be in comparison with the solution for the continuous case. This paper, as the Ph.D. thesis by Mogens Dyhr-Nielsen, represents an initial inquiry into the effects and consequences of discretization, assumed to represent an additional uncertainty in planning and operating water resources projects. More research is deemed necessary in the future, particularly related to practical examples in current applications.

An assessment is needed for the penalty which is paid as new uncertainties are generated by various calculations in water resources optimization, based on discretization of variables and series. For the penalty in the form of a loss of information, as this thesis demonstrates, the question also arises of which concept of information should be used in hydrology and water resources. The particular question is whether an economic optimization of information procurement for each particular water resource project should be the basis of collecting hydrologic information, or whether another criteria should be developed because of yet unidentified future uses of collected long-range hydrologic data. Therefore, the study presented in this report should be viewed as a first attempt at throwing light on where the present trend of discretizations, and consequently, of procurement of hydrologic information and water resources optimization, may lead in the future. By introducing this paper, in the capacity of project principal investigator and advisor, it does not imply that all statements and approaches used in this paper are shared. The progress on the subject of this paper will be much faster under a diversity of concepts, approaches and attempts.

Vujica Yevjevich
Professor of Civil Engineering
Colorado State University

CHAPTER 1

INTRODUCTION

With an increasing interest in rational development and management of water resources, the complexity and sophistication of analytical tools to solve water related problems have increased accordingly. Such techniques as systems analysis, mathematical modeling, stochastic processes, and time series analysis have been introduced in the different fields of hydrology in the last decade, often transferred from other sciences such as communication engineering. It has become evident, however, that the value of such modeling techniques may be limited in hydrologic applications because of insufficient data to identify the model parameters. Therefore, an efficient data collection system is of prime importance for successful use of advanced mathematical techniques. This study treats one aspect of this data problem: the use of discretized data as the basis for stochastic modeling of continuous hydrologic variables and processes.

1.1 Purpose of Study

Most measurements of hydrologic phenomena may be considered to be continuous random variables occurring as continuous stochastic time series. Hence, both the marginal distribution $f_X(x)$ of a hydrologic variable X and its realization in time X(t) are basically continuous curves as shown in Fig. 1.1.

Such series can be observed and analyzed continuously by various analog methods. However, most techniques used for data collection and processing usually introduce discretization which destroys the continuous character of the actual series.

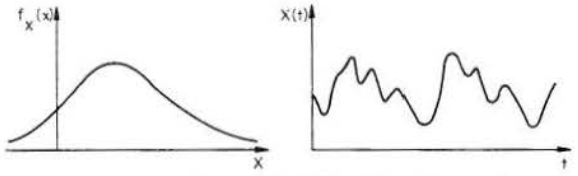

Fig. 1.1 Distribution and realization of a continuous stochastic process.

The continuous random variable itself may be discretized by dividing its total range into class intervals and considering, for example, the midpoint of each interval as representative for all the values in the interval. In this way, the continuous variable is transformed to a discrete variable. Similarly, the continuous time series may be observed at discrete, intermittent times or transformed by averaging the actual process over a given time period, giving rise to a discontinuous realization of the underlying process.

Even though the sampling introduces discretization, it is still possible to make inference about such properties of the continuous stochastic process as mean, variance, autocovariance and extremes. The ultimate aim of a discretization procedure would be to insure that the results obtained are statistically nearly equivalent to the results obtained based on the continuous sample.

Although discretization is used almost exclusively in analysis of hydrologic data, very few attempts have been made to evaluate the effect of such a transformation of the original continuous process. By taking samples of water quality or sediment load once or twice a day, how much can actually be inferred about these processes, compared to a continuous sampling? Does such sampling destroy important information in the sample? It is obvious that the answer to these questions depends upon what one is looking for. If the property of interest is only the mean concentration of some pollutant, once-daily samples, taken over a long period might be sufficient. With increasing time of sampling, the error of the estimate will converge to zero. But if inference has to be made on extreme properties such as the number of times the concentration is above or below a certain standard, daily observations might be seriously inadequate. Such properties are governed by the very fast fluctuations in the process, and a too wide sampling spacing cannot detect them no matter how long the sample is. In such cases the proper choice of discretization is of utmost importance.

Similarly, the practice of publishing data as daily, monthly or annual averages is sufficient for some purposes, whereas much information about for instance extreme values may be lost by averaging. This is not so serious if the averaging was performed under the data processing and if documentation of the continuous series still exists, but if the averaging is done through the sampling procedure, as in the case of a cumulative raingauge, the information is lost forever. In general the introduction of discretization introduces a certain loss of information about the continuous process, and it is the purpose of this investigation to develop tools to quantify this loss as a function of the discretization interval.

1.2 Outline of Investigations

First, different possible quantitative measures of information content are investigated. Several definitions have been used in the literature, and their relevance for hydrologic studies are discussed. Based on this, the decision theoretical concept of information content has been selected as most advantageous.

Next, three of the most commonly used discretization procedures are presented: the discrete point sampling technique, where the process is observed instantaneously at discrete time intervals: the average sampling technique, where the discretization is performed by averaging the process over a certain time period: and the quantization of the random variable by pooling its values in a given interval into a common value.

The loss of information in estimation of selected stochastic properties has been determined as a function of the discretization interval for each of the three sampling techniques. One of the most fundamental characteristics of a process is its marginal probability distribution function, and first the effect on estimation of this function is investigated.

The first and second order moments, or the mean and variance, are the most important parameters of a process, and particularly for normal processes these parameters describe uniquely the marginal distribution. Therefore, the loss of information is investigated on these two parameters.

To describe the time dependence of a process, the basic property is described by the autocovariance $\gamma(u)$. It is possible to postulate simplified mathematical models for the autocovariance function, such as Markov models, fractional noise models, or broken-line models, but these restrict the applicability of results. It has been shown by Quimpo (1967) that, for instance, that the first and second order Markov models cannot be used for modeling of daily streamflows. It has, therefore, been decided to present the results in terms of the autocovariance function itself without any model restrictions. Based on this, it is possible to obtain results for any model.

Some properties of particular importance in hydrology and water resources are investigated. The probability that the flood exceeds a certain stage u is a standard problem in hydrology, and it is evident that too large a sampling interval may miss some exceptional events. For this reason the effect of discretization on the estimation of this probability is investigated.

Often, it is not only the instantaneous intensity of an extreme event that is of importance, but also its duration and total magnitude, expressed as the run-length L_u and the run-sum S_u of the exceedance. The design of a flood retention reservoir strongly depends on these random variables. Furthermore, the expected number N_u of occurrence of events exceeding a given level in a given sample or time period may be included in characterizing the exceedances. A complete description of these properties of a continuous stochastic process is still lacking. A rather general approximation to the exceedance probability for a stationary normal process has been developed by Ditlevsen (1971) and the expected numbers of L_u, S_u and N_u have been found by Cramer and Leadbetter (1967). This study is limited to the effect of discrete sampling on these four properties. Even though this does not give full description of sampling effects on extreme properties, it does provide an indication of the importance of frequent sampling when such characteristics are of interest.

Finally, a set of streamflow series is analyzed to demonstrate the procedure and the effect of discretization for particular cases. It should be pointed out, however, that the procedure is general, and may be similarly applied to other hydrologic processes, such as water quality or rainfall.

CHAPTER 2

REVIEW OF LITERATURE

The purpose of this literature review is twofold: first, to find a useful concept of information content to form the basis for the present study; second, to address the investigation more directly to the effect of discretization on estimation efficiency.

The results of the search for a concept of information are presented in detail in Chapter 3. Here, it is sufficient to mention the three major concepts:

(l) R. A. Fisher's information concept was first presented in his two papers (Fisher, 1921, 1925) and later in a more descriptive version (Fisher, 1966). Fisher's concept was developed directly for use in statistical estimation theory which is also the object of this study.

(2) C. E. Shannon's information concept (Shannon 1948) actually was developed from Hartley (1928), but since it was Shannon who showed its usefulness in the theory of communications it is usually credited to him. Numerous texts in communication theory have introductory presentations of Shannon's information concept. The clearest one for readers outside the field of communication theory may be McMullen (1968). Shannon studied the information in transmitted messages in a communication system; however, it was also shown (McMillan, 1953, Khinchin, 1957, Kullback, 1959) that the concept is useful in statistical applications.

(3) The decision theoretical concept of information is based on the work of Savage (1950) and introduced by Raiffa and Schlaifer (1961). A less mathematically oriented introduction to this concept may be found in Raiffa (1968) or Benjamin and Cornell (1970). The general scope of decision theory is so broad that the estimation problem may be considered a special case of application. A presentation more specifically oriented toward mathematical statistics is given by DeGroot (1970). It should further be mentioned that Davis (1971) has studied the applicability of decision theory in hydrology in great detail, including a thorough literature review.

The second part of the literature review relates to hydrologic studies which consider the information loss by sampling. Although there exists a substantial number of papers studying the spatial sampling problem as a basis for data network planning, relatively few studies consider the effect of the in-station sampling in time, and of those available, most are more concerned about how long a time to sample instead of how often.

Matalas and Langbein (1962) studied Fisher's information content in the mean for stochastic series satisfying the first-order Markov model. They defined as a measure of information the effective number of observations, i.e., the number of observations of an independent process that contains the same information as the dependent process, and they found that the information content in a dependent series is smaller than the content in an independent series of the same length and with same variance.

This concept was used for the mean of a second-order Markov model by Reiher and Huzzen (1967) and Quimpo (1969). Yevjevich (1972) presents the effective number of years for estimating the variance of a first-order Markov model. Common for these studies is the assumption of a Markov model, which makes the results less applicable to closely sampled data when this assumption does not always hold.

A more direct attempt to study the loss due to increased sampling interval was presented by Knisel and Yevjevich (1967). They studied series of average values such as daily, weekly, monthly, and annual flows and analyzed independent residual series, where both the deterministic component and the dependency in the stochastic component were removed. They concluded that the variance of the estimate of the mean increases with an increase in the averaging interval; however, this conclusion is misleading. When the mean is estimated it should make no difference whether the sampling average is estimated based on daily average, monthly average, or annual average values; the estimate will always be the same, and, therefore, its sampling variance should be independent of the sampling interval. The conclusion above is a result of the removal of the dependent component, and this removal is not permissible because the information content in the total series must be considered to include both the dependent and the independent component. If this is done, the information content does not vary with the sampling interval.

A paper by Quimpo and Yang (1970) addresses the same problem as this dissertation, the proper

choice of the sampling interval. Based on their results, the conclusion can be drawn that the information content increases with the increased sampling interval, a somewhat staggering result. The reason is that their concept of information content was developed based on discrete point sampling, whereas the data analysis was based on average value sampling in form of daily, 2-day, 3-day, etc. gloed, thereby changing the characteristics of the underlying process. Strictly speaking, the optimal sampling interval according to Quimpo's paper would be infinite, i.e., one should never sample.

A different approach to the sampling problem was proposed by Eagleson and Shack (1966) based on concepts used in spectral analysis. They proposed to choose a sampling frequency large enough to include "most of the spectrum" and to use as a criterion for the sampling frequency that the spectral density at the cutoff frequency is only 5 percent of the density in the origin. This approach attempts to obtain "all" the information and does not consider the problem losses when the sampling interval increases.

The papers mentioned above are the only hydrology oriented studies of the problem that has been found. Some interesting work on discrete sampling has been done in queueing theory by Riordan (1951) and Benes (1961) in relation to the measurement of traffic loads in telephone systems. However, the processes they considered are different from those usually encountered in hydrologic time series. Because the problem of time series sampling is equivalent to the non-random sampling problem, an abundance of references could be found in the statistical literature. However, it is more practical to refer to these when they are actually used in the following text. Here only the work of Tick and Shaman (1966) is mentioned, which illustrates the effect of discrete sampling on the estimation of crossing properties. They show that a sufficiently frequent sampling is essential for the preservation of such properties as the expected number of crossings of a given level, the number of maxima in a given time period, etc. This work plus the works of Cramer and Leadbetter (1967) and Ditlevsen (1971) have pointed out the importance of the high-frequency terms in the spectrum when extreme and crossing properties are of interest and hence the importance of a sufficiently close sampling in such cases.

CHAPTER 3

CONCEPTS OF INFORMATION CONTENT

The word information is frequently found in hydrologic literature. Generally speaking all research in hydrology is aimed at increasing information about hydrologic phenomena, even though one seldom finds attempts to define this increase in quantifiable terms.

Before talking about information it is important to emphasize that this word can have different, although related meanings. The dictionary defines information in two ways:

(1) The act of informing or communication of knowledge.
(2) Knowledge derived from study or knowledge about a specific situation.

The first definition is of importance for problems concerning transmission of knowledge or messages such as in electrical communication theory. The latter seems more relevant for studies of knowledge of messages at hand; this problem is represented in statistical inference, where knowledge about stochastic properties has to be inferred from samples. However, the advantage of using a particular definition, when it comes to applications, is not so clear, and both definitions have actually been applied in statistical inference.

There is a close relation between the concept of information and the concept of uncertainty. Availability of information should, hopefully, reduce uncertainties, and one might, therefore, define information as the decrease in uncertainty about a parameter that has taken place after the information has been obtained, for example, by sampling of a stochastic process. As increase of information is associated with decrease of uncertainty, the two concepts are complementary, and the amount of uncertainty in an estimate of a parameter should be inversely proportional to the amount of information defined as knowledge about the parameter.

The information under consideration in this study is that contained in a sample estimate of a particular parameter of the continuous process such as mean, variance or extreme properties. Sample estimates will always be subject to some uncertainty, even if the estimate is based on a continuous sample, and this gives an upper limit to the amount of information about the parameter that can be obtained.If the estimation is based on a discrete sample, the behavior of the process between the sampled values is unknown, and an additional uncertainty is introduced relative to the continuous sample. It is this additional loss of information that is of principal interest in this study.

When there can be some dichotomy in talking about information qualitatively, the arbitrariness increases in trying to define a quantitative measure of information content. In general, any monotonic function that maps a high information into a high numerical value is a candidate as a measure of information. In any particular situation, the selection of appropriate definitions depends on the use of the information that is to be extracted from the data.

Three different concepts of information content are presented: Fisher's information content, Shannon's information content and a concept about expected information loss that has been developed in statistical decision theory. All three may be used as objective measures of information, but the decision theoretical approach has the advantage that it includes the effect of a biased estimator on the information content; and it provides a basis for economic assessment of the value of information. The latter becomes particularly important in practice where a tradeoff between the cost of lost information and the cost of sampling and processing the data is necessary in order to determine the optimal sampling rate. Although it is not the intent of this investigation to address directly the problem of selection of an optimal sampling rate, it is important to use an information concept that can be applied in such cases. Therefore, the decision theoretical concept of information loss will be selected for use throughout this study as a measure of information.

3.1 Fisher's Information Content

About 50 years ago R. A. Fisher (1921, 1925) presented the first major attack on the problem of extracting the maximum amount of information out of a given set of data. He defines a high amount of information as a large amount of knowledge about a particular parameter, that has to be estimated based on the sample.

Fisher's information content in n independent observation of the random variable X about a population parameter α is defined by (Rao, 1965),

$$I_f = n \int f_X(x;\alpha) \left[\frac{\partial \ln f_X(x;\alpha)}{\partial \alpha}\right]^2 dx , \qquad (3.1)$$

where $f_X(x;\alpha)$ is the probability density function of X. Equation 3.1 is an expression for the maximum amount of information that is contained in the sample. If the n observations are combined into a statistic $\hat{\alpha}$ to be used as an estimate of $\hat{\alpha}$, it can be shown that the information in $\hat{\alpha}$ is smaller than or equal to I_f of Eq. 3.1. If it is equal, $\hat{\alpha}$ is called an efficient estimate.

In general, the estimate $\hat{\alpha}$ is not efficient, and in such cases a loss of information about α occurs by replacing the original sample by $\hat{\alpha}$. The information in an inefficient estimate $\hat{\alpha}$ can be found by considering $\hat{\alpha}$ as a single observation taken from the sample distribution $f_A(\hat{\alpha};\alpha)$ of $\hat{\alpha}$,(Fisher, 1966):

$$I_f(\hat{\alpha}) = \int f_A(\hat{\alpha};\alpha) \left[\frac{\partial \ln f_A(\hat{\alpha};\alpha)}{\partial \alpha}\right]^2 d\hat{\alpha} \qquad (3.2)$$

The distribution of $\hat{\alpha}$ varies with both the estimator itself and the sample size. For large samples, however, many estimates have an approximate normal distribution, specified by the mean μ and the variance V of $\hat{\alpha}$ only (Cramer, 1951).

If the estimate is unbiased, $\mu = \alpha$, and if it has a bias B, $\mu = \alpha + B$. In general for large samples

$$f_A(\hat{\alpha}) \simeq \frac{1}{\sqrt{2\pi V}} \exp\left[- \frac{(\hat{\alpha} - (\alpha + B))^2}{2V} \right] . \qquad (3.3)$$

This distribution has the first two moments exact even for small samples, but the shape is an approximation only.

Combining Eqs. 3.2 and 3.3 it is easily shown that Fisher's information content becomes

$$I_f = \frac{(1 + \frac{\partial B}{\partial \alpha})^2}{V} . \qquad (3.4)$$

Equation 3.4 clearly shows the relation between information and uncertainty. A large sample variance of the estimate $\hat{\alpha}$ is associated with a small information content. The bias of the estimator enters the equation as the derivative with respect to α. This implies that if for example the bias is proportional to α, I_f is independent of the absolute magnitude of B. In general, however, an estimate with a large bias should be expected to contain less information than one with small bias, given the sample variances are equal. This will not be the case in the example above, and this is a drawback for this particular measure of information content.

Actually, Fisher never used his concept to study biased estimates. He stated explicitly (Fisher, 1925): "-knowledge of the exact form of the distribution of $\hat{\alpha}$ will enable us to eliminate any disadvantages from which a statistic might seem to suffer by reason of bias."

In short, he does not consider the bias an inaccuracy, because, if known, it can be compensated for exactly. If the bias is assumed to be zero, Fisher's information content reduces to the well known equation

$$I_f = \frac{1}{V} , \qquad (3.5)$$

This expression has been used in hydrology by Matalas and Langbein (1962) and Knisel and Yevjevich (1967). If the bias of a given estimate is estimated in advance, a new unbiased estimate can be formed by subtracting this bias. For this new estimate, the information content may be expressed as a function of V^2 only, and Eq. 3.5 can be used. However, in this study it has been found more advantageous to consider the information as a function of both bias and variance of the estimate.

Based on its similarity to Eq. 3.5, the following expression may be proposed as a measure of information in the estimate $\hat{\alpha}$:

$$I_{mse} = \frac{1}{mse(\hat{\alpha})} , \qquad (3.6)$$

where $mse(\hat{\alpha})$ is the mean square error of $\hat{\alpha}$. As $mse(\hat{\alpha}) = V + B^2$, this concept takes into account the influence of both bias and variance on the information content. It should be noted, however, that this concept is a somewhat subjective choice, and it is not directly related to Fisher's concept.

A major disadvantage in using I_f or I_{mse} as measures of information is, that they are not easily related to the value of information, as expressed in monetary terms. From a practical point of view this is of significance because it is the value of the information that ultimately is of interest, when a decision about a given discretization level has to be made.

3.2 Shannon's Information Concept

Although the information concept discussed here is mostly associated with Shannon (1948), it actually is almost as old as Fisher's concept and was introduced in 1928 by Hartley (1928).

The word information, in this theory, is used in a sense that should not be confused with the concept Fisher used. Where Fisher associated a high amount of information with a high amount of knowledge or accuracy, Shannon associates high information with a high amount of uncertainty. The reason for this is that the fundamental problem of communication theory is that of reproducing at one point either exactly or approximately a message selected at another point from a set of possible messages. The word message should here be understood in general terms, including not only written or spoken statements but also a statistic computed from a sample. In this context, a transferred message that is certain does not contain any information, as it did not increase our knowledge. On the other hand, the transfer of a very unlikely message carries a large amount of information. In this way, the information content is related to the measure of uncertainties or the probabilities of the messages. Hartley proposed somewhat subjectively that the natural choice of measure of information content in a message is a logarithmic function of its probability. Shannon states that it is near to our intuitive feelings as to the proper measure, as it satisfies the properties mentioned above, and it is additive for independent messages. The main support for its usefulness comes, however, from its mathematical simplicity and from some important theorems, that has been developed about the capacity of transmission channels. Furthermore, its mathematical equivalence to the concept of entropy in statistical mechanics and thermodynamics seems appealing.

If one considers the information not in a single message, but in the information source that produces the messages, it is natural to define the information content in this source as the expected information in a message,

$$I_s = - E[\ell n\, P[\hat{\alpha}]] \tag{3.7}$$

$$= - \sum_{i=1}^{n} P[\hat{\alpha}_i]\, \ell n\, P[\hat{\alpha}_i] , \tag{3.8}$$

$\hat{\alpha}$ discrete and

$$I_s = - \int_A f_A(\hat{\alpha})\, \ell n\, f_A(\hat{\alpha})\, d\hat{\alpha} \tag{3.9}$$

for $\hat{\alpha}$ continuous where $f_A(\hat{\alpha})$ is the density function of the messages. The minus sign is added to make I_s positive. This is the definition of Shannon's information content, or, as it is called also, due to its mathematical similarity with the physical concept of entropy, the entropy of the information source.

In this study the sampled part of a stochastic process is considered as an information source that transmits uncertain messages in the form of estimates $\hat{\alpha}$ of the parameter α. The Shannon information content in the sample is then determined by Eq. 3.9, where $f_A(\hat{\alpha})$ is the sampling distribution of $\hat{\alpha}$.

If, like above, it is assumed that the statistic $\hat{\alpha}$ is distributed approximately normal with mean, α+B and variance V, where B is the bias, then

$$I_s = - E[\ln [\frac{1}{\sqrt{2\pi V}} \exp (- \frac{(\hat{\alpha} - (\alpha+B))^2}{2V})]] \tag{3.10}$$

$$= - E[\ln \frac{1}{\sqrt{2\pi}} - \frac{1}{2} \ln \sigma^2 - \frac{(\hat{\alpha} - (\alpha+B))^2}{2V}] \tag{3.11}$$

$$= - \ln \frac{1}{\sqrt{2\pi}} + \frac{1}{2} \ln V - \frac{1}{2} \frac{V}{V} \tag{3.12}$$

$$= k + \frac{1}{2} \ln V \tag{3.13}$$

Equation 3.13 shows that in general Shannon's information concept does only include the effect of sampling variance on the information content, and that the effect of bias is not accounted for. This is a serious disadvantage, because in some cases, such as in estimating probabilities of extremes based on discrete sampled data, the bias is a major contributor to the uncertainty about the population value. In such cases, it is important that the bias enters as a factor in the definition of information content.

Although the concept of entropy has been used in many applications in communications theory, in economics (Theil 1967), and in mathematical statistics (Kullback, 1959), it has been concluded that it is not the most advantageous approach in this study. It does not account for information loss due to bias, and, similar to Fisher's concept, it is not related to the value of the information. This is clearly stated in the beginning of Shannon's paper, as this aspect is

considered irrelevant to the communication engineering problems. However, in hydrologic practice, knowledge of the value of the information is essential in order to evaluate if it is economically feasible to analyze a process with a given discretization, and it has, therefore, been necessary to search for an information concept that takes this into account.

3.3 Expected Information Loss

Statistical decision theory has been developed as a tool for making decisions about actions to be taken when the state of the world is uncertain, but when further information about it may be obtained by experimentation. The objective of a decision theoretical analysis is to identify a course of action (which may or may not include experimentation) that is consistent with the decisionmaker's own preferences for consequences, as expressed by a numerical loss function and with the weights he attaches to the possible states of the world, as expressed by numerical probabilities (Raiffa and Schlaifer 1961). It is not a rigid theory, but rather a general framework to advance evaluation of the consequences of given actions. It is build on two main foundations, namely, that the consequence of all possible actions for all possible outcomes can be specified through a loss function, and that the probability distribution of all possible states is known.

Here the concept is used to evaluate the effect of different discretization intervals on the estimation of a given population parameter α. Due to sampling variation, the estimate, $\hat{\alpha}$, will generally be different from the true value, α, and when $\hat{\alpha}$ is used as if it actually is the population value, a certain loss will occur. This loss might be expressed in monetary values, for example, as the losses that occur in a water supply system due to a wrong estimate of the average supply; here a simple linear loss function is used

$$l(\hat{\alpha}) = k|\hat{\alpha} - \alpha| \qquad (3.14)$$

so that the loss is directly proportional to the deviation of $\hat{\alpha}$ from its true value, α. It might be considered the first term in a Taylor expansion of the actual loss function. An inclusion of higher order terms would not increase the mathematical difficulties to any great extent, but it has not been judged necessary to do so here.

It is not known in advance what value $\hat{\alpha}$ will have, as it is a random variable with stochastic properties determined by both the basic population from which the sample is taken and by the discretization performed in the sampling procedure. However, if this distribution is found, the expected information loss is defined as the expected value of the loss function:

$$\bar{l}(\hat{\alpha};\Delta t,\Delta x) = kE[|\hat{\alpha}(\Delta t,\Delta x) - \alpha|] \qquad (3.15)$$

By sampling a given stochastic process with sampling intervals Δt and Δx, Eq. 3.15 gives a measure of the loss of information about α that occur due to this particular discretization.

This concept will account for the effects of both bias and variance of the estimate. Furthermore, it may assess the value of information and form the basis for an optimization of the discretization interval, if a realistic loss function can be developed. The optimal sampling interval should then minimize the sum of the expected loss and the cost of data collection and processing.

To find the expected loss of Eq. 3.15, the distribution of the estimate $\hat{\alpha}$ must be developed as a function of the discretization interval. This problem may be solved in three ways: by an exact mathematical solution in closed form, by an empirical solution using the Monte Carlo method, or by an approximation to the exact solution. For most parameters under consideration, the first alternative seems impossible; the second will probably give solutions close to the exact; however, a severe disadvantage is that for practical reasons the number of cases that may be considered is limited.

The last alternative should give the mean and the variance of the sample estimate and assume that the sampling distribution may be approximated by a normal distribution, specified by this mean and variance. This assumption is satisfied for large samples (Cramer, 1951). Furthermore, as the distribution of the estimate is only used to develop the expectation of the loss function, too much sophistication in the derivation of the sampling distribution seems unnecessary. Because it is possible to express the mean and the variance of an estimate as a function of Δt under very general assumptions about the process, the latter approach is selected for this study.

Given the bias $B(\Delta t)$ and variance $V(\Delta t)$ of the estimate $\hat{\alpha}$ of α for sampling interval Δt, the

computation of the expected information loss $l(\hat{\alpha},\Delta t0$ is straight forward, Fig. 3.1:

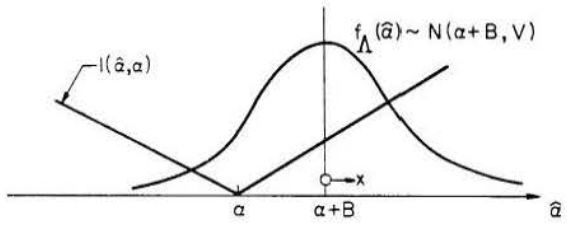

Fig. 3.1 Computation of $\bar{l}(\hat{\alpha},\Delta t)$.

$$\bar{l}(\hat{\alpha};\Delta t,\Delta x)=E[k|\hat{\alpha}-\alpha|]$$

$$= k\{\int_{-\infty}^{-B} -(x+B)\frac{1}{\sqrt{V}}\phi(\frac{x}{\sqrt{V}})\,dx$$

$$+\int_{-B}^{\infty}(x+B)\frac{1}{\sqrt{V}}\phi(\frac{x}{\sqrt{V}})\,dx\}$$

where $\phi(u)$ is the density function of the standardized normal distribution. The single terms give

$$-k\int_{-\infty}^{-B} x\frac{1}{\sqrt{V}}\phi(\frac{x}{\sqrt{V}})\,dx = -k\sqrt{V}\int_{-\infty}^{-B/\sqrt{V}} u\phi(u)\,du$$

$$= k\sqrt{V}\,\phi(\frac{B}{\sqrt{V}})$$

$$-kB\int_{-\infty}^{-B}\frac{1}{\sqrt{V}}\phi(\frac{x}{\sqrt{V}})\,dx = -kB\int_{-\infty}^{-B/\sqrt{V}}\phi(u)\,du$$

$$= -kB\,\Phi(-\frac{B}{\sqrt{V}}) = kB(\Phi(\frac{B}{\sqrt{V}})-1)$$

where $\Phi(u)$ is the standardized normal distribution function.

$$k\int_{-B}^{\infty} x\frac{1}{\sqrt{V}}\phi(\frac{x}{\sqrt{V}})\,dx = k\sqrt{V}\int_{-B/\sqrt{V}}^{\infty} u\phi(u)\,du$$

$$= k\sqrt{V}\,\phi(\frac{B}{\sqrt{V}})$$

$$kB\int_{-B}^{\infty}\frac{1}{\sqrt{V}}\phi(\frac{x}{\sqrt{V}})\,dx = kB\int_{-B/\sqrt{V}}^{\infty}\phi(u)\,du$$

$$= kB\,\Phi(\frac{B}{\sqrt{V}})$$

so that

$$\bar{l}(\hat{\alpha};\Delta t,\Delta x) = k\{2[B(\Delta t,\Delta x)\,\Phi(\frac{B(\Delta t,\Delta x)}{\sqrt{V(\Delta t,\Delta x)}}) + \sqrt{V(\Delta t,\Delta x)}\,\phi(\frac{B(\Delta t,\Delta x)}{\sqrt{V(\Delta t,\Delta x)}}]\cdot B(\Delta t,\Delta x)\} \quad (3.16)$$

When B and V are found as functions of Δt, the expected information loss can easily be found from standard tables of $\phi(u)$ and $\Phi(u)$.

This approach establishes the measure of information content used in this study. The extension to more complicated loss functions is easy. If these are expressed as higher order polynominals, the expected uncertainty loss can still be expressed in terms of $\phi(u)$ and $\Phi(u)$. If not, numerical integration may be necessary.

CHAPTER 4

SAMPLING PROCEDURES

The purpose of sampling a stochastic process is to enable an inference about the population from which the samples are generated, and then predicting of behavior of future realizations of this process. Intuitively, it seems that the only way to obtain "all" information contained in a realization of a continuous process is to perform a continuous recording. However, it seems just as obvious that if a continuous process is sampled sufficiently closely, "most" of the relevant information should be present in the discrete sample by obtaining a close approximation by straight line interpolation between the observations.

Some complexities and cost in recording and processing continuous data have prevented continuous sampling and processing techniques from being currently used, in particular in hydrologic data collecting. Even when the data are collected continuously, as is the case with some stream flow gaging, the original recordings are not easily accessible. Finally, the wide-spread use of digital computers instead of analog computers has made direct use of continuous data impractical.

The traditional presentation of hydrologic data as discrete series made it feasible to apply stochastic techniques which have been developed for discrete series, neglecting the fact that the underlying process is actually continuous. However, when the sampling is discrete both in time and for the variable, it is still possible to make some statistical inferences about the underlying continuous process. With tools for generation of continuous processes becoming more and more available (Mejia, 1971) it is a logical approach to analyze hydrologic time series as continuous stochastic processes.

A brief review of three common sampling and processing techniques is given below with an assessment of implications the use of these techniques have upon the inference about a continuous process.

4.1 Discrete Point Sampling

If a continuous process is observed only as instantaneous values at discrete points in a time period T_s, this is referred to as the discrete point sampling, Fig. 4.1. It is the classical nonrandom sampling scheme, and it has been extensively treated in the literature.

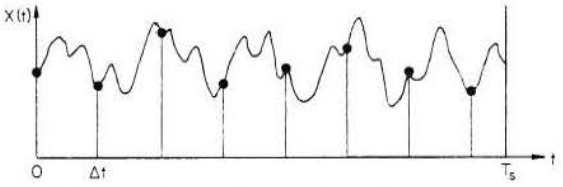

Fig. 4.1 Discrete point sampling scheme.

Sampling of this type is used in monitoring water quality variables, where water samples are collected at certain times of the day and analyzed for sediment and mineral content. Another example is once-a-day observations of water levels in rivers or ground water aquifers, a gaging procedure that still is widely used outside the U.S.A. Although wind and temperature data may be measured continuously, the instruments usually are read only at discrete time intervals, and therefore these data falls into this category. Many other examples are easily found in hydrologic data collections.

Finally, the use of digital computers in the data processing phase has led to the introduction of discrete sampling of series such as stream flow stages that previously were sampled continuously.

In general, discrete sampling of continuous processes may have the time distance between observations either constant or random. Although random sampling may be considered, it is rather unfeasible when one is interested in estimating the correlation structure as a function of time lags. Hence, only sampling with regular intervals is considered, by further assuming that the sampling is extended uniformly over the total sampling period, T_s. Throughout the study, T_s is considered constant.

In certain cases, in which economic or other constraints limit the number of observations to be taken, it might be profitable to sample only part of the period T_s with smaller intervals between sampling points in order to get a better estimate of the high frequency properties of the process. This problem has been studied by Jones (1962) and Parzen (1963), but is not considered here.

Only the loss of information by uniform discrete sampling over the complete time interval T_s is investigated. This problem was studied early in the century in the fields of electrical engineering and

communication theory. The practical advantages of transmitting continuous signals as discrete impulses raised the question of how much this procedure distorted the received message. Nyquist (1924) showed that for a very wide class of stochastic processes a sufficiently close discrete sampling will preserve all the information, and that it is possible to recover the original continuous trace from the discrete sample. This result is presented in the following version of the famous sampling theorem: "The spectrum of a frequency limited signal, which has no spectral components above the frequency f_c, is uniquely determined by its samples taken at uniform intervals less than $\Delta t_c = 1/2f_c$ apart."

The proof of this theorem can be made by considering a convolution between the original continuous signal $f_c(t)$ and a periodic unit impulse function $\delta(t)$ with frequency, $f_s = 1/\Delta t$ (Lahti, 1968). In the time domain, this operation will result in the discrete sampled process, $f_s(t)$, Fig. 4.2. It can be seen that this operation produces a periodic copy of the original spectrum if $f_s > 2f_c$. If, however, $f_s < 2f_c$ the tails of the sampled spectrum will overlap, and the resulting spectrum is distorted.

To recover the original spectrum from the discretized series one should neglect everything outside the frequency interval $-f_c < f < f_c$ or, mathematically speaking, filter the sampled spectrum by unit height in the interval $-1/2\Delta t_c < f < 1/2\Delta t_c$ and zero outside. If a similar operation is performed on the discrete series in the time domain, the original continuous series will be recovered, as there is a one to one correspondence between the signal and its Fourier transform. It should be noted here that, in general, the original series is not recovered by interpolating straight lines between the sampled points. It has been shown by Tick and Shaman (1966) that if only this is done, the appearance of the recovered process in terms of run properties such as number of crossings will be distorted, even if one samples with a frequency twice the limiting frequency f_c. However, for spectra usually encountered in hydrology, where the spectral density decreases with an increase of frequency, this distortion is relatively small. Furthermore, it is important to emphasize that even if the reconstruction of the original signal is incomplete due to the approximate linear interpolation, no distortion is present in the spectrum. Therefore, the inference based on the spectrum of the discrete sampled process will be identical to inference based on the continuous sampled process, if the sampling frequency is above the critical value $2f_c$. In this sense, the discretized process is exactly equivalent to the continuous process. Therefore, in this study where the characteristics to be studied are functions of the spectrum, one need only to measure the loss of information relative to a sufficiently dense sampled process instead of the original continuous signal. This allows for some simplification in the analysis, as only discrete processes are to be studied.

A critical point in the sampling theorem is the assumption of a frequency limited signal. It can be shown that a finite signal can never be frequency-limited, so no realizable signal can satisfy this assumption exactly. However, experience has shown that most realizations of hydrologic time series can be considered approximately frequency-limited, so that components with frequencies above a certain value can be neglected without introduction of any significant error. Spectral densities of instantaneous rainfall intensities show a fast decrease with increased frequency (Eagleson and Shack, 1966), and this will be even more pronounced for stream flow records that are the filtered output of the rainfall input into the catchment system. In general, the catchment will filter out the highest frequencies in the rainfall spectrum and produce an even closer approximation to the frequency-limited case. Therefore, it is assumed that a sampling frequency of twice the largest significant frequency will satisfy the conditions of the sampling theorem, with an acceptable error, and, therefore, be equivalent to the continuous sample.

If the sampling frequency is smaller than the critical frequency, three things will change the information that was contained in the complete sample and give rise to an information loss:

(1) Increase of the sampling interval will usually decrease the autocovariance between the observations, making them more independent. This tends to increase the information in the sample.

(2) Reduction of the number of sampled points in the time period T tends to decrease the information content.

(3) The high frequency components in the sample will not be detected if the sampling interval is too wide. This may give rise to bias in certain estimates, decreasing the information content.

The loss of information is due to the joint effect of these three factors. It should be noted that if 1) dominates 2) and if 3) is negligible, the information content in the discretized sample may be greater than in the continuous sample. In such

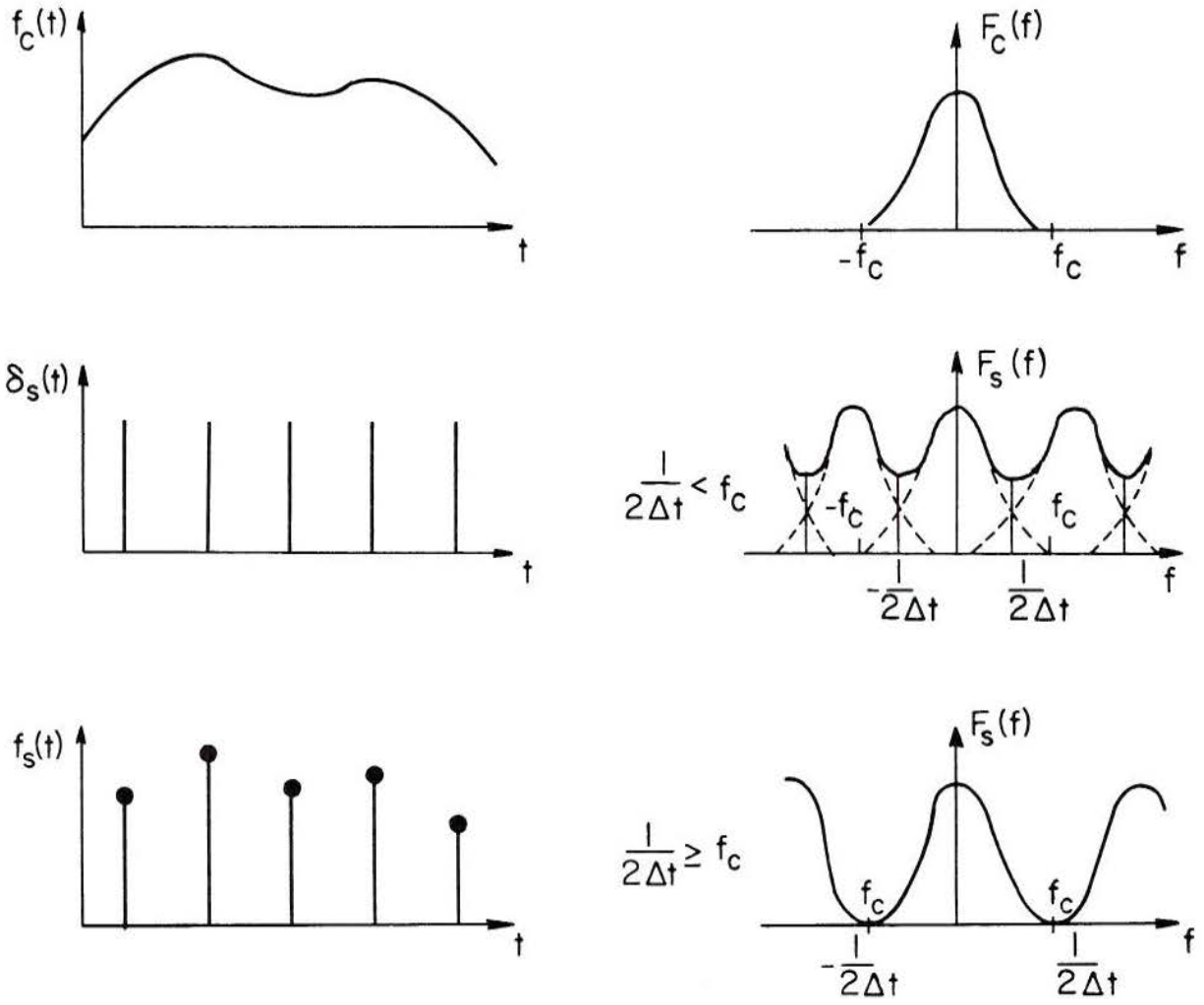

Fig. 4.2 Effect of pond sampling on the spectrum of the process.

cases discrete sampling is actually superior to continuous sampling.

4.2 Average Sampling

While the previous section was concerned with discrete observations of instantaneous values of a process, the focus is now on another common technique of sampling, namely time-average or integrated sampling. Here the sampling is performed over a certain time interval Δt and the result is the integral of the process over the interval, ordinarily presented as an average intensity in the time interval Δt, Fig. 4.3.

Averaging sampling devices are frequently used in hydrology, for example, for the measurement of precipitation and evaporation. The bucket rain gauge that is emptied at fixed times every day is probably one of the most used rain gauging techniques in the world, and the evaporation pan is by far the most used evaporation gauge.

Similarly, samples that were obtained continuously are often integrated in the data processing procedure and are presented as hourly, daily, monthly, annual total or average values. If these data are used uncritically as a representation for the continuous process, some important properties may be changed.

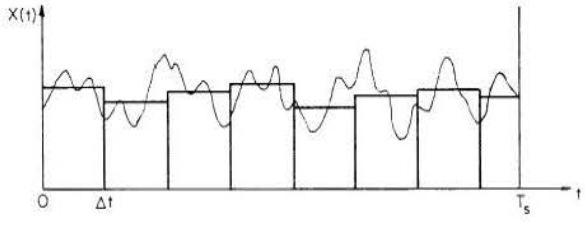

Fig. 4.3 Average sampling scheme.

It is obvious from Fig. 4.3 that average sampling may seriously degrade the characteristics of the original continuous sample. The variability in the sample is attenuated by the smoothing effect of averaging, so that the magnitude of extreme events will be reduced. Actually, the average sampling, with a sampling interval Δt, corresponds to discrete point sampling with interval Δt of a process formed by smoothing the original process. So, in addition to the information loss due to discrete point sampling as demonstrated above, one may introduce an additional uncertainty due to the distortion of the basic process.

This attenuation is easily demonstrated by comparing the original continuous process X(t) and a derived continuous process Y(t) formed by a moving average process of X(t) over the interval Δt.

$$Y(t) = \frac{1}{\Delta t} \int_{t-\Delta t/2}^{t+\Delta t/2} X(t)\, dt \qquad (4.1)$$

If Y(t) is sampled at discrete intervals Δt, a discrete series A_i is obtained, and it is easily seen from Fig. 4.4 that A_i is equivalent to an average sampled realization of X(t) over Δt. The loss of information in going from Y(t) to A_i can be treated as a discrete sampling of the attenuated process, Y(t). The additional loss due to the attenuation by transforming X(t) to Y(t) can be analyzed by evaluating the bias of the properties of Y(t) in relation to the true properties of X(t). Since bias is an additive property, there is no difficulty in combining the effects of the two steps.

It is difficult to tell in advance which of the two methods, point or average sampling, is associated with the least loss of information for a given sampling interval. For a property such as the mean, which is not distorted by the average procedure, the average sampling is obviously the best as it actually incorporates the total information in the continuous trace. But for more frequency sensitive properties such as crossings of a given level, the attenuation due to averaging may be so critical that point sampling should be preferred.

It is clear that when the sampling interval approaches the critical frequency f_c, the two approaches are almost equivalent, so that, for example, the daily instantaneous sampling might be approximated with daily average values if the attenuation due to averaging is negligible. This will, however, seldom be the case with monthly or annual values, and often even daily values will be a gross approximation.

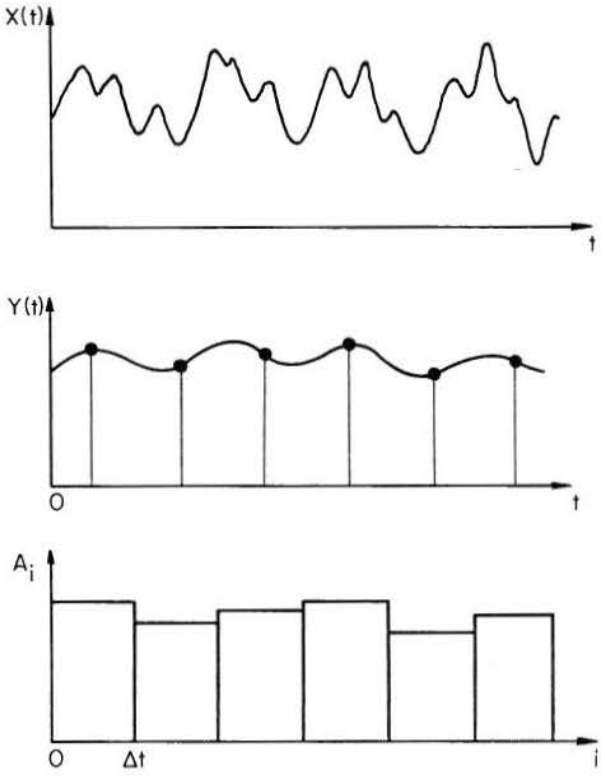

Fig. 4.4 Illustration of the effect of integrated sampling in the form of an attenuation of the original continuous process.

4.3 Quantization

The two previous subsections considered discretization along the time axis of a continuous process. A similar discretization can be performed on the random variable itself, or along the axis of the random values. If all values in a given interval, Δx of the X-axis are considered equal to one value, say the midpoint of the interval, and the whole X-axis is partitioned into such intervals, the procedure represents the quantizing of a random variable, Fig. 4.5.

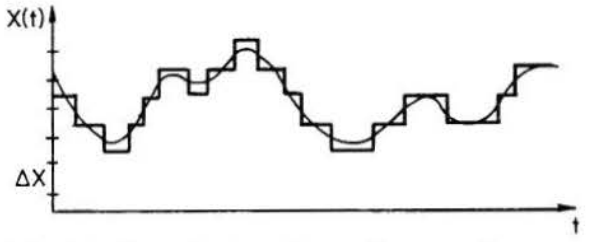

Fig. 4.5 Quantization of a random variable.

The most common reason for quantization is the presence of errors in observations, instruments or errors introduced by the data processing. The limited accuracy automatically approximates the continuous variable by a discrete variable defined down to the last reliable digit, and the size of the measurement error defines the interval Δx. Usually this error is small compared to the total range of variation of the random variable, and the error introduced by such quantization may not be significant.

Another common use of quantization is applied in computation of sample frequencies or histograms. The random variable is grouped into class intervals, and relative frequencies are associated with the central value of the class interval.

In recent years, the concepts of systems engineering have been applied extensively in the analysis of water resources problems. A fundamental property of many of these techniques is the grouping of the variables values into discrete states. Due to limitations in computer storage, it is often necessary to limit the number of states an actual continuous variable may assume, and this may give rise to a rather coarse quantization (Hall and Dracup, 1970). Similarly, the introduction of discrete Markov chains (Moran, 1959) to model stochastic processes such as reservoir levels introduces quantization.

In this study, the effect of quantization on the estimation of stochastic properties is emphasized. It has been shown (Widrow, 1956) that the quantization can be considered as a nonlinear operator having the input-output relationship shown in Fig. 4.6.

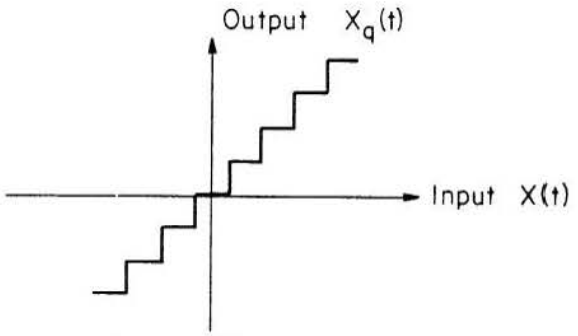

Fig. 4.6 Effect of quantization.

Based on this, it is shown that introduction of quantization corresponds mathematically to addition of a quantization noise n_q to the original signal, i.e.

$$X_q = X + n_q \tag{4.2}$$

where X_q is the quantized random variable. n_q has been found to be uniformly distributed over the interval

$$X_q - \Delta x/2 < n_q \leq X_q + \Delta x/2 \tag{4.3}$$

and it introduces a bias in the estimation of the variance. It will be shown, however, that this bias is small for even very coarse quantization, and that the dominating loss of information is introduced through the discretization of the time axis.

CHAPTER 5

LOSS OF INFORMATION BY DISCRETE POINT SAMPLING

In this chapter, expressions are developed for determination of the expected information loss as a function of the sampling interval, Δt. It is assumed throughout the chapter that the sampling is performed by discrete periodic observations of instantaneous values.

In subchapter 3.3 the expected uncertainty loss was found to be a function of bias and variance of the estimate. These two properties have been evaluated for a set of selected estimates as functions of Δt and of the mean μ, variance, σ^2, and autocovariance, $\gamma(u)$, of the underlying continuous process. This provides the most general framework for evaluating the loss of information in a given process. No attempts have been made to express the loss for given dependence models, because it is believed that so far, no generally accepted model exists for continuous hydrologic time series. If future research should develop theoretically based autocovariance functions for hydrologic processes, these might be substituted in the equations presented to express the information losses in closed form. At present this is not possible, so instead it is suggested that empirical covariance functions be estimated from continuous samples of the processes, and, based on this, the information losses can be computed numerically from the equations presented.

Even though the topic of this study is a comparison between continuous and discrete sampling of stochastic processes, it is sufficient to consider discrete sampling only. The reason for this is the approximate frequency-limited behavior of the spectrum of most, if not all, hydrologic time series. According to the sampling theorem discussed in subchapter 4.1, a sufficiently dense discrete sampling of a frequency-limited signal contains all the information of the continuous signal. This is in the sense that the sample spectrum is undistorted and that the original signal may be completely recovered from the discrete sample by a suitable filtering transformation. Therefore, it has been decided to exploit this property in this study and to consider a continuous trace and a discrete trace sampled at twice the critical frequency to be equivalent, and to measure the information loss due to coarser discretization relative to this latter case. This gives the convenience of limiting the analysis strictly to discrete processes. It should be emphasized, however, that this simplification is made mainly because mathematically the main difference between the discrete and continuous processes is the change of summation signs to integral signs, and that by considering discrete processes only, the use of the integral versions of results becomes unnecessary.

5.1 Loss of Information in Estimating Distribution Functions

The distribution function $F_X(x)$ of a random variable is estimated by the relative frequency distribution curve determined from the sample. Siddiqui (1962) found the bias and variance for such an estimate by considering the process

$$Y_i = 1 \quad \text{if} \quad X_i \leq x \tag{5.1}$$

$$Y_i = 0 \quad \text{otherwise}$$

If a time series, $X(t)$ of length T_s is sampled periodically with sampling interval Δt, the number of sampled values is $T_s/\Delta t$. The estimate may now be written as

$$\hat{F}_X(x) = \frac{\Delta t}{T_s} \sum_{i=1}^{T_s/\Delta t} Y_i \tag{5.2}$$

so $\hat{F}_x(x)$ is actually a sample mean of the random variable Y_i. The bias of the sample mean is always zero, so the bias of $\hat{F}_x(x)$ is zero for any sampling interval Δt. The variance of a sample mean is given in Eq. 5.12, which for the Y_i process becomes

$$V[\hat{F}_X(x)] = \frac{\Delta t}{T_s}\left[\sigma_Y^2 + 2 \sum_{u=1}^{T_s/\Delta t - 1} (1 - \frac{u\Delta t}{T_s})\, \gamma_Y(u)\right] \tag{5.3}$$

Here σ_Y^2 and $\gamma_Y(u)$ are respectively the variance and autocovariance functions of the random variable, Y with

$$\sigma_Y^2 = F_X(x)[1 - F_X(x)] \tag{5.4}$$

and

$$\gamma_Y(u) = P[X_i \leq x, X_{i+u\Delta t} \leq x] - F_X^2(x) \tag{5.5}$$

for $\gamma(u\Delta t) \neq 0$. If $\gamma(u\Delta t)$ equals zero, $\gamma_Y(u)$ vanishes. The value of $P[X_i \leq x, X_{i+u\Delta t} \leq x]$ may be found for a normal process from tables of the

bivariate normal distribution, since the mean μ, the variance σ^2, and the autocovariance $\gamma(u)$ of the X(t) process are assumed known. For non-normal cases either a transformation to normal variables, or the bivariate Gamma distribution may be used as the population distribution function for the computation of $\gamma_Y(u)$. In this latter case, the shape parameter α and scale parameter β are computed by

$$\alpha = \frac{\mu^2}{\sigma^2} \tag{5.6}$$

and

$$\beta = \frac{\sigma^2}{\mu} \tag{5.7}$$

With the bias $B[\hat{F}_x(x)]$ equal to zero and variance $V[\hat{F}_x(x)]$ determined in Eq. 5.3 as a function of Δt, the expected information loss of Eq. 3.16 simplifies to

$$\bar{I}(\hat{F}_X(x),\Delta t) = .3989\, k \sqrt{V[\hat{F}_x(x)]} \tag{5.8}$$

Based on Eqs. 5.3 and 5.8 the information loss may be evaluated for any given covariance function and discretization interval.

5.2 Loss of Information in Estimating the Mean

The mean μ of a stochastic process is one of its most important characteristics. It is estimated by the sample mean $\hat{\mu}$ of N observations of the process.

The bias of the sample mean is zero, and the variance of the estimate is expressed by

$$V[\hat{\mu}] = E[(\hat{\mu}-\mu)^2] = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} E[(X_i-\mu)(X_j-\mu)]$$

or

$$V[\hat{\mu}] = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} C[X_i,X_j] \tag{5.9}$$

where $C[X_i,X_j]$ is the covariance between the i'th and the j'th observations. If $\gamma(u)$ is the autocovariance of the continuous process

$$C[X_i,X_j] = \gamma((i-j)\Delta t) \quad , \tag{5.10}$$

and using the symmetry in $\gamma(u)$, Eq. 5.9 can be written as

$$V[\hat{\mu}] = \frac{1}{N} [\sigma^2 + 2 \sum_{u=1}^{N-1} (1 - \frac{u}{N}) \gamma(u\Delta t)] \quad , \tag{5.11}$$

or, by introduction of $N = T_s/\Delta t$

$$V[\hat{\mu}] = \Delta t [\frac{\sigma^2}{T_s} + \frac{2}{T_s} \sum_{u=1}^{T_s/\Delta t-1} (1 - \frac{u\Delta t}{T_s}) \gamma(u\Delta t)] \tag{5.12}$$

The variance of $\hat{\mu}$ can also be expressed in terms of the spectral density function $\Gamma(f)$ (Siddiqui 1962). If $\Gamma(f)$ is known, the following asymptotic expression is useful for large T_s and small Δt:

$$V[\hat{\mu}] = \Delta t \frac{\Gamma(0)}{T_s} \tag{5.13}$$

For any given sampling interval Δt and autocovariance function $\gamma(u)$, the expected information loss can now be determined by

$$\bar{I}(\hat{\mu},\Delta t) \tag{5.14}$$
$$= .3989\, k \sqrt{\frac{\Delta t}{T_s} [\sigma^2 + 2 \sum_{u=1}^{T_s/\Delta t-1} (1 - \frac{u\Delta t}{T_s}) \gamma(u\Delta t)]}$$

The information loss includes the factor $\sqrt{\Delta t}$; however the number of terms in the summation is diminished with increased Δt, which decreases the magnitude of the factor in parenthesis. It is actually possible to construct autocovariance functions, where the latter factor dominates over the increase in Δt, so the information loss decreases with increased Δt. But it is easily shown that, for the monotonically decreasing, convex autocovariance functions that are most common for hydrologic processes, information losses of the mean estimate will always increase with increase of Δt.

If the series is a first-order Markov model, the variance can be found in closed form (Brooks and Carruthers, 1953). For a second-order Markov model, the variance has been found by Reiher and Huzzen (1967) and Quimpo (1969), but the analytical expression is rather complex, and the results are given in tabular form. In general, the first and second-order Markov models are often insufficient for modeling of closely sampled hydrologic series. In such cases, a direct sample estimate of the autocovariance function $\gamma(u)$ of the process under investigation should be used in Eq. 5.14 and the change in information with changes in Δt determined numerically.

5.3 Loss of Information in Estimating Variance and Autocovariance

The variance σ^2 and autocovariance function $\gamma(u)$ define the second-order moments of the

process. Since the variance is equal to $\gamma(0)$, expressions for bias and variance of the estimates $\hat{\gamma}(u)$ of $\gamma(u)$ are developed as functions of Δt. From these, the properties of the variance estimate $\hat{\sigma}^2$ may be found as a special case.

The autocovariance may be estimated based on N equally spaced observation of X as

$$\hat{\gamma}(u) = \frac{1}{N-u} \sum_{i=1}^{N-u} (X_i - \hat{\mu})(X_{i+u} - \hat{\mu}) \qquad (5.15)$$

To find the expectation of $\hat{\gamma}(u)$ Eq. 5.15 can be rewritten as

$$\hat{\gamma}(u) \cong \frac{1}{N-u} \sum_{i=1}^{N-u} (X_i - \mu)(X_{i+u} - \mu) - (\hat{\mu} - \mu)^2 \qquad (5.16)$$

The mean μ may be assumed equal to zero without loss of generality. The expectation of the first term becomes, by the defintion of $\gamma(u)$,

$$E\left[\frac{1}{N-u} \sum_{i=1}^{N-u} (X_i - \mu)(X_{i+u} - \mu)\right] = \gamma(u)$$

According to Eq. 5.12 the expectation of the second term becomes

$$E[(\hat{\mu} - \mu)^2] = V[\hat{\mu}] = \frac{1}{N}\left[\sigma^2 + 2 \sum_{u=1}^{N-1} (1 - \frac{u}{N})\, \gamma(u\Delta t)\right]$$

So by introducing $T_s/\Delta t = N$ the expected value of $\hat{\gamma}(u)$ is

$$E[\hat{\gamma}(u)] = \gamma(u) - \frac{\Delta t}{T_s}\left[\sigma^2 + 2 \sum_{u=1}^{T_s/\Delta t-1} (1 - \frac{u\Delta t}{T_s})\, \gamma(u\Delta t)\right], \qquad (5.17)$$

and the bias for a given Δt is

$$B[\hat{\gamma}(u)] = -\frac{\Delta t}{T_s}\left[\sigma^2 + 2 \sum_{u=1}^{T_s/\Delta t-1} (1 - \frac{u\Delta t}{T_s})\, \gamma(u\Delta t)\right]. \qquad (5.18)$$

An approximate expression for the variance of $\hat{\gamma}(u)$ has been determined by Bartlett (1946) as

$$V[\hat{\gamma}(u)] \cong \frac{1}{N} \sum_{i=-N+u}^{N-u} [C^2[X_j, X_{j+i}] \qquad (5.19)$$

$$+ C[X_j, X_{j+i-u}]\, C[X_j, X_{j+i+u}]] .$$

For a sampling interval Δt, the autocovariances are

$$C[X_j, X_{j+i}] = \gamma(i\Delta t),$$

$$C[X_j, X_{j+i-u}] = \gamma(i\Delta t - u), \text{ and}$$

$$C[X_j, X_{j+i-u}] = \gamma(i\Delta t + u),$$

so that

$$V[\hat{\gamma}(u)] \simeq \frac{\Delta t}{T_s} \sum_{i=-T_s/\Delta t+u}^{T_s/\Delta t-u} [\gamma(i\Delta t)^2 + \gamma(i\Delta t - u)\, \gamma(i\Delta t + u)]. \qquad (5.20)$$

or by using the symmetry in $\gamma(u)$:

$$V[\hat{\gamma}(u)] \simeq \frac{\Delta t}{T_s}\left[\sigma^4 + \gamma(u)^2 + 2 \sum_{i=1}^{T_s/\Delta t-u} [\gamma(i\Delta t)^2 + \gamma(i\Delta t - u)\, \gamma(i\Delta t + u)]\right] \qquad (5.21)$$

Equations 5.18 and 5.21 apply to small samples and can, therefore, be used without discrimination. For large samples, simplified approximations such as those presented by Siddiqui (1962), can be obtained. The reason for using the more complex, small sample formulas is that the sample size decreases with an increase in the sampling interval Δt, and, therefore, it is important not to use asymptotic expressions for the bias and the variance. Bias and variance of $\hat{\gamma}(u)$ may now be determined as functions of Δt for any given autocovariance function and by Eq. 3.16, the expected information loss $\overline{I}(\hat{\gamma}(u), \Delta t)$ may be found.

It should be noted that $\gamma(u)$ can be estimated only if Δt is smaller than u and if u is a multiple of Δt. If these conditions are not fulfilled, no information about $\gamma(u)$ is contained in the sample.

For the variance estimate $\hat{\sigma}^2$, the bias is found from Eq. 5.18:

$$B[\hat{\sigma}^2] = \frac{\Delta t}{T_s}\left[\sigma^2 + 2 \sum_{i=1}^{T_s/\Delta t-1} (1 - \frac{i\Delta t}{T_s})\, \gamma(i\Delta t)\right] \qquad (5.22)$$

and the variance $\hat{\sigma}^2$ is

$$V[\hat{\sigma}^2] = 2\frac{\Delta t}{T_s}\left[\sigma^4 + 2 \sum_{i=1}^{T_s/\Delta t-1} \gamma^2(i\Delta t)\right] \qquad (5.23)$$

For a given autocovariance function $\overline{I}(\hat{\sigma}^2, \Delta t)$ is found by substitution of Eqs. 5.22 and 5.23 into Eq. 3.16.

As was the case for the mean, the bias and variance of the estimators above tend to increase with an increased Δt through the factor Δt, whereas the reduction of terms in summation of the autocovariances has the opposite effect. If the latter is dominating, the information content may increase with an increase of Δt, so that very close sampling is actually detrimental; however, if the autocovariance

function is convex, as is mostly the case in hydrology, the information content in the variance and autocovariance estimate will always increase with a decrease of Δt.

5.4 Analysis of the First Derivative of the Process

The first derivative of a stochastic process is a measure of the rate of change per unit time. It is intuitively evident that a process with rapid changes should be sampled more closely than a more slowly changing process in order not to lose important information. The derivative is estimated by the difference between adjacent sample values; too open sampling may not show all the variability in the underlying continuous realization and therefore tend to make the process appear more smooth than it really is. Being a function of the stochastic process X(t), the derivative X'(t) is itself a stochastic process that is characterized by its mean, variance and autocovariance. A realization of the derivative process illustrates the changes in the original process as large values of X'(t) are associated with rapid changes, whereas small values are associated with slow changes, as shown in Figs. 5.1 and 5.2.

Cramer and Leadbetter (1967) presented a general analysis of the derivative X'(t) of a stochastic process by showing that X'(t) has the expected value

$$E[X'(t)] = 0 \tag{5.24}$$

and the autocovariance

$$\gamma_{X'}(u) = -\frac{\partial^2 \gamma(u)}{\partial u^2} = -\gamma''(u) \tag{5.25}$$

i.e. the autocovariance of the derivative process is the negative value of the second derivative of the autocovariance of the original process. In particular, the variance of the derivative becomes

$$V[X'(t)] = -\gamma''(0) \quad . \tag{5.26}$$

a parameter that is of great significance for some extreme and crossing properties. Intuitively, it is not difficult to understand the implications of changes in $-\gamma''(0)$. A large value implies that the derivatives cover a wide range, and that many rapid changes are present in the original process. This makes the process appear very irregular, whereas a process with small $-\gamma''(0)$ has a more smooth and wavy appearance. Hence $-\gamma''(0)$ can be considered as a measure of the irregularity of a process. It should be noted

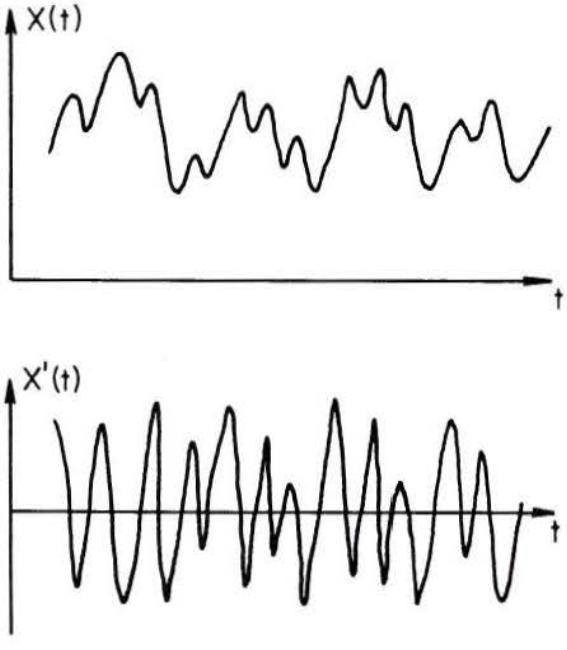

Fig. 5.1 Irregular process with rapid changes.

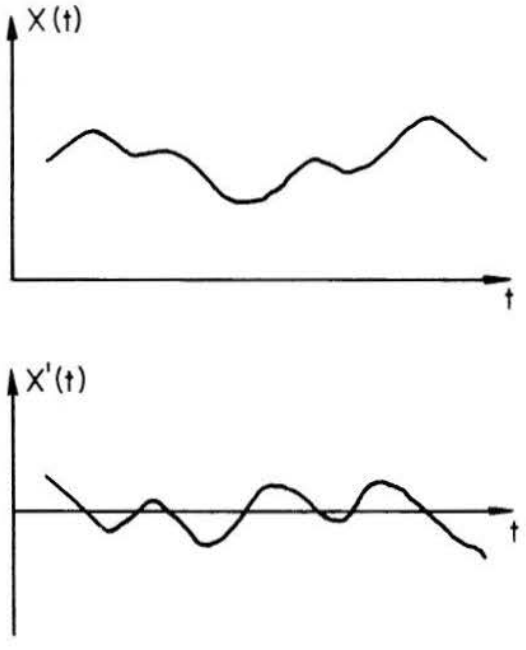

Fig. 5.2 Smooth process with slow changes.

that this measure is nonstructural without making any other assumption about the process except that it is stationary and differentiable.

Another measure of irregularity, or of the high-frequency behavior of the process, is found in the spectrum. As might be expected, there is a close relation between $-\gamma''(0)$ and the spectrum $\Gamma(f)$.

Be definition, if f denotes the angular frequency,

$$\gamma(u) = \int_{-\infty}^{\infty} \Gamma(f) \cos(uf)df \qquad (5.27)$$

so

$$\gamma''(u) = - \int_{-\infty}^{\infty} f^2 \Gamma(f) \cos(fu)df \qquad (5.28)$$

and

$$\gamma''(0) = - \int_{-\infty}^{\infty} f^2 \Gamma(f)\, df\,, \qquad (5.29)$$

i.e., $\gamma''(0)$ is the negative value of the second moment of the spectrum with respect to the origin of f. However, the most important implication of $\gamma''(0)$ is that it is an important parameter affecting certain extreme and crossing properties of a continuous stochastic process, and loss of information in estimation of $\gamma''(0)$ will therefore be carried over into the estimation of these properties.

$\gamma''(0)$ may be estimated by computing the second moment of the sample spectrum, taken about the origin. However, estimation of the spectrum is as known connected with some difficulties, in particular for short samples often encountered in hydrologic applications. Hence, it is more advantageous to base the estimations on the sample autocovariance.

Because of discrete sampling, the sample autocovariance is a discrete function and cannot give $\gamma''(0)$ directly by differentation. Rodriguez (1968) proposed a finite difference approximation to $\gamma''(0)$ by

$$\hat{\gamma}''(0) = \frac{2}{\Delta t^2} [\hat{\gamma}(\Delta t) - \hat{\gamma}(0)] \qquad (5.30)$$

This estimate is equivalent to a direct estimation of the variance of the differences between adjacent values of the discrete sampled process, and is as such an efficient estimate of that particular parameter. However, the actual parameter of interest is the variance of the derivative of the continuous process, and an estimate of this parameter has been proposed by Ditlevsen (1971).

For a discrete sample, the spectrum is defined only within the limits $0 \leqslant f \leqslant 1/2\Delta t$. However, if the sampling rate is equal to or higher than the maximum frequency f_c in the process, i.e., $\Delta t_* \leqslant 1/2f_c$, the spectrum of the discretely sampled process is identical to the spectrum of the continuous process. Without any loss of accuracy, then,

$$\gamma''(0) = - 4\pi^2 \int_{-\infty}^{\infty} f^2 \Gamma(f)df = - 8\pi^2 \int_{0}^{1/2\Delta t_*} f^2 \Gamma(f)df, \qquad (5.31)$$

where f is the ordinary frequency.

By definition

$$\Gamma(f) = 2\Delta t_* \sum_{u=-\infty}^{\infty} \gamma(u\Delta t_*) \cos 2\pi f u \Delta t_*\,,$$

$$0 \leqslant f \leqslant \frac{1}{2\Delta t_*}\,. \qquad (5.32)$$

Introducing Eq. 5.32 into Eq. 5.31

$$\gamma''(0) = - \frac{1}{\Delta t_*^2} [\frac{\pi^2}{3} \sigma^2 + 4 \sum_{u=1}^{\infty} \frac{(-1)^u}{u^2} \gamma(u\Delta t_*)]\,. \qquad (5.33)$$

Eq. 5.33 gives the exact value $\gamma''(0)$ of a continuous autocovariance $\gamma(u)$ expressed by a discrete representation $\gamma(u\Delta t_*)$, and it shows that continuous sampling and discrete sampling give identical results as long as $\Delta t_* \leqslant 1/2f_c$. For larger sampling intervals, a bias will be introduced into $\gamma''(0)$. For a given Δt, the estimate of $\gamma''(0)$ becomes

$$\hat{\gamma}''(0) = - \frac{1}{\Delta t^2} [\frac{\pi^2}{3} \hat{\sigma}^2 + 4 \sum_{u=1}^{N} \frac{(-1)^u}{u^2} \hat{\gamma}(u\Delta t)] \qquad (5.34)$$

The advantage of this estimate is, that for sampling frequencies higher than the critical frequency it gives results equivalent to those obtained by estimating $\gamma''(0)$ based on a continuous sample. If the estimates of the autocovariances can be assumed unbiased, such as may be the case for large samples, the estimate of Eq. 5.34 is unbiased too. It can be seen that the coefficients reduce the influence of the large-lag autocovariance estimates, so that an accurate estimate of these are of less importance.

If the sampling frequency is below the critical frequency, the estimate of $\gamma''(0)$ becomes biased. It is possible to express the bias directly as a function of autocovariances in the analysis of the sampling properties of $\hat{\gamma}''(0)$. However, the resulting equation becomes so complex that it is practical only to present the expectation of $\hat{\gamma}''(0)$ and determine the bias based on a direct computation of $\gamma''(0)$ for the approximation to the underlying process sampled with the sampling interval $\Delta t_* = 1/2f_c$. The expectation of Eq. 5.34 is

$$E[\hat{\gamma}''(0)] = -\frac{1}{\Delta t^2}\left\{\frac{\pi^2}{3} E[\hat{\sigma}^2] + 4 \sum_{u=1}^{T_s/\Delta t} \frac{(-1)^u}{u^2} E[\hat{\gamma}(u\Delta t)]\right\} \quad (5.35)$$

where $E[\hat{\gamma}(u)]$ can be found from Eq. 5.17 and the bias is

$$B[\hat{\gamma}''(0)] = E[\hat{\gamma}''(0)] + 4f_c^2 \left[\frac{\pi^2}{3}\sigma^2 + 4 \sum_{u=1}^{2T_s f_c} \frac{(-1)^u}{u^2} \gamma\left(\frac{u}{2f_c}\right)\right] . \quad (5.36)$$

The variance is found by considering the variance of the sum

$$V[\hat{\gamma}''(0)] = \frac{1}{\Delta t^4} V\left[\sum_{u=0}^{T_s/\Delta t} a_u \hat{\gamma}(u\Delta t)\right] \quad (5.37)$$

where

$$a_o = \frac{\pi^2}{3} \text{ , for } u = 0$$

and

$$a_u = 4 \frac{(-1)^u}{u^2} \text{ , for } u \neq 0 ,$$

so that

$$V[\hat{\gamma}''(0)] = \frac{1}{\Delta t^4} \sum_{u=0}^{T_s/\Delta t} \sum_{v=0}^{T_s/\Delta t} a_u a_v C[\hat{\gamma}(u\Delta t) . \hat{\gamma}(v\Delta t)] . \quad (5.38)$$

According to Bartlett (1946), the autocovariances are

$$C[\hat{\gamma}(u\Delta t), \hat{\gamma}(v\Delta t)] \cong \frac{\Delta t}{T_s} \sum_{i=-\infty}^{\infty} [\gamma(i\Delta t)\gamma((i+u-v)\Delta t) + \gamma((i+u)\Delta t)\gamma((i-v)\Delta t)] . \quad (5.39)$$

With the autocovariance $\gamma(u)$ of the continuous process given, $B[\gamma''(0)]$ and $V[\gamma''(0)]$ may be determined for different values of Δt, and by substituting them into Eq. 3.16, the expected information loss $\bar{I}[\gamma''(0),\Delta t]$ can be found. In Chapter 8 it is shown that, for large sampling intervals, the bias of the estimate of $\gamma''(0)$ can become the dominating contributor to the information loss. In this case, inclusion of the bias is essential for a proper measure of information content, and concepts such as Shannon's would not give a true picture of the changes of information with increase of the sampling interval.

5.5 Loss of Information in Estimating Probabilities of Extremes

The estimation of the probability that a hydrologic variable such as a flood discharge exceeds a certain value is a very important practical problem. Evidently, loss of information in estimating the probability of an extreme can occur if the sampling interval is so wide that extreme events occur between two adjacent observations. Hence, for a very irregular process with rapid changes, close sampling is essential, whereas the extremes of more smooth processes can be detected by a relatively wide sampling interval.

Several different approaches have been applied in hydrologic studies of probabilities of extremes. Beginning with the work of Hazen (1914), a much used technique applies the fitting to empirical frequency curves of standard probability functions such as the normal, lognormal or Pearson Type III distributions. Another much used approach, first presented by Fisher and Tippett (1928) and introduced into hydrologic studies by Gumbel (1945), uses the double exponential distribution

$$F_X(x) = e^{-e^{-x}} \quad (5.40)$$

to model the distribution of the extremes in a time interval $0 < t < T$, under the assumption that T is large.

More recently, Zelenhasic (1970) has demonstrated still another approach by considering the stream of flood events as an intermittent random process. This approach is based on the contributions by Todorovic (1970) on the problem of probabilities of a random number of random variables.

Common to the approaches above is that they consider the extremes as new random variables derived from the underlying continuous process, and this complicates an analysis of the influence of the choice of sampling interval. Therefore, the present

investigation is based on an approach that relates the properties of the underlying continuous process directly to the probability of extremes, as presented by Ditlevsen (1971). It has been used in flood analysis by Mejia (1971).

In an analysis of the extremes of a continuous normal process Cramer and Leadbetter (1967) shows that there is an important relationship between the second spectral moment λ_2 and the probability of extremes. They develop the asymptotic relation for a standardized normal process in a time interval of length T as

$$P\left[\max_{0<t<T} X(t) \leq (2\ell n\ T)^{1/2} + \frac{A+z}{(2\ell n\ T)^{1/2}}\right] \cong e^{-e^{-z}} \tag{5.41}$$

for large T. Here

$$A = \ell n\ \frac{\sqrt{\lambda_2}}{2\pi}\ , \tag{5.42}$$

and

$$\lambda_2 = \int_{-\infty}^{\infty} f^2\ \Gamma\ (f)\ df = -\gamma''(0) \tag{5.43}$$

Hence, the parameters of the classical double exponential extreme distribution of Eq. 5.40 are related to $\gamma''(0)$. This parameter was studied in subchapter 5.4, where its dependence on the high frequency terms in the spectrum was demonstrated.

Based on Eq. 5.41 and the results of subchapter 5.4, the effect of discrete sampling on the double exponential distribution may be investigated. However, the asymptotic assumption about large T is unfortunate in hydrologic studies, where the seasonal variation over the year limits the time period during which the process can be considered stationary. Ditlevsen (1971) studied the probabilities of extremes of a continuous process in the interval $0 < t < T$ without the assumption of a large T. He developed the following general expression for the probability that the maximum value of the process X(t) in the time interval $0 < t < T$ does not exceed a given level u:

$$P\left[\max_{0<t<T} X(t) \leq u\right] = P[X(0) \leq u] \exp\left[-\frac{E[U_u(0,1)]T}{P[X(0) \leq u]}\right] \exp\ [G(T)] \tag{5.44}$$

where $E[U_u(0,1)]$ is the expected number of upcrossings of level u in a unit time period. This value was found by Rice (1945) for a normal process with mean μ and variance σ^2

$$E[U_u\ (0,1)] = \sqrt{\frac{\lambda_2}{2\pi}}\ \phi\ (u_s)\ . \tag{5.45}$$

where again λ_2 is the second moment of the spectrum and $\phi(u_s)$ is the standardized normal density function evaluated in $u_s = (u-\mu)/x$.

Furthermore,

$$P[X(0) \leq u] = \Phi\ (u_s) \tag{5.46}$$

where $\Phi(u_s)$ is the standardized normal distribution function. The only unknown now is the factor, exp G(T). Ditlevsen realized the practical difficulties in finding G(T), for even simple processes. He therefore assumed that in practical applications G(T) may be equal to zero and supported this assumption with extensive checking by means of data generation. Furthermore, he showed that even with this assumption Eq. 5.44 converges to the double exponential distribution of Eq. 5.40 when $T \to \infty$. Hence, for a normal process X(t) with mean μ, variance σ^2 and second spectral moment λ_2:

$$P\left[\max_{0<t<T} X(t) \leq u\right] = \Phi(u_s) \exp\left[-\frac{T\,\phi\,(u_s)}{\sqrt{2\pi}\,\Phi\,(u_s)}\ \sqrt{\lambda_2}\right] \tag{5.47}$$

Again the importance of the high frequency parameter λ_2 is evident.

An estimate of the probability $P[\text{Max } X(t) \leq u]$ for a given sampling interval can be found by substituting the estimate $\hat{\lambda}_2$ for λ_2. To find the exact sampling mean and variance of this estimate as a function of Δt is extremely difficult, due to the complexity of Eq. 5.47. Instead, this equation is expanded in a Taylor series around the mean μ_* of λ_2, and the mean and variance of this approximation is found.

The second order expansion is of the form

$$f(x) = f(\mu_*) + f'(\mu_*)\ (x - \mu_*\ ) + \frac{1}{2}\ f''\ (\mu_*)\ (x - \mu_*)^2\ , \tag{5.48}$$

where x is substituted for λ_2.

The expected value of f(x) is found by taking the expectation of Eq. 5.48. $E[x-\mu_*] = 0$, and when x is normally distributed with the variance σ_*^2, with the skewness $E[X-\mu_*)^3$ zero, the approximation for the mean is

$$E[f(x)] \cong f(\mu_*) + \frac{1}{2} f''(\mu_*)\, \sigma_*^2 \quad . \qquad (5.49)$$

The approximate second moment of f(x) becomes

$$E[f(x)^2] \cong E[f(\mu_*)^2 + 2f(\mu_*)f'(\mu_*)(x - \mu_*)$$

$$+ (f(\mu_*)f''(\mu_*) + f'(\mu_*)^2)(x-\mu_*)^2$$

$$+ f'(\mu_*)f''(\mu_*)(x - \mu_*)^3 + \frac{1}{4} f''(\mu_*)^2 (x-\mu_*)^4]$$

$$\cong f(\mu_*)^2 + (f(\mu_*)f''(\mu_*) + f'(\mu_*)^2)\,.$$

$$\sigma_*^2 + \frac{3}{4} f''(\mu_*)^2\, \sigma_*^4 \quad , \qquad (5.50)$$

because for X normal, $E[(x-\mu_*)]^4 = 3\sigma_*^4$.

The variance is found from

$$V[f(x)] = E[f(x)^2] - E^2[f(x)]$$

$$= f'(\mu_*)^2 \sigma_*^2 + \frac{1}{2} f''(\mu_*)^2\, \sigma_*^4 \,. \qquad (5.51)$$

and the expression for the probability of extremes, Eq. 5.47 then becomes

$$f(x) = \Phi(u_s) \exp\left[- \frac{T\ \phi(u_s)}{\sqrt{2\pi}\ \Phi(u_s)} \sqrt{x} \right] \quad , \qquad (5.52)$$

$$f'(x) = - \frac{T\phi(u_s)}{2\sqrt{2\pi}\sqrt{x}} \exp\left[- \frac{T\phi(u_s)}{\Phi(u_s)} \sqrt{x} \right] \quad . \qquad (5.53)$$

and

$$f''(x) = \frac{T\ \phi(u_s)}{4\sqrt{2\pi}\ x} \left[\frac{1}{\sqrt{x}} + \frac{T\ \phi(u_s)}{\sqrt{2\pi}\ \Phi(u_s)} \right]$$

$$\exp\left[- \frac{T\ \phi(u_s)}{\sqrt{2\pi}\ \Phi(u_s)} \sqrt{x} \right] \qquad (5.54)$$

By substituting $\mu_* = E[\lambda_2]$ for x in Eqs. 5.52 to 5.54 and $\sigma_*^2 = V[\lambda_2]$ in Eqs. 5.49 and 5.51, the expectation and variance of the estimate is given in terms of the mean and variance of λ_2. In subchapter 5.4 the mean and variance of the estimate $\hat{\gamma}''(0)$ of $\gamma''(0)$ are found as functions of the sampling interval Δt. As $\lambda_2 = -\gamma''(0)$, then

$$\mu_* = E[\hat{\lambda}_2] = - E[\hat{\gamma}''(0)] \qquad (5.55)$$

and

$$\sigma_*^2 = V[\hat{\lambda}_2] = V[\hat{\gamma}''(0)] \,. \qquad (5.56)$$

The above analysis permits computation of the expectation and variance of the estimate of the probability of extremes of a continuous normal process for any given autocovariance $\gamma(u)$ and sampling interval Δt. The bias of the estimate is found by subtraction of the population value, and the expected uncertainty loss $\overline{I}[\hat{P},\Delta t]$ is found by substitution of bias and variance into Eq. 3.16.

The significant bias that is characteristic for estimates of $\gamma''(0)$ when the sampling interval is large will influence the estimation of the probability of extremes accordingly. This bias is a measure of the amount of extreme events that are not detected by the discrete sampling; it should be accounted for in the information concept, just as has been done by the expected information loss.

5.6 Loss of Information in Estimating Run Properties

Although it seems to be of similar importance as the flood frequency analysis, the analysis of runs has attracted less attention in hydrologic studies. A run is defined as a sequence of the stochastic process with a specified property, and in particular, a positive run is an uninterrupted sequence that exceeds a certain level u. The interest may be concentrated on the number of runs N_u in particular time $0 < t < T$, or on the length L_u and area S_u of the run, all considered as random variables, Fig. 5.3. An example to illustrate the importance of including all these properties in a study of extremes may be found in the design of a flood retention structure, as its dimension is a function not only of the maximum instantaneous value of the flood event, but also of the duration and the volume of water associated with the event. Another example is found in degradation for which a relatively low flood flow of extended duration may be more critical than a large flood of a short duration. Similar problems arise in analyses of deficit situations in water supply or water quality problems.

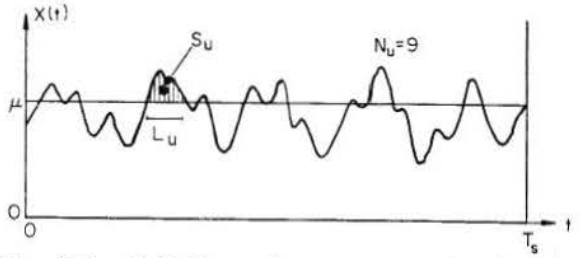

Fig. 5.3 Definition of runs, as run-lengths L_u, run-sums S_u and number of runs N_u.

Studies of run properties of continuous stochastic processes have been made by Rice (1945) and Cramer and Leadbetter (1967). Rodriques (1968), Nordin and Rosbjerg (1970) and Nordin (1971) presents applications of their results to hydrologic time series. In general, the stochastic properties of N_u, L_u and S_u are complex functions of the distribution and autocovariance of the process. However, for normal processes the mean values $\overline{N}_u$, $\overline{L}_u$ and $\overline{S}_u$ depend upon the variance σ^2 and the second spectral moment λ_2 only. Because λ_2 is particularly sensitive to the choice of sampling interval, the estimate of $\overline{N}_u$, $\overline{L}_u$ and $\overline{S}_u$ will be affected accordingly. As in the case with the extremes, a too open sampling interval may not reveal all the variability in the process and consequently may give biased estimates of the true properties.

The average number of positive runs $\overline{N}_u$ has a definite practical importance by indicating how many times the process exceeds a given level u uninterrupted. If the process of interest is a water quality parameter, it denotes how many times a year a given quality standard may be expected to be violated. For a discharge series, it gives the number of separate flood events in a period. Examples of similar importance are easily found.

The mean number of positive runs of a normal process with mean μ and variance σ^2 in a time period T was developed by Rice (1945):

$$\overline{N}_u = T \sqrt{\frac{\lambda_2}{2\pi}} \; \phi(u_s) \; , \tag{5.57}$$

with $u_s = (u-\mu)/\sigma$. An estimate of $\overline{N}_u$ is found by substituting the estimate $\hat{\lambda}_2$ for λ_2. In this case, the mean and variance of the estimate $\hat{\overline{N}}_u$ are found by using the Taylor expansion approximation given in Eqs. 5.49 and 5.51; then

$$E[\hat{\overline{N}}_u] = \frac{T}{\sqrt{2\pi}} \; \phi(u_s) \; [\mu_*^{1/2} - \frac{1}{8} \; \mu_*^{-3/2} \; \sigma_*^2] \tag{5.58}$$

and

$$V[\hat{\overline{N}}_u] = \frac{T^2}{8\pi} \; \phi^2 \; (u_s) \; [\mu_*^{-1} \; \sigma_*^2 + \frac{1}{8} \; \mu_*^{-3} \; \sigma_*^4 \;] \; . \tag{5.59}$$

Again

$$\mu_* = E[\hat{\lambda}_2] = - \; E[\hat{\gamma}''(0)] \tag{5.60}$$

and

$$\sigma_*^2 = V[\hat{\lambda}_2] = V[\hat{\gamma}''(0)] \tag{5.61}$$

which are related to the sampling properties of $\hat{\gamma}''(0)$ found in subchapter 5.4 as functions of the sampling interval Δt. As in the case of extremes, an explicit presentation of the effect of the sampling interval on the properties of N_u is not practical. The expected information loss must be found by numerical computation of the single factors step by step.

Given the mean of $\overline{N}_u$ from Eq. 5.58, the bias is found by subtraction of the population value of $\overline{N}_u$. The variance of $\overline{N}_u$ is determined by Eq. 5.59, and by substituting it into Eq. 3.16, the information loss can be determined for different sampling intervals.

The knowledge about the mean of the run-length $\overline{L}_u$ is of importance in the analysis of particular exceedances such as flood events or water pollution occurrences. It gives the expected duration of the adverse situation, and as such influences economic losses that may be associated with the events.

The expectation of the run-length of a normal process with mean μ and variance σ^2 is given by Cramer and Leadbetter (1967)

$$\overline{L}_u = \frac{T}{\overline{N}_u} \; P[X(0) > u \;] \tag{5.62}$$

or

$$\overline{L}_u = \sqrt{\frac{2\pi}{\lambda_2}} \; \frac{1 - \Phi(u_s)}{\phi(u_s)} \quad , u_s = \frac{u-\mu}{\sigma} \tag{5.63}$$

An estimate of $\overline{L}_u$ is found by substituting the estimate $\hat{\lambda}_2$ for λ_2. In this case the mean and variance of $\hat{\overline{L}}_u$ are found by using the approximation given in Eqs. 5.49 and 5.51

$$E[\hat{\overline{L}}_u] = \sqrt{2\pi} \; \frac{1-\Phi(u_s)}{\phi(u_s)} \; [\mu_*^{-1/2} + \frac{3}{8} \; \mu_*^{-\frac{5}{2}} \; \sigma_*^2 \;] \tag{5.64}$$

and

$$V[\hat{\bar{L}}_u] = \frac{\pi}{2}\left(\frac{1-\Phi(u_s)}{\phi(u_s)}\right)^2 \left[\mu_*^{-3}\sigma_*^2 + \frac{9}{8}\mu_*^{-5}\sigma_*^4\right] \quad (5.65)$$

μ_* and σ^2 are defined by Eqs. 5.60 and 5.61. Equations 5.64 and 5.65 form the basis for computing the expected information loss for different sampling intervals Δt by a method similar to that mentioned for the mean number of runs, $\bar{N}_u$.

The mean run-sum $\bar{S}_u$ is defined as the expected area between the crossing level u and the process itself. For a runoff record, this corresponds to the expected volume of water than must be stored if the riverflow must be kept below the flood level u. Similarly, when a water quality parameter violates a given standard, the mean run-sum denotes the additional amount of waste that created the critical situation. Many more examples of the practical significance of this parameter can be easily found.

Cramer and Leadbetter (1967) developed an expression for the total expected sum TS_u of the normal process, exceeding u in a time period T

$$TS_u = T \int_u^\infty (x-u)\,\phi(x_s)\,dx \;,\; x_s = \frac{x-\mu}{\sigma} \quad (5.66)$$

or

$$TS_u = T\,(\sigma\phi(u_s) - u + u\Phi(u_s)) \;,\; u_s = \frac{u-\mu}{\sigma} \quad (5.67)$$

Nordin and Rosbjerg (1970) used Eq. 5.67 to determine the expected value of the run-sum $\bar{S}_u$ by

$$\bar{S}_u = \frac{TS_u}{N_u} \quad (5.68)$$

or

$$\bar{S}_u = \sqrt{\frac{2\pi}{\lambda_2}}\;\frac{\sigma\phi(u_s) - u + u\Phi(u_s)}{\phi(u_s)} \quad (5.69)$$

An estimate of $\bar{S}_u$ is found by substituting the estimate $\hat{\lambda}_2$ for λ_2. The mean and variance of $\hat{\bar{S}}_u$ are obtained by using the approximations given in Eqs. 5.49 and 5.51 as

$$E[\hat{\bar{S}}_u] = \sqrt{2\pi}\;\frac{\sigma\phi(u_s) - u + u\Phi(u_s)}{\phi(u_s)}\left[\mu_*^{-1/2} + \frac{3}{8}\mu_*^{-\frac{5}{2}}\sigma_*^2\right] \quad (5.70)$$

and

$$V[\hat{\bar{S}}_u] = \frac{\pi}{2}\left(\frac{\sigma\phi(u_s) - u + u\Phi(u_s)}{\phi(u_s)}\right)^2 \left[\mu_*^{-3}\sigma_*^2 + \frac{9}{8}\mu_*^{-5}\sigma_*^4\right] \quad (5.71)$$

It should be noted that as

$$\bar{S}_u = k(u)\,\bar{L}_u \quad (5.72)$$

and

$$k(u) = \frac{\sigma\,\phi(u_s)}{1-\Phi(u_s)} - u_s \quad (5.73)$$

then

$$E[\hat{\bar{S}}_u] = k(u)\,E[\hat{\bar{L}}_u] \quad (5.74)$$

and

$$V[\hat{\bar{S}}_u] = k(u)^2\,V[\hat{\bar{L}}_u] \quad (5.75)$$

The expected information loss can be found by a step-by-step procedure as for the other parameters above.

CHAPTER 6

LOSS OF INFORMATION BY AVERAGE SAMPLING

When statistical analysis of a continuous process of instantaneous values is based on samples of integrated interval values such as the average daily or weekly values, an additional information loss is added to the losses introduced by discrete sampling as described in Chapter 5. However, this additional loss can be accounted for, and with a slight modification it is possible to find the information losses for the average sampling case based on the equations developed in Chapter 5.

6.1 Effect of Average Sampling on the Sampling Properties of a Continuous Process

To find this additional loss, consider the basic continuous process X(t) as an equivalent and discretized realization X_i, with the sampling frequency at or above the critical spectral frequency f_c. This sampling interval is used as the time unit, $\Delta t_* = 1$, so that a period of length T_s has T_s observations of X_i.

As mentioned in subchapter 4.2, the average sampling of X_i is equivalent to discrete point sampling of the moving average process

$$Y_j = \frac{1}{\Delta t} \sum_{i=j}^{j+\Delta t-1} X_i \, , j = 1,2,...T_s/\Delta t \qquad (6.1)$$

where Δt is the interval over which the average is taken, as shown by Fig. 4.4. The variable Y_j is a linear moving average process of X_i and its expected value is

$$E[Y] = E[X] \qquad (6.2)$$

so that the expectation of the Y_j process is undistorted. Its autocovariance function $\gamma_Y(u)$ can be determined from the autocovariance $\gamma(u)$ of the X_i process (Jenkins and Watts 1968) as

$$\gamma_Y(u) = \frac{1}{\Delta t^2} \sum_{i=1}^{\Delta t} \sum_{j=1}^{\Delta t} \gamma(u+i-j). \qquad (6.3)$$

In particular, the variance is

$$V[Y] = \frac{1}{\Delta t} \left[\sigma^2 + 2 \sum_{u=1}^{\Delta t-1} (1 - \frac{u}{\Delta t}) \, \gamma(u)\right] \, . \qquad (6.4)$$

where in general V[Y] becomes smaller than σ^2. However, the process Y_j is not the actual series of average values, but this series can be formed by deriving a new discrete process A_i by sampling Y_j with interval Δt. This process has the expectation

$$E[A] = E[X] \qquad (6.5)$$

variance

$$\sigma_A^2 = V[Y] \qquad (6.6)$$

and the autocovariance

$$\gamma_A(u\Delta t) = \gamma_Y(u\Delta t)$$

$$= \frac{1}{\Delta t^2} \sum_{i=1}^{\Delta t} \sum_{j=1}^{\Delta t} \gamma(u\Delta t+i-j) \qquad (6.7)$$

The smoothing effect of average sampling introduces a distortion of the basic autocovariance of X_i as illustrated by Eq. 6.7.

When the series of average values is used for estimating the properties of a continuous process, this distortion affects both the bias and the variance of the estimate. This effect can be separated into two components:

(1) the bias introduced into the basic process by the averaging procedure, and,

(2) the bias and variance introduced by the estimation procedure in using the average interval values for statistical inference.

The bias of an estimate $\hat{\alpha}$ is the sum of the bias $B_A[\alpha]$ introduced by Eq. 6.7 for the basic parameter α, and the bias $B_E[\hat{\alpha}_A]$ introduced in estimating the parameter α_A of the average process by discrete sampling of the Y_j process, so that

$$B[\hat{\alpha}] = B_A[\alpha] + B_E[\hat{\alpha}_A] \qquad (6.8)$$

The variance of $\hat{\alpha}$ is equal to the variance of a discrete point sampling estimate of the average process parameter α_A

$$V[\hat{\alpha}] = V[\hat{\alpha}_A] \qquad (6.9)$$

The bias and variance due to discrete point sampling $B_E[\alpha_A]$ and $V[\alpha_A]$ can be found from the expressions developed in Chapter 5 when the autocovariance function for the average process, Eq. 6.7, is used instead of the autocovariance of the basic pro-

cess. The bias $B_A[\alpha]$ due to the distortion may be found simply by evaluating the parameter based on the average process and subtracting it from the actual parameter. Thus, both the bias and the variance can be obtained for the average sampling case, and the expected information loss can then be determined as previously from Eq. 3.16.

6.2 Loss of Information in Estimating Distribution Functions

When the distribution function $F_X(x)$ is estimated from the average series A_i, the information loss is found by the variance of $\hat{F}_A(x)$ from Eq. 5.3 with

$$\sigma_Y^2 = F_A(x)\,[1 - F_A(x)] \tag{6.10}$$

and

$$\gamma_Y(u) = P[A_i \leqslant x, A_{i+u\Delta t} \leqslant x] \tag{6.11}$$

The bias due to averaging is

$$B_A[F_X(x)] = F_A(x) - F_X(x) \tag{6.12}$$

If $F_X(x)$ is a normal distribution with a given mean μ and variance σ^2 $F_A(x)$ is normal with mean μ and the variance given by Eq. 6.6, so that $F_A(x)$ can be determined from tables. Bivariate normal tables are used to determine $P[A_i \leqslant x, A_{i+u\Delta t} \leqslant x]$ in Eq. 6.11, where the autocovariance may be computed by Eq. 6.7.

If $F_x(x)$ is assumed to be distributed as a two-parameter Gamma distribution with the shape parameter α and the scale parameter β, the procedure outlined above is still applicable. The average process is distributed as a two-parameter Gamma distribution with

$$\alpha_A = \frac{\mu^2}{\sigma_A^2} \tag{6.13}$$

and

$$\beta_A = \frac{\sigma_A^2}{\mu} \tag{6.14}$$

since μ^2 and σ_A are known, α_A and β_A can be determined and, using tables for the bivariate Gamma function the procedure is similar so that for the normal case.

6.3 Loss of Information in Estimating the Mean

Compared to continuous sampling, the average sampling does not introduce any additional information loss when the mean is estimated. Whether the sample mean is computed based on the continuous sample or based on daily, monthly or annual values makes no change in the value of the estimate, so all these estimates are identical random variables. It then follows that both their expectation and their variance must be equal and hence that the expected information loss does not change with an increase of the sampling interval.

6.4 Loss of Information in Estimating Variance and Autocovariance

The averaging procedure will have a tendency to decrease the total variation in the sample and so to decrease both the variance and the autocovariance. On the other hand, the dependence between two points in time, as expressed by the correlation coefficient, increases. The bias due to averaging may be determined by Eq. 6.7 as

$$B_A[\gamma(u\Delta t)] = \frac{1}{\Delta t^2} \sum_{i=1}^{\Delta t} \sum_{j=1}^{\Delta t} \gamma(u\Delta t + i - j) - \gamma(u\Delta t) \tag{6.15}$$

The estimation bias $B_E[\hat{\gamma}(u\Delta t]$ and variance $V[\hat{\gamma}_A(u\,\Delta t)]$ are found by Eqs. 5.18 and 5.21 if $\gamma_A(u\Delta t)$ of Eq. 6.7 is substituted for the autocovariance $\gamma(u)$ of the underlying process. The total sampling bias and the variance are evaluated and the expected information loss is then found for different values of Δt by Eq. 3.16.

6.5 Analysis of the First Derivative of the Process

Because the average sampling procedure tends to smooth out the largest irregularities of the process, it is evident that the parameter $\gamma''(0)$ is affected by this averaging.

With the distorted autocovariance $\gamma_A(u\Delta t)$, given by Eq. 6.7, the distortion bias is

$$B_A[\gamma''(0)] = \gamma_A''(0) - \gamma''(0) \tag{6.16}$$

where

$$\gamma_A''(0) = -\frac{1}{\Delta t^2}\left[\frac{\pi^2}{3}\,\sigma_A^2 + 4 \sum_{u=1}^{\infty} \frac{(-1)^u}{u^2}\,\gamma_A(u\Delta t)\right] \tag{6.17}$$

and $\gamma''(0)$ is found by Eq. 5.33.

The estimation bias may be found by Eq. 5.36 as

$$B_E[\hat{\gamma}_A''(0)] = -\frac{1}{\Delta t^2}\{\frac{\pi^2}{3} E[\hat{\sigma}_A^2] + 4 \sum_{u=1}^{\infty} \frac{(-1)^u}{u^2} E[\hat{\gamma}_A(u\Delta t)]\} - \gamma_A''(0) \quad (6.18)$$

with $E[\hat{\gamma}_A(u)]$ determined as in subchapter 6.4.

Similarly, $V[\hat{\gamma}_A(0)]$ is found by substituting $\gamma_A(u)$ for $\gamma(u)$ in Eq. 5.39 and using these autocovariances in Eq. 5.37. The total bias and variance of the estimate may now be found and the information loss determined for different Δt.

6.6 Loss of Information in Estimating Probabilities of Extremes

It is well known that average sampling in many cases degrades important information about the extremes. Even for relatively large catchment areas the instantaneous flood flow may be 2-3 times greater than the average daily value of the flood event. So, if the flood analysis is based on daily flows, serious error may be introduced. Similarly, it is impossible to infer much about the instantaneous rainfall intensities from traditional measurements by integration of the bucket raingauge, if these are emptied, say, only every 3 or 6 hours. However, these intensities are important for urban drainage design, and similar problems.

As mentioned in subchapters 6.4 and 6.5, both the variance σ^2 and the parameter $\gamma''(0)$ are distorted by the averaging procedure, and this distortion affects the probabilities of extremes accordingly. If the computation is based on average values, an averaging bias is introduced as

$$B_A[P[\max X(t) \leq u]] = P[\max X_A(t) \leq u] - P[\max X(t) \leq u] \quad (6.19)$$

where the latter term can be determined by Eq. 5.47. The first term may be found from

$$P[\max X_A(t) \leq u] = \Phi(u_{s,A}) \exp\left[-\frac{T\phi(u_{s,A})}{\sqrt{2\Pi}\,\Phi(u_{s,A})}\sqrt{\lambda_{2,A}}\right], \quad (6.20)$$

with $u_{s,A} = \frac{u-\mu}{\sigma_A}$

Here σ_A^2 and $\lambda_{2,A} = -\gamma_A''(0)$ may be found based on Eq. 6.6 and Eq. 6.7, respectively.

The additional uncertainty due to the estimation error, B_E and V, may be found as described in subchapter 5.5, if the variance and autocovariances of the average process are substituted for the corresponding parameters in the underlying process.

Given B_A, B_E, and V, the expected information loss can be computed by Eq. 3.16.

6.7 Loss of Information in Estimating Run Properties

As in the case with probabilities of extremes, the average sampling may cause significant distortions of run properties.

Average sampling will tend to underestimate the number of runs, but overestimate the run lengths. If the mean number of runs of the underlying process is estimated from the average values, the bias due to averaging is

$$B_A[\bar{N}_u] = \frac{T}{\sqrt{2\pi}}[\sqrt{\lambda_{2,A}}\,\phi(u_{s,A}) - \sqrt{\lambda_2}\,\phi(u_s)] \quad (6.21)$$

with $u_{s,A} = \frac{u-\mu}{\sigma_A}$ and $u_s = \frac{u-\mu}{\sigma}$

The estimation bias and variance of $\hat{\bar{N}}_u$ are obtained by Eqs. 5.58 and 5.59, using the parameters of the average process.

Similarly, for the mean run-length, the bias of averaging becomes

$$B_A[\bar{L}_u] = \sqrt{2\pi}\left[\frac{1}{\sqrt{\lambda_{2,A}}}\,\frac{1-\Phi(u_{s,A})}{\phi(u_{s,A})} - \frac{1}{\sqrt{\lambda_2}}\,\frac{1-\Phi(u_s)}{\phi(u_s)}\right] \quad (6.22)$$

and the estimation bias $B_E[\hat{\bar{L}}_u]$ and variance $V[\hat{\bar{L}}_u]$ can be determined by Eq. 5.64 and 5.65.

Finally, the averaging bias of the mean run-sum is

$$B_A[\bar{S}_u] = \sqrt{2\pi}\left[\frac{1}{\sqrt{\lambda_{2,A}}}\,\frac{\sigma_A\phi(u_{s,A}) - u + u\,\Phi(u_{s,A})}{\phi(u_{s,A})} - \frac{1}{\sqrt{\lambda_2}}\,\frac{\sigma\,\phi(u_s) - u + u\,\phi(u_s)}{\phi(u_s)}\right] \quad (6.23)$$

and the estimation bias $B_E[\hat{\bar{S}}_u]$ and the variance $V[\hat{\bar{S}}_u]$ of the run-length can be obtained by Eqs. 5.70 and 5.71.

Given the total bias as the sum of B_A, and B_E and the variance V, the expected information loss can be found for all three parameters by Eq. 3.16.

CHAPTER 7

LOSS OF INFORMATION BY QUANTIZATION

Quantization refers to discretization of the random variable itself. The values of the X-axis are pooled into a set of discrete class intervals

$$i\Delta x - \frac{\Delta x}{2} < X(t) \leq i\Delta x + \frac{\Delta x}{2} \quad , \; i = 0, \pm 1, \pm 2, \ldots,$$

so that the values of $X(t)$ in the interval are replaced by the class interval center $i\Delta x$. This procedure evidently introduces a distortion of the underlying process, but studies have shown that in most cases, the bias of quantization in the population moments averages out, approximately, or can be predicted and corrected for even with a surprisingly coarse quantization.

7.1 Effect of Quantization on the Properties of the Continuous Process

The problem of quantization was studied as early as 1898 (Sheppard, 1898) with the introduction of the Sheppard corrections in the estimated moments of a distribution function. In its present form, the theory is due essentially to Widrow (1956) and Watts (1961), but the following discussion is taken mainly from a presentation by Korn (1966).

Basically, the theory of quantization is based on arguments similar to the reasoning behind the sampling theorem presented in subchapter 4.1, although they are not identical. Consider the continuous probability density function $f_X(x)$ of X of Fig. 7.1.

The distribution function of the quantized variable $f_{X_q}(x_q)$ consists of a series of impulse functions with distance Δx, where the height of each impulse equals the area of $f_X(x)$ for the corresponding interval. The similarity to the discrete sampling situation of Fig. 4.1 is evident, with the only exception that the "discrete sampling" is an areal sampling instead of a point sampling.

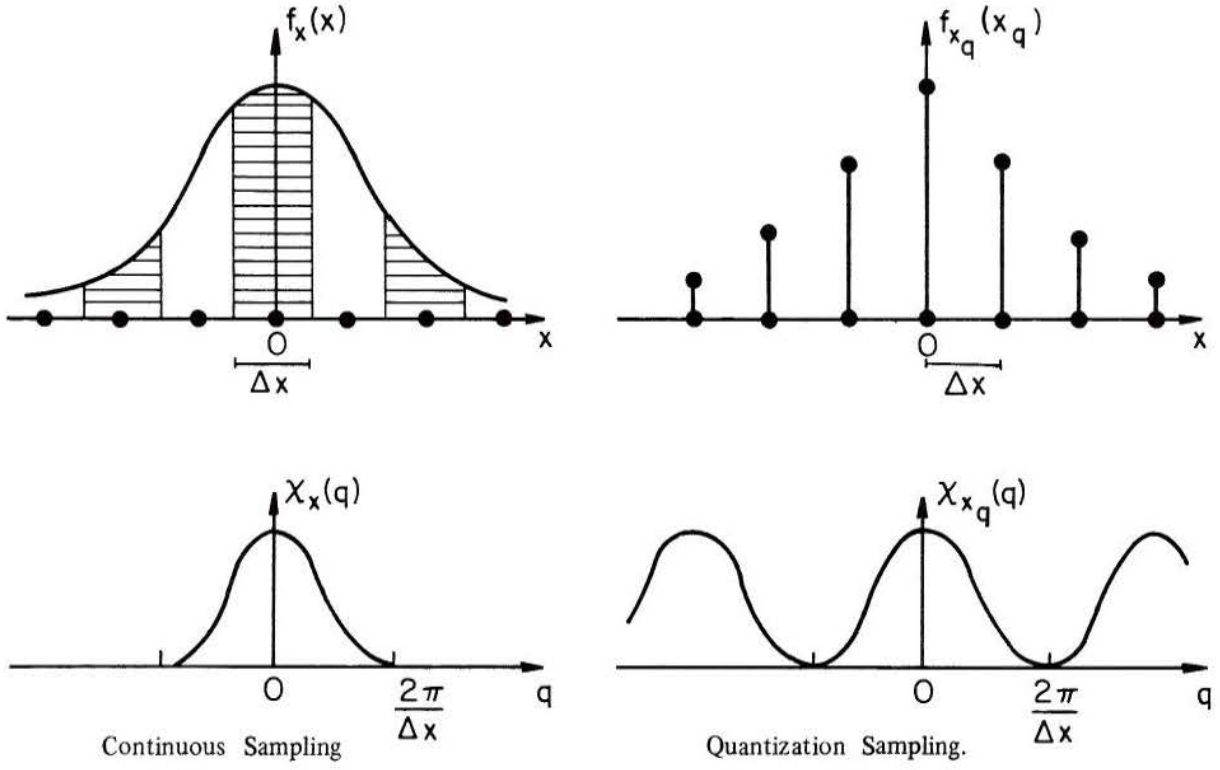

Fig. 7.1 Effect of quantization of a continuous random variable.

Like the case of the sampling theorem, this problem may be studied more easily in the frequency domain by using the Fourier transforms $\chi_X(p)$ and $\chi_{X_q}(p)$ of $f_X(x)$ and $f_{X_q}(x_q)$, or, equivalently, by analyzing the characteristic functions of these distributions. It is then possible to derive the quantization theorem, with its evident analogy to the time sampling theorem, Korn (1966):

"If the density function $f_X(x)$ is 'frequency limited' so that its characteristic function $\chi_X(p) = E[e^{ipx}]$ vanishes for $|p| \geq 2\pi/\Delta x$, then every existing moment $E[x^n]$ of $f_X(x)$ is completely determined by moments of the quantized process $f_{X_q}(x_q)$, and the quantization noise

$$n_q = X_q - X \tag{7.1}$$

is independently and uniformly distributed between $-\Delta x/2$ and $\Delta x/2$."

A similar theorem applies to a bivariate distribution so that the results can be extended to cover the joint moments such as the autocovariances of a stochastic process (Korn, 1966):

"If the joint density $f_{X,Y}(x,y)$ is 'frequency limited' so that the joint characteristic function $\chi_{X,Y}(p_1,p_2) = E[e^{i(p_1x+p_2y)}]$ is zero for $|p_1| \geq 2\pi/\Delta x$ and $|p_2| \geq 2\pi/\Delta y$, then every existing joint moment $E[X^nY^m]$ is completely defined by the joint moments of the quantized process."

The proof of these two theorems comes from the fact that the quantized characteristic function is a periodic copy of the original characteristic function, and as seen in Fig. 7.1, if Δx is small enough, the tails of the characteristic functions do not overlap, and the original characteristic function can be recovered. Because the moments can be found as derivatives of the characteristic functions, it follows that the moments may also be recovered. The sampling was an "areal sampling" in contrast to the point sampling in the time-sampling theorem, so a correction for this distortion must be added, but this correction turns out to be surprisingly small even for coarse quantization intervals.

In particular, it has been shown by Korn (1966) that

$$E[X_q] = E[X] \quad , \tag{7.2}$$

$$V[X_q] = V[X] + \frac{1}{12}\,\Delta x^2 \quad , \tag{7.3}$$

and

$$\gamma_Q(u) = \gamma(u) \tag{7.4}$$

For third and fourth order moments (Widrow, 1957),

$$E[X_q^3] = E[X^3] + \frac{1}{4}\,E[X]\,\Delta x^2 \tag{7.5}$$

and

$$E[X_q^4] = E[X^4] + \frac{1}{2}\,E[X_q^2]\,\Delta x^2 - \frac{7}{240}\,\Delta x^4 \tag{7.6}$$

If the assumption of "frequency limitation" applies, the above equations indicate that the quantization affect on a random variable is either negligible or easily corrected. Actually, the real random variables cannot possibly satisfy this condition exactly, since physical variables are bounded. Nevertheless, many distributions satisfy the quantization theorems so nearly that excellent approximations result. The reason for this is that most continuous distribution functions are relatively smooth, i.e. they lack any significant high-frequency components, so the characteristic function dies out rapidly. In particular, if Eqs. 7.2 through 7.6 are applied to a normal process, this approximation has negligible errors for quantization intervals up to the same order of magnitude as the standard deviation.

The results indicating that the quantization noises are independent cannot hold exactly either. However, even for highly correlated observations this is not a serious limitation. The correlation coefficient between two adjacent observations has been found for the normal case (Widrow 1957)

$$\rho_{n_q} \cong \exp\left[-4\pi^2\sigma^2/\Delta x^2\,(1-\rho_x)\right] \tag{7.7}$$

where ρ_x is the correlation coefficient between the two adjacent values of X. It can be seen from Fig. 7.2 that, even for $\Delta x = 0.5\sigma$, the quantization noises are virtually uncorrelated for $\rho_x < .95$.

If the interval is so large that the quantization effect must be accounted for in estimating a parameter α, it corresponds to an additional bias $B_Q[\alpha]$ similar to the effect of introducing the averaging bias $B_A[\alpha]$ as described in Chapter 6. However, in this case, the bias is much smaller, as only the variance is changed, whereas the mean and the autocovariances are the same.

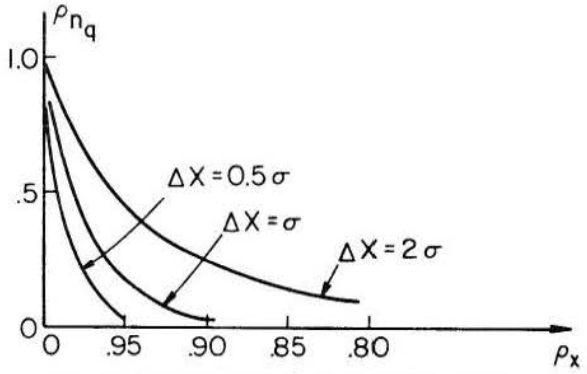

Fig. 7.2 Correlation of quantization noises.

Including the bias from quantization, average sampling and point sampling, the total bias of the sample estimate becomes

$$B[\alpha] = B_Q[\alpha] + B_A[\alpha] + B_E[\hat{\alpha}_{Q,A}] \qquad (7.8)$$

with $B_E[\hat{\alpha}]$ is developed in Chapter 5 and $B_A[\alpha]$ in Chapter 6. Note that it can make a difference whether the averaging or the quantization is performed first. If the quantization is first, $B_A[\alpha]$ should be found based on the quantized properties and vice versa.

The variance of $\hat{\alpha}$ is found from equations in Chapter 5, by using the properties of the distorted process, including both the quantization and averaging effects, or

$$V[\hat{\alpha}] = V[\hat{\alpha}_{Q,A}] \qquad (7.9)$$

where

$$\alpha_{Q,A} = \alpha + B_Q[\alpha] + B_A[\alpha] \qquad (7.10)$$

Hence, both the bias and the variance may be found for given Δt and Δx, and the expected information loss can be evaluated by Eq. 3.16.

7.2 Loss of Information in Estimating Distribution Functions

The distribution function is biased because the bias is introduced into the moments, as shown by Eqs. 7.2 through 7.6. For a normal process with mean μ, and variance σ^2, only the bias in the variance is relevant or

$$B_Q[F_X(x)] = \Phi\left(\frac{x-\mu}{\sqrt{\sigma^2 + \frac{1}{12}\Delta x^2}}\right) - \Phi\left(\frac{x-\mu}{\sigma}\right) \qquad (7.11)$$

and a similar expression can be found for the Gamma distribution.

The bivariate distribution, used to compute σ_Y and $\gamma_Y(u)$ of Eqs. 5.4 and 5.5 as basis for the sampling variance, has the same mean and autocovariances as the basic process, and only the variance is changed according to Eq. 7.3.

Both the bias and the variance of $\hat{F}_x(x)$ can then be found as demonstrated above and the expected information loss evaluated.

7.3 Loss of Information in Estimating the Mean

The bias of the sample mean is zero, but the variance of the estimate increases as

$$V[\hat{\mu}_Q] = V[\hat{\mu}] + \frac{1}{12}\frac{\Delta t}{T_s}\Delta x^2 \qquad (7.12)$$

For an independent process $V[\hat{\mu}] = \Delta t/T_s\ \sigma^2$, so in this case a quantization interval of 0.5σ will introduce an error of only about 2 per cent. For a dependent process, $V[\hat{\mu}]$ is even larger and the relative impact of the error is less. Hence, the influence of quantization on the estimates of mean may be neglected in most cases.

7.4 Loss of Information in Estimating Variance and Autocovariance

Only the variance becomes biased due to the quantization as

$$B_Q[\sigma^2] = \frac{1}{12}\Delta x^2 \qquad (7.13)$$

and

$$B_Q[\gamma(u)] = 0 \quad \text{for} \quad u \neq 0 \qquad (7.14)$$

The estimation bias and variance can be found by equations in subchapter 5.3. For the bias one finds

$$B[\hat{\sigma}_Q^2] = B_E[\hat{\sigma}^2] + \frac{1}{12}\frac{\Delta t}{T_s}\Delta x^2 \qquad (7.15)$$

and the variance is

$$V[\hat{\sigma}_Q^2] \cong V[\sigma^2] + \frac{1}{72}\frac{\Delta t}{T_s}\Delta x^4 \qquad (7.16)$$

For the estimate of autocovariance, $\gamma(u)$, Eq. 5.18 shows that the bias is of the same order of magnitude as in Eq. 7.15, and the effect on the variance can be found from Eq. 5.19. The additional bias and variance introduced by quantization is negligible in most

practical cases, which again illustrates that the uncertainty introduced by the time sampling is by far the most important.

7.5 Analysis of the First Derivative of the Process

The first derivative introduced by quantization into a process is, strictly speaking, always zero. Where the process jumps to another level, it is not even defined. However, if $\gamma''(0)$ is estimated by Eq. 5.42 with quantized data, virtually the same result is obtained as without quantization, as only the variance is biased with the amount $1/12\ \Delta x^2$, and this usually has a negligible influence on the summation. Consequently, quantization does not introduce any significant additional uncertainty in the estimation of $\gamma''(0)$.

7.6 Loss of Information in Estimating Probabilities of Extremes and Run Properties

The probability of extremes $P[\max X(t) \leqslant u]$ and the run properties $\overline{N}_u$, $\overline{L}_u$ and $\overline{S}_u$ are shown in Chapter 5 to depend on the mean μ, variance σ^2 and the second spectral moment λ_2. Only the variance has any significant bias, and as shown above, even this in most cases is negligible. Hence, for these properties, most of the commonly used quantization intervals do not introduce any significant additional uncertainty.

7.7 Joint Effect of Quantization of a Random Variable and Discrete Sampling in Time

The analysis above has considered only the effect of quantization on the loss of information, neglecting the losses introduced by discrete sampling in time presented in Chapter 5 and 6. However, these losses will always occur together, so that their joint effect must be evaluated.

It has been stressed above that the losses due to quantization become surprisingly small even for very coarse x-intervals. For an interval as large as $\Delta x = .5\sigma$, the quantization bias introduced in the variance is only

$$B_Q[\sigma^2] = \frac{1}{12}(0.5\sigma)^2 = .02\sigma^2$$

or only 2% of the total variance. Most hydrologic variables are actually quantized even finer than in this example, so it is evident that, compared to the information loss introduced by time sampling, the quantization effect may usually be neglected completely.

Much effort has been applied on attempts to increase the accuracy of measurements of hydrologic variables. However, if the purpose of data collection is to make inference about stochastic parameters, the results above indicate that the inaccuracies introduced by measurement errors are insignificant compared to time sampling errors, and that in this light very accurate measurements may often be unwarranted.

A sampling program should, therefore, primarily emphasize sampling of many observations in time and relax accuracy requirements to the measurements. This may be particularly important for water quality sampling, where expense of accurate measurements tends to limit the number of samples that are taken. However, the information extracted might increase by trading the measurement accuracy for more frequent sampling.

A similar conclusion about the dominating influence of time sampling errors in contrast to measurement errors has been reached by Moss (1970).

It might be noted, however, that if the data are to be used for deterministic modeling, for example by computation of transfer-functions like unit hydrographs, accurate measurement of the variable is essential.

CHAPTER 8

APPLICATION

The procedures developed in the previous chapters are applied to a stream flow series for a demonstration of the effect of discrete point and average sampling in this particular case. The effect of quantization is negligible here and has not been included in the analysis.

8.1 General Description of the Stream Flow Series

The stream studied is the Davidson River near Brevard in North Carolina. This 40.4 sq. mi. undeveloped catchment area lies in a climatic region with frequent rain storms in the months of July and August, giving rise to a very irregular hydrograph, Fig. 8.1. It must be expected that the loss of information by discrete sampling can be very significant in this case. The catchment is equipped with a digital water-stage recorder installed in the Fall of 1960. From the recorder, 10 years of instantaneous stage observations at 15 min. intervals are available on punched paper tape. The accuracy of the records is excellent, according to reports on surface water data issued by U. S. Geological Survey.

The expressions for information loss are all based on the assumption of stationarity of the stochastic process. However, most hydrologic time series are nonstationary due to the presence of seasonal variations. In such cases, linear transformations may be used to obtain a new stationary series; or the series can be studied over a limited time period only, so that it may be considered approximately stationary.

The latter approach is used here, so the stream flow series is considered as a stationary stochastic process during the flood season in July and August. No deterministic trend in the mean is found during this period. A weak daily periodicity is present, but its contribution to the total variance of the series is negligible, so it has been ignored. Figure 8.1 shows the presence of a cycle in the mean with a period of about a month. This cycle is not present every year and may therefore be considered as a random low frequency component. As this study is primarily an investigation of the effect of discrete sampling on the high frequency properties of the process, the influence of such low frequency components on the information loss is not important.

8.2 Estimation of Parameters of the Continuous Stream Flow Series

The expressions for information loss are all functions of the population parameters of the continuous process, so in order to find the information loss by discrete sampling of Davidson River, estimates of the continuous population parameters must be made. It is then assumed, that these estimates are identical to the population parameters.

Based on the observed series of stream flow data for the months of July and August during the period 1961-1970, the cumulative distribution function was estimated. A Smirnov-Kolmogorov test on a 5 per cent confidence level accepted the hypothesis that the data follows a log-normal distribution. Accordingly, a logarithmic transformation of the original data will satisfy the normal assumption for the equations for probabilities of extremes and run properties developed in previous chapters.

To determine the critical frequency f_c, above which the spectrum can be assumed zero, the spectrum was first estimated based on the 15 minute sampling interval. Neglecting spectral ordinates of a magnitude less than one percent of the spectral ordinate at the origin, the critical sampling interval Δt_*, was found to be 2 hours. Therefore, a discrete series, X_i, derived by sampling the continuous process $X(t)$ every 2 hours is considered the equivalent to the original continuous series.

The mean, variance and autocovariance function of the discrete process, X_i, were estimated by the estimates given in Chapter 5. For each of the 10 years an estimate was made, and the final estimate was formed as the average of these ten estimates.

The estimate of the autocovariance function is presented in Fig. 8.2. The ordinates have been scaled with the variance and correspond to the autocovariance of a dimensionless standardized variable with zero mean and unit variance. This transformation does not affect the generality of the results. The computation of the autocovariance estimates was stopped at the 24 hour lag, where the estimate is close to zero. For larger lag the autocovariances are assumed to be zero.

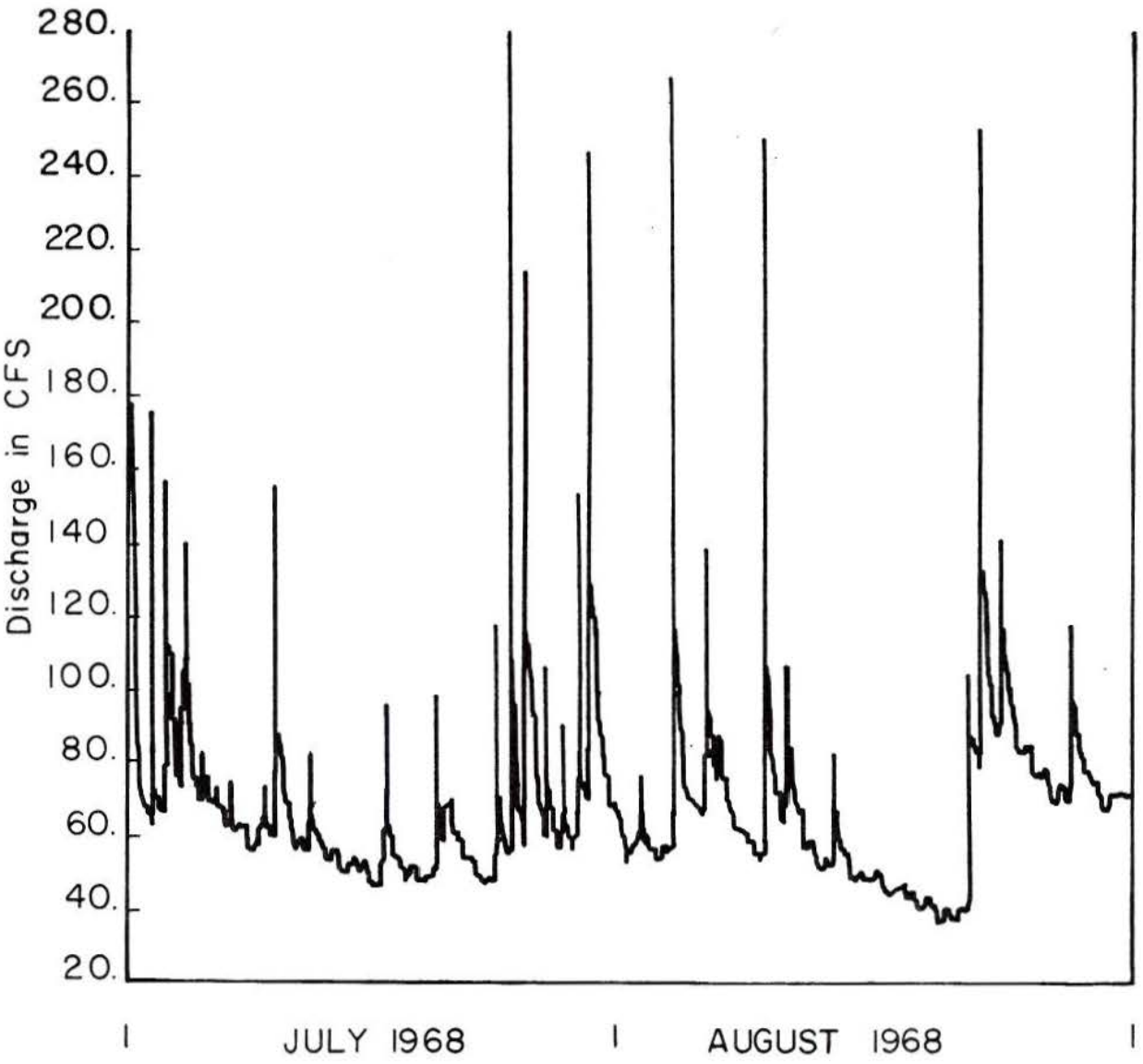

Fig. 8.1 Stream flow series of the Davidson River near Brevard, North Carolina, in the period July-August 1968.

The estimate of the spectrum of the discrete process, X_i, formed from observations every 2 hours, is shown in Fig. 8.3. It can be seen that the assumption about a frequency limit seems reasonable, since the major part of the variance is contributed by frequencies below f = 0.2.

The second derivative of the autocovariance in the origin was estimated to $\hat{\gamma}(0) = -.231$. Based on this value, the probabilities of extremes in a 24 hour period for exceedance levels zero, one, two and three times the standard deviation were estimated. Similarly, estimates of the mean number of runs, the mean run length and the mean run sum were obtained. These results are presented in Tables 8.1 and 8.2.

8.3 Determination of Expected Information Loss

A computer program has been written for computation of the expected information losses based on the expressions developed in Chapters 5 and 6. For given estimates of mean, variance and autocovariance function of the continuous process, the sampling bias, variance and information loss are determined as functions of the sampling interval Δt for both a discrete sampling scheme and an average sampling scheme. The factor k in the information loss equation 3.16 is assumed to be one.

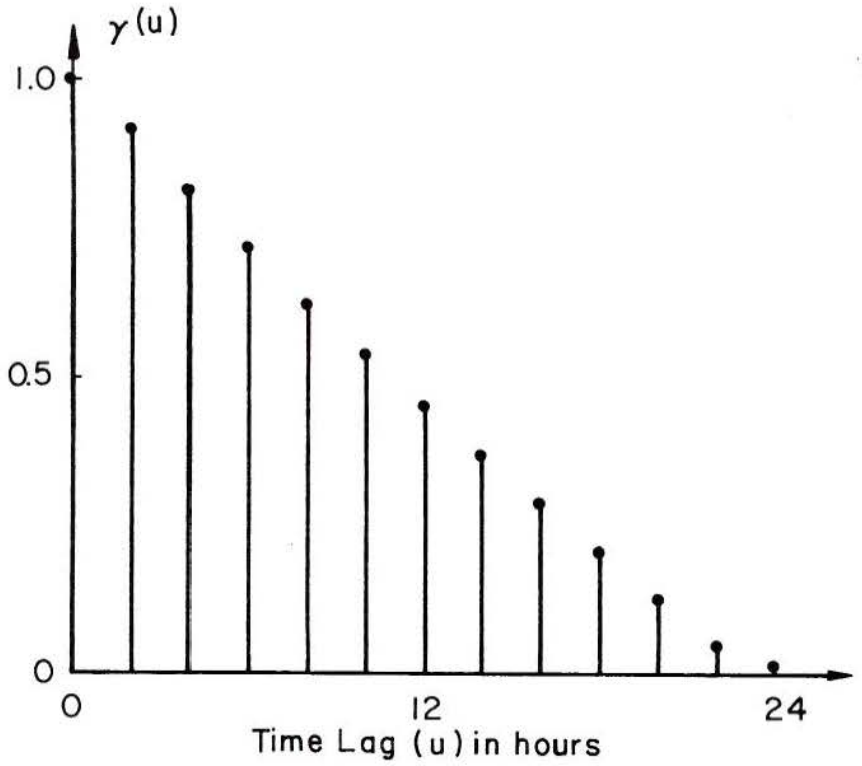

Fig. 8.2 Estimated autocovariance function $\hat{\gamma}(u)$ of Davidson River.

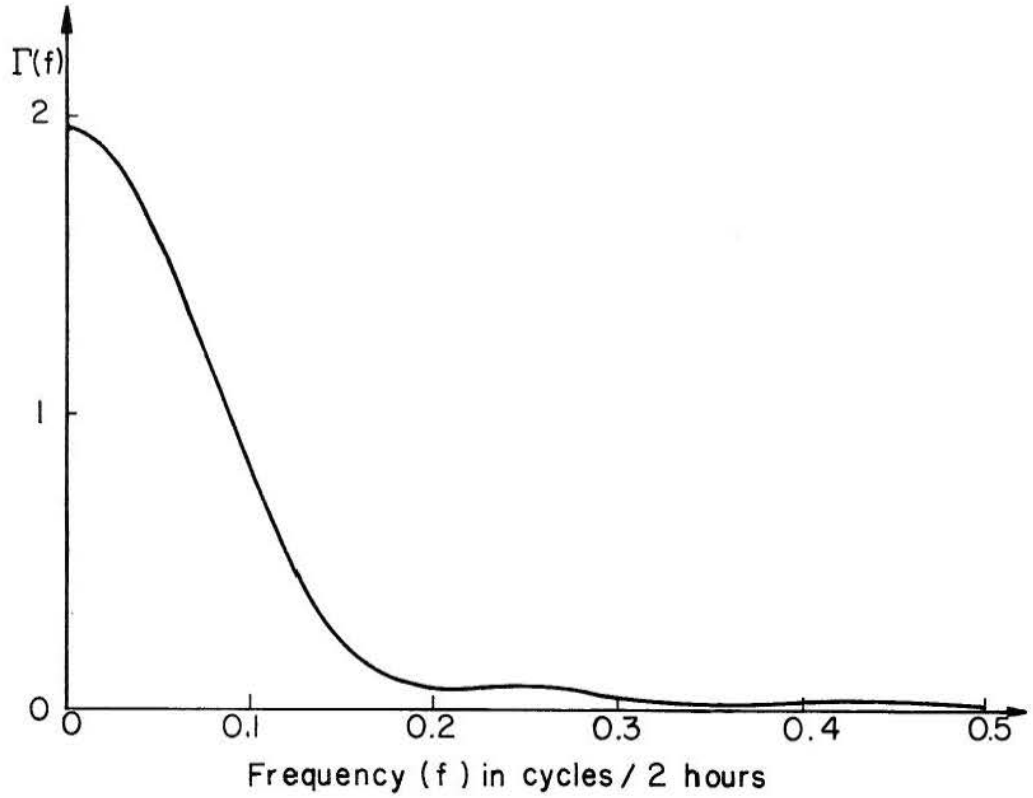

Fig. 8.3 Estimated spectrum $\hat{\Gamma}(f)$ for Davidson River.

TABLE 8.1 Estimates of Probabilities of Extremes

Exceedance level u in standard deviations	Probability of extremes $P(\max X(t) \leq u)$ $0<t<24$ hrs.
0.0	.0799
1.0	.4343
2.0	.8607
3.0	.9883

TABLF 8.2 Estimates of Run Properties

Crossing level u in standard deviations	Mean number of runs per day $\overline{N}_u$	Mean run length $\overline{L}_u$ Time unit = 2 hrs.	Mean run sum $\overline{S}_u$ Time unit = 2 hrs.
0.0	.9172	6.542	5.219
1.0	.5562	3.422	1.797
2.0	.1241	2.199	.820
3.0	.0101	1.590	.450

The bias, variance and expected information loss in estimates based on 20 years of samples of the months July and August of the Davidson River are presented in Appendix A. In this chapter only graphical presentations are shown of the expected information loss as it changes when the sampling interval increases from 2 hours to 24 hours.

Figure 8.4 shows that the information loss of the mean estimate increases very little when the sampling interval is increased from 2 hours to 24 hours. The information content about the mean based on daily observations is virtually the same as if the inference were based on observations every two hours.

The size of sampling interval is more critical for the variance estimate, Fig. 8.5. An increase of Δt from 2 hours to 6 hours does not introduce any significant loss, but if only daily samples are taken, the information loss increases by about 60 percent. If average sampling is introduced, a serious loss of information about the variance results, due to the biasing effect of the sampling procedure. Figure 8.6 shows that the loss is two orders of magnitude larger than for the discretely sampled case, so in this case daily average value contains only a small amount of information about the variance of the underlying continuous process. If only daily average values are available, very little may be inferred about the actual variance. A sample of once a day observations would be much superior in this respect.

The same comments apply to estimation of the autocovariance function, Figs. 8.7 and 8.8. With point sampling, an increase to 6 hour sampling intervals gives negligible loss of information. But if average sampling is introduced, the loss becomes significant for any sampling rate. Hence, if a proper estimation of variance and covariance is of interest, average sampling should be avoided, whereas even daily discrete sampling introduces a moderate information loss.

The information loss by estimation of the second derivative of the covariance is shown in Fig. 8.9. This loss is very sensitive to an increase in Δt, even for discrete point sampling. The loss increases five times when Δt is increased from 2 to 4 hours, an illustration of the contributions of high frequency components that have been neglected by the increase. Further increase results in less drastic increases, and 12 hour sampling and 24 hour sampling are almost equivalent in this respect. Average sampling introduces an additional loss, but the difference between the two sampling schemes is not nearly as large as was the case with the variance and autocovariance, in particular for large sampling intervals.

Figures 8.10 through 8.13 demonstrate the information losses caused by estimation of the probability of extremes for exceedance levels 0.0, 1.0, 2.0 and 3.0 respectively. The general tendencies are the same as for estimation of $\gamma''(0)$. The major loss of information occurs by increasing Δt to 4 hours, and this effect is particularly serious for the highest exceedance levels. Therefore, frequent sampling is of utmost importance, when the probabilities of extremes are of interest. Further increase of Δt increases the loss, but at a slower rate. For example, for the highest level an increase of Δt from 12 hours to 24 hours increases the information loss by only about 15 percent. So in this case an increase of Δt from 12 hours to 24 hours will result in only a minor increase in the information loss and these sampling intervals are almost equivalent. But a increase from 2 hours to 4 hours will increase the information loss by almost two orders of magnitude and the 2 hour sampling interval should be used. If the average sampling scheme is used, the loss will be about twice as much as for discrete point sampling, so the latter is superior for estimation of probabilities of extremes.

Figures 8.14 through 8.17 show the information losses in estimation of the mean run length for crossing levels 0.0, 1.0, 2.0 and 3.0 respectively. The general characteristics are the same as for the probability of extremes. For the lower crossing levels, the rate of increase in information loss by increasing Δt from 2 hours to 4 hours is about half as big as it was for the probability of extremes, but for the high levels it is practically the same. Again, a frequent sampling is essential to prevent large loss of information. The losses by average sampling are about twice as big as for discrete point sampling.

Figures 8.18 and 8.19 show the information loss of the estimate of run lengths. Here, the characteristic property is the steady increase in the loss with increase in Δt, contrasted to the estimates of probability of extremes and mean run length, where the rate of increase was diminishing for larger Δt. The loss introduced by average sampling is about three times the loss by point sampling, so the averaging loses more information than was the case with the two previous properties.

For estimation of the mean run sum, information losses are given in Figs. 8.20 and 8.21. The general characteristics are similar to those for the run length, but the losses increase faster than in that case. The losses due to average sampling are again about three times as large as those caused by discrete point sampling.

In summary it should be noted that for estimation of mean, variance and autocovariance frequent point sampling seems of relatively small importance. For variance and autocovariance estimates, however, average sampling over even relatively small time intervals can give rise to a significant loss of information. But when probabilities of extremes or run properties are of interest, frequent sampling is always essential; and the additional loss introduced by average sampling is a relatively smaller component than was the case for variance and autocovariance estimates.

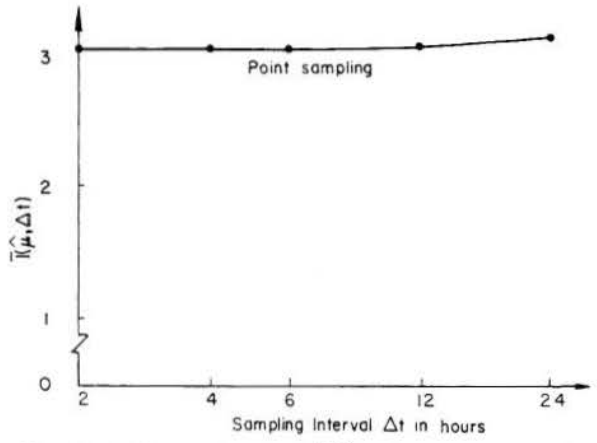

Fig. 8.4 Information loss $\bar{I}(\hat{\mu},\Delta t)$ in estimating the mean.

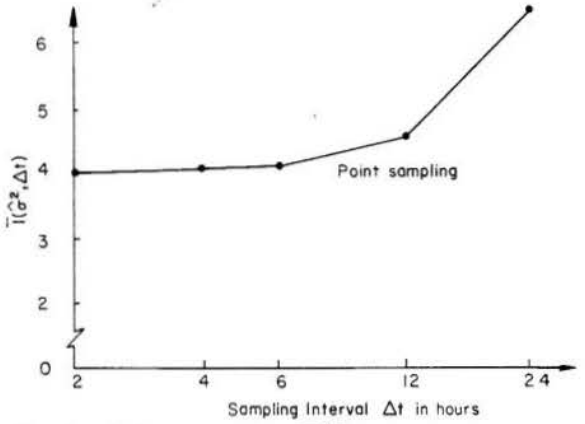

Fig. 8.5 Information loss $\bar{I}(\hat{\sigma}^2,\Delta t)$ in estimating the variance.

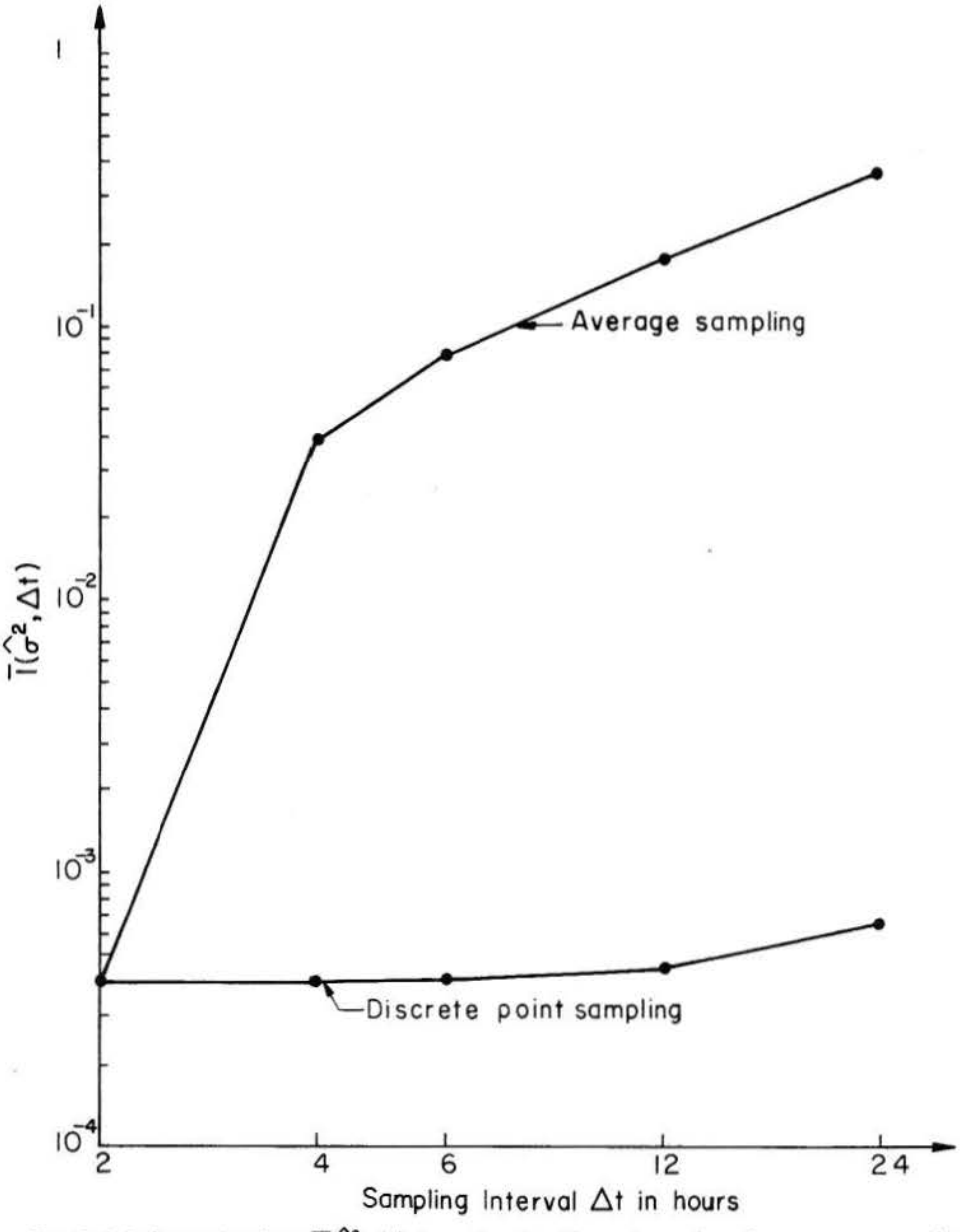

Fig. 8.6 Information loss $\bar{I}(\hat{\sigma}^2,\Delta t)$ in estimating the variance based on average sampling.

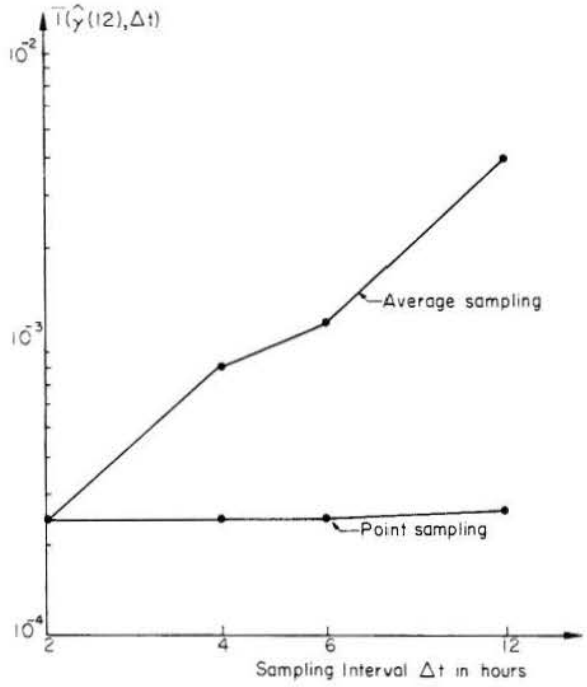

Fig. 8.7 Information loss $\bar{I}(\hat{\gamma}(12),\Delta t)$ in estimating the 12 hour-lag autocovariance.

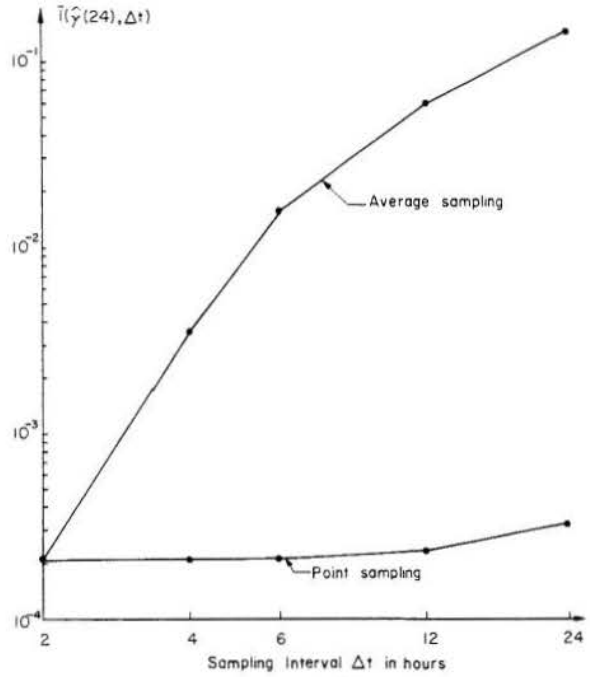

Fig. 8.8 Information loss $\bar{I}(\hat{\gamma}(24),\Delta t)$ in estimating the 24 hour-lag autocovariance.

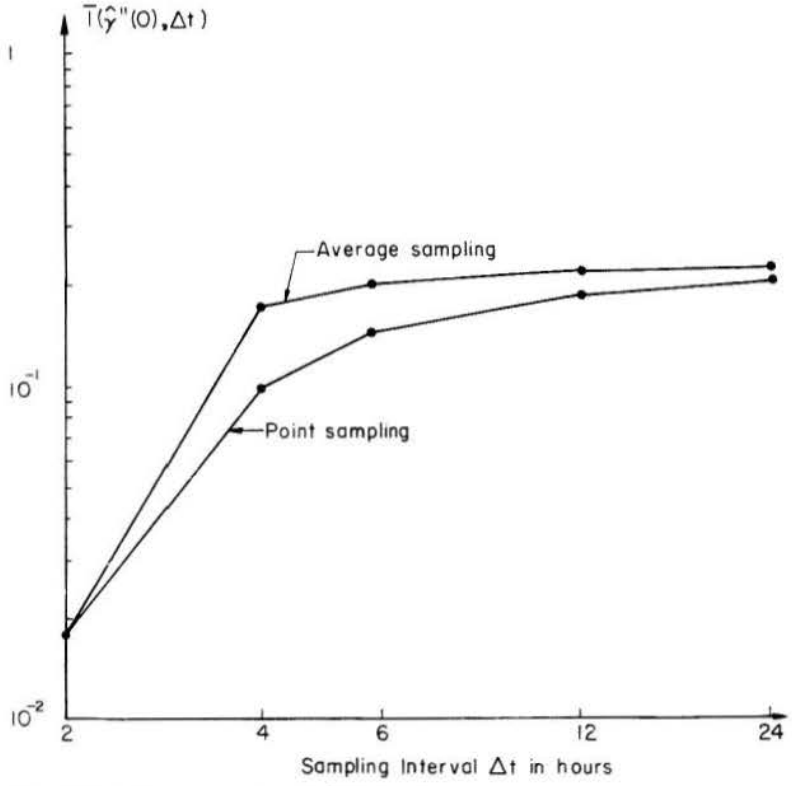

Fig. 8.9 Information loss $\bar{I}(\hat{\gamma}''(0),\Delta t)$ in estimating the second derivative of the autocovariance in the origin.

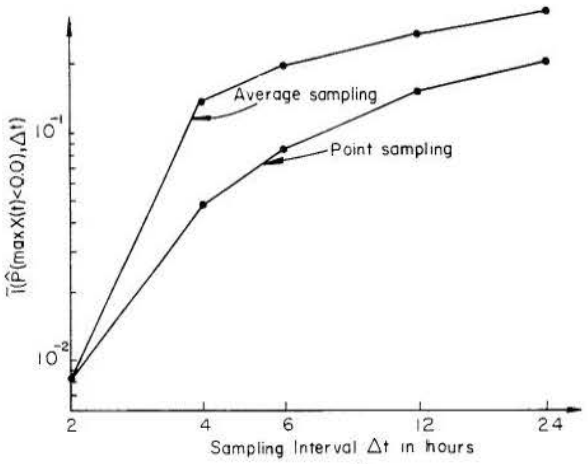

Fig. 8.10 Information loss $\bar{I}(\hat{P}(\max X(t)\leqslant 0.0),\Delta t)$ in estimating probabilities of extremes, Exceedance level u = 0.0.

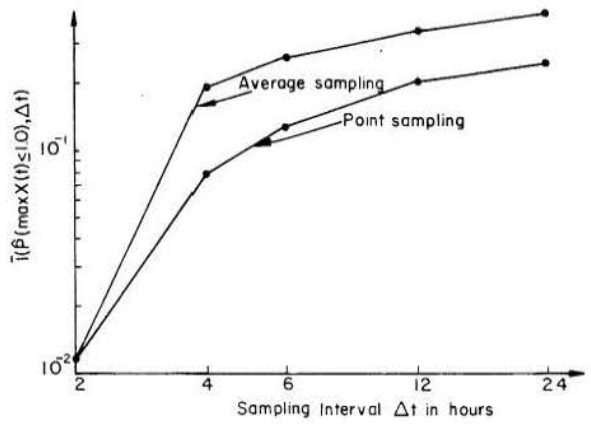

Fig. 8.11 Information loss $\bar{I}(\hat{P}(\max X(t)\leqslant 1.0),\Delta t)$ in estimating probabilities of extremes. Exceedance level u = 1.0 standard deviation.

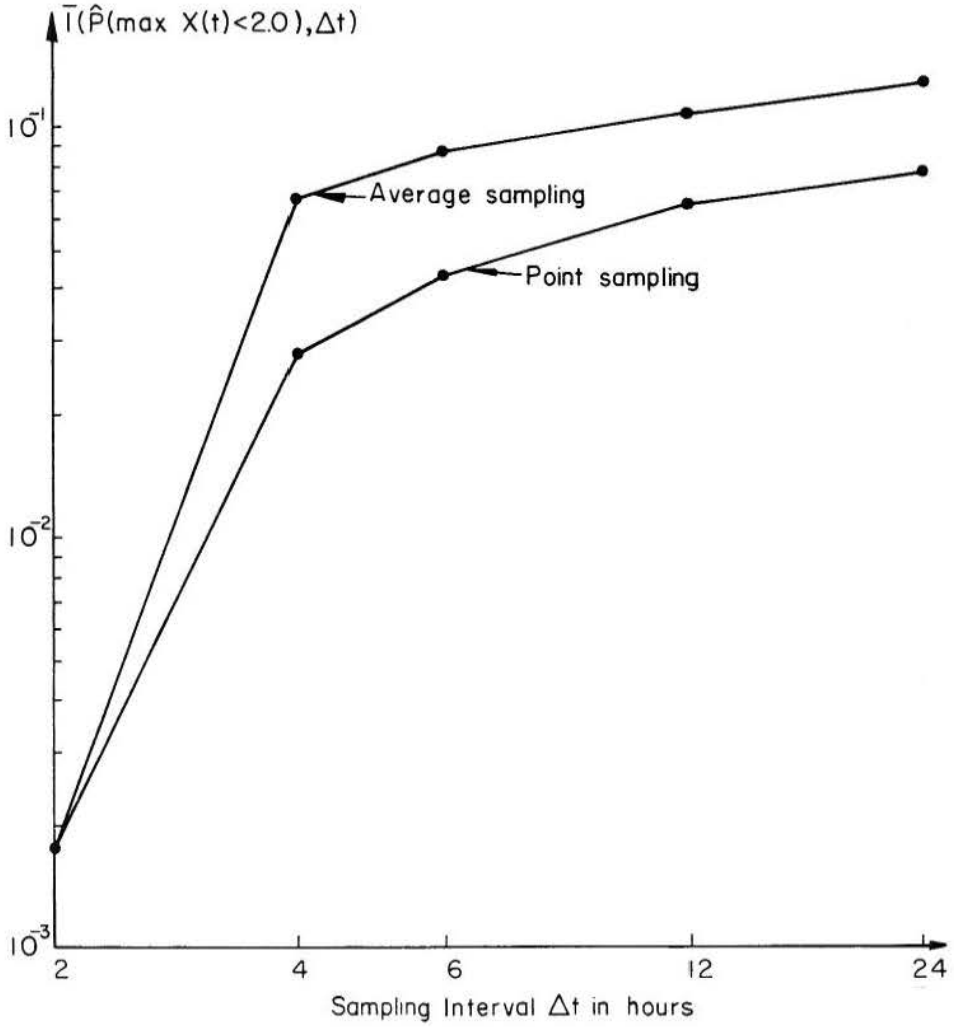

Fig. 8.12 Information loss $\bar{I}(P(\max X(t)<2.0),\Delta t)$ in estimating probabilities of extremes. Exceedance level u = 2.0 standard deviations.

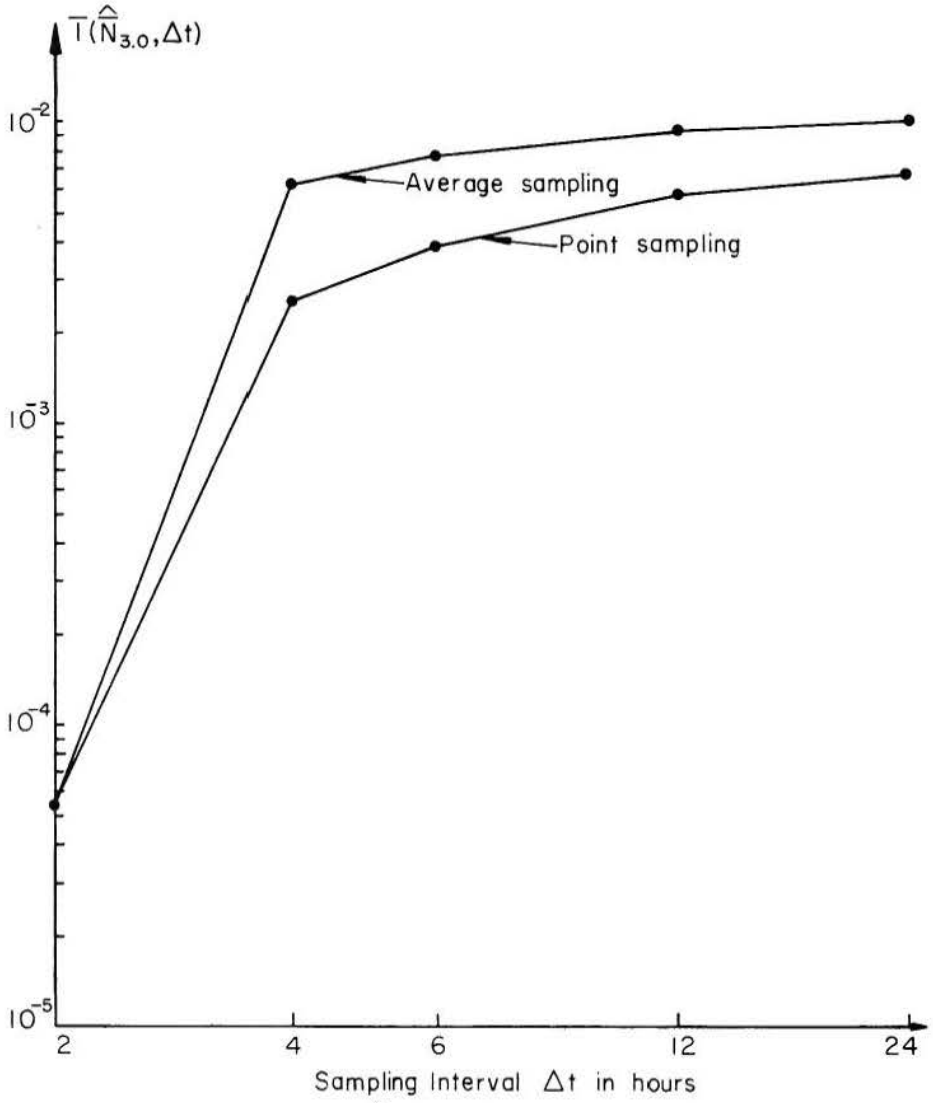

Fig. 8.13 Information loss $\bar{I}(\hat{P}(\max X(t)<3.0), t$ in estimating probabilities of extremes. Exceedance level u = 3.0 standard deviations.

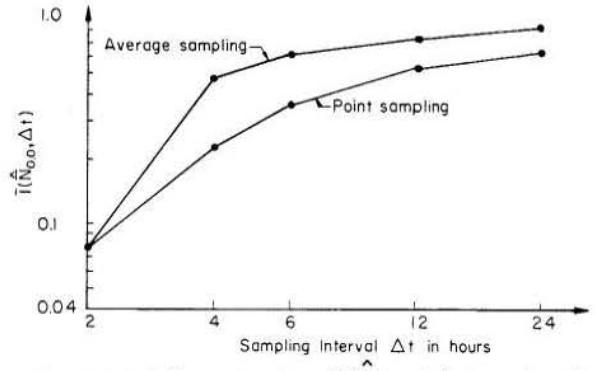

Fig. 8.14 Information loss $\bar{I}(\hat{N}_{0.0},\Delta t)$ in estimating the mean number of runs. Crossing level = 0.0.

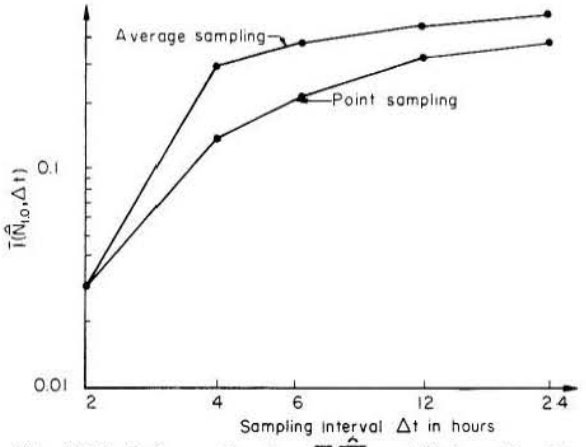

Fig. 8.15 Information loss $\bar{I}(\hat{N}_{1.0},\Delta t)$ in estimating the mean number of runs. Crossing level u = 1.0 standard deviation.

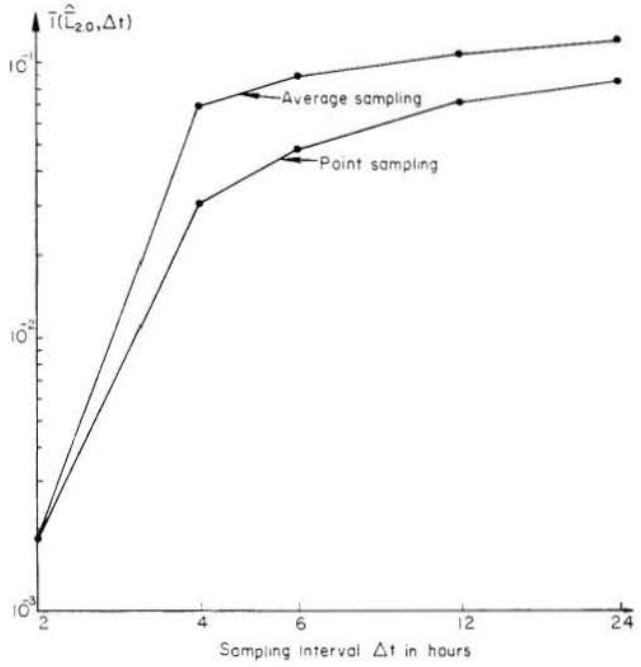

Fig. 8.16 Information loss $I(N_{2.0},\Delta t)$ in estimating the mean number of runs. Crossing level u = 2.0 standard deviations.

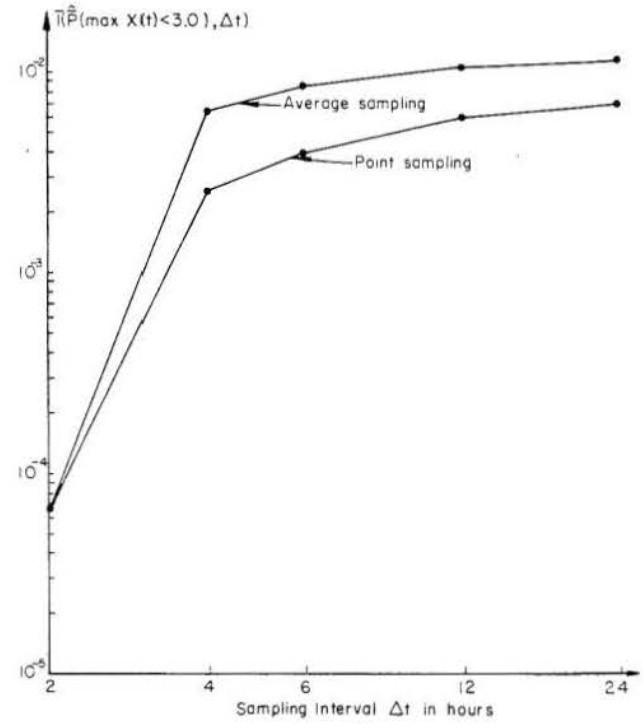

Fig. 8.17 Information loss $\bar{I}(\hat{N}_{3.0},\Delta t)$ in estimating the mean number of runs. Crossing level u = 3.0 standard deviations.

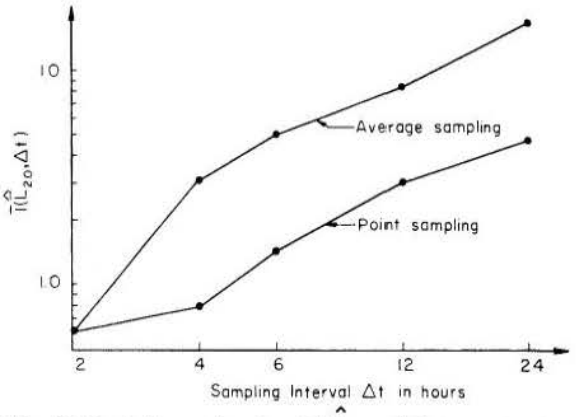

Fig. 8.18 Information loss $\overline{I}(\hat{L}_{2.0},\Delta t)$ in estimating the mean run length. Crossing level u = 2.0 standard deviations.

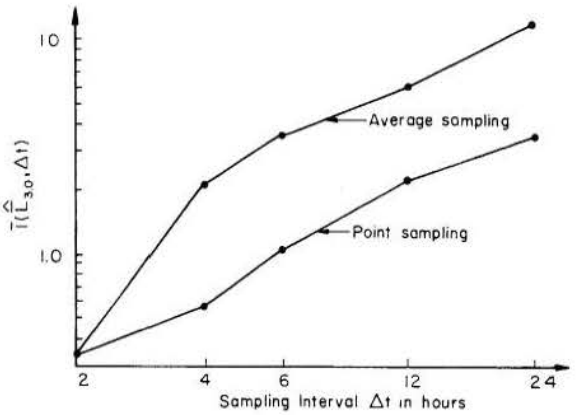

Fig. 8.19 Information loss $\overline{I}(\hat{L}_{3.0},\Delta t)$ in estimating the mean run length. Crossing level u= 3.0 standard deviations.

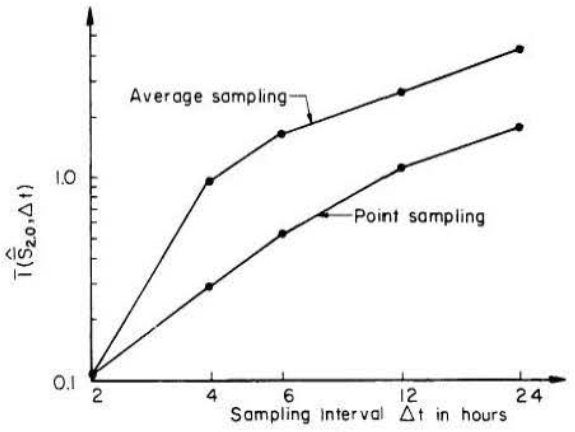

Fig. 8.20 Information loss $\overline{I}(\hat{S}_{2.0},\Delta T)$ in estimating the mean run sum. Crossing level u = 2.0 standard deviations.

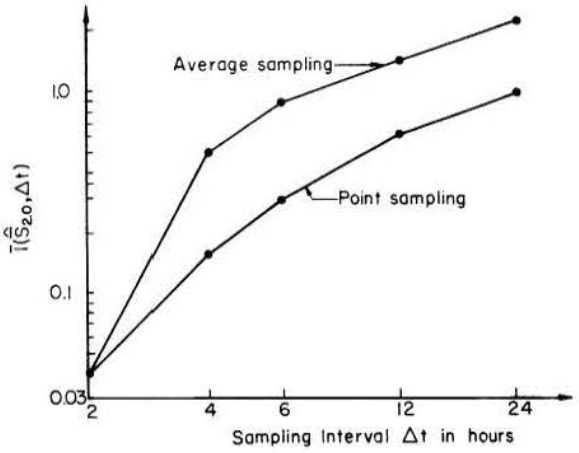

Fig. 8.21 Information loss $\overline{I}(\hat{S}_{3.0},\Delta t)$ in estimating the mean run sum. Crossing level u = 3.0 standard deviation.

CHAPTER 9

SUMMARY AND CONCLUSIONS

9.1 Summary

A general procedure for a quantitative evaluation of loss of information in estimating parameters of a continuous stochastic process based on discrete sampled data has been developed. Three different discretization procedures are considered:

(1) The discrete point sampling scheme, where the process is sampled in time as a series of instantaneous values at periodic time intervals.

(2) The average sampling scheme, where the process is sampled in time as a series of average values over a given time period.

(3) The quantization sampling scheme, where the random variable itself is discretized by lumping its values into class intervals.

As a measure of information content, the decision theoretical concept of "expected information loss," based on a linear loss function, was chosen. It includes the information loss due to both bias and variance of the estimate, and it may eventually provide a framework for assessing the value of the information, as expressed in monetary terms. For the three discretization procedures outlined, general expressions were developed for the expected information loss of selected stochastic properties as functions of the sampling interval, the sampling period, and the mean, variance and autocovariance of the continuous process.

Expressions for information losses were developed for estimates of mean, variance and autocovariance without any assumptions about the distribution of the variable. For normal processes the information losses are found for estimates of the distribution function, of the probabilities of extremes and of the mean number of runs, the mean run length and the mean run sum.

9.2 Conclusions

It has been demonstrated that, for a large majority of hydrologic applications, the information loss due to quantization of the variable is negligible compared to the losses introduced by sampling in time. This implies that when the purpose of the data collection is to make inference about stochastic parameters, frequent sampling in time in order to reduce the time sampling loss is much more important than accurate measurements of the variable itself.

The general expressions developed is applied to a 20 year sampling period of the Davidson River near Brevard in N. Carolina in order to find the information losses due to discrete sampling in this particular case.

It is shown that an increase of the sampling interval from 2 hours up to 24 hours does not introduce any significant loss of information in estimating the mean based on discrete point samples. Therefore, mean estimates based on daily observations are virtually just as accurate as mean estimates based on bi-hourly observations. Average sampling does extract the same amount of information about the mean as continuous sampling.

The information loss in estimating the variance from a sample of instantaneous values does not increase significantly if the sampling interval is increased from 2 hours to 6 hours. However, an increase to a daily sampling interval increases the information loss with about 60 percent. Therefore a relatively frequent sampling rate is more important, when the variance is of interest. A similar conclusion can be drawn about estimation of the autocovariance. In contrast to estimation of the mean, average sampling gives rise to a significant loss of information in estimating variance and autocovariance. Even for a 4 hour sampling interval, the bias due to averaging becomes a dominating factor in the information loss. For a 24 hour interval the information loss about the variance is three orders of magnitude larger than the corresponding loss for a daily instantaneous observation. So when variance and autocovariance of the continuous process are of interest, average sampling can be extremely detrimental. In the case studied here, a series of samples taken once a day contains considerably more information about these properties than a series of daily average values. Nevertheless, it is the latter series that is published; it seems that some consideration should be given to making sampled series of instantaneous values just as readily available, as these actually give a better description of the underlying process than the daily average values.

For estimation of the probabilities of extremes and the mean run length, an increase of the sampling interval gives rise to significant loss of information both for discrete sampling and average sampling. The

loss is particularly large if the sampling interval is increased from 2 hours to 4 hours, whereas an increase of the interval from 12 hours to 24 hours is associated with a relatively small increase of the information loss. It is concluded that frequent sampling is essential for efficient extraction of the information about extreme and run properties. This is the case particularly for estimation of these properties at high exceedance and crossing levels, i.e., for estimation of the rarest events.

A significant characteristic is the relatively small difference between point and average sampling, when the probabilities of extremes and the mean number of runs are considered. In these cases the major loss of information evidently occur because observations with a large sampling interval are unable to detect the high frequency components in the process; the bias due to averaging is of less importance. Average sampling, however, still gives rise to a bigger information loss than discrete point sampling with the same time interval.

The information loss in estimating the mean run length and the mean run sum also rapidly increases with increased sampling interval, but not nearly as fast as the case for the probabilities of extremes and the mean run length. In particular, for point sampling the information loss in the run length increases relatively little by increasing the sampling interval from 2 hours to 4 hours. So frequent sampling is of somewhat less importance here than when probabilities of extremes are of interest. On the other hand, the additional loss due to averaging is of relatively larger significance.

The average sampling scheme always gives rise to an additional loss of information about extremes and run properties as compared to discrete point sampling with the same interval; but the increase is much smaller than was the case for estimation of the variance. It is again concluded, that a series of daily observations actually contains more information about the extremes and runs of the process than a series of daily average values. However, where a series of daily instantaneous observations has lost a relatively small amount of information about the variance, the loss of information about extremes and runs is so large that very little may be inferred about these properties, based on such a series.

9.3 Recommendations for Further Research

The expressions for expected information loss have been applied only in a demonstration of the effect of discretization on one particular stream flow series. A more systematic investigation of a wider selection of streams with different characteristics could be made. Other hydrologic series, such as temperature, sediment load or other water quality parameters might be analyzed.

The information losses developed for extremes and runs are based on the assumption of normal processes. Further research should be directed toward elimination of this assumption.

The information content used here is always related to a specific parameter, not to the stochastic process as such. An information concept could be investigated that would include all or at least several of the most important parameters at the same time. The use of multidimensional loss functions and multivariate distributions of the sample estimates might be a possible approach to this problem.

Finally, the extension of the decision theoretical approach by incorporation of realistic loss functions can form a basis for the optimal choice of discretization procedure and sampling interval. Much research is needed to establish such loss functions as they are a basic factor in the evaluation of the worth of data and for rational investment in data collection systems.

REFERENCES

Bartlett, M. S. (1946), On the theoretical specification and sampling properties of autocorrelated time series: Journ. Roy. Stat. Soc., v.B 8, p. 27-41.

Benes, V. E. (1961), Covariance function of a simple trunk group, with applications to traffic measurements: Bell System Tech. Journ., v. 40, p. 117-148.

Benjamin, J. R. and C. A. Cornell (1970), Probability, statistics and decision for civil engineers: McGraw-Hill, New York.

Brooks, C. E. P., and N. Carruthers (1953), Handbook of statistical methods in meteorology: M. O. 538, Meteorological Office, London.

Cramer, H. (1951), Mathematical methods of statistics: Princeton University Press, Princeton.

Cramer, H., and M. R. Leadbetter (1967), Stationary and related stochastic processes: John Wiley and Sons, New York.

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

DeGroot, M. H. (1970), Optimal statistical decisions: McGraw Hill, New York.

Ditlevsen, O. (1971), Extremes and first passage times with applications in civil engineering: Doctoral thesis, Technical University of Denmark, Copenhagen.

Eagleson, P. S., and W. J. Shack (1966), Some criteria for the measment of rainfall and runoff: Water Resources Research, v. 2, no. 3, p. 427-436.

Fisher, R. A. (1921), The mathematical foundations of theoretical statistics: Phil. Trans. Roy. Soc., Ser. A, v. 222, p. 309-368.

Fisher, R. A. (1925), Theory of statistical estimation: Proc. Cambridge Phil. Soc., v. 22, p. 700-725.

Fisher, R. A., and L. H. C. Tippett (1928), Limiting forms of the frequency distribution of the largest and smallest number of a sample: Cambridge Philos. Soc. Proc., v. 24, p. 180-190.

Fisher, R. A. (1966), The design of experiments: 8th edition, Oliver and Boyd, London.

Gumbel, E. J. (1945), Floods estimated by probability method: Engineering News-Record, June 14.

Hall, W. H. and J. A. Dracup (1970), Water resources systems engineering: McGraw-Hill, New York.

Hartley, R. V. L. (1928), Transmission of information: Bell System Tech. Journ., v. 7, no. 3, p. 535-563.

Hazen, A., (1914), Storage to be provided in impounding reservoirs for municipal water supply: Am. Soc. Civil Eng. Trans., v. 77, p. 1539-1640.

Jenkins, G. and D. G. Watts (1968), Spectral analysis and its applications: Holden-Day, San Francisco.

Jones, R. H. (1962), Spectral analysis with regularly missing observations: Ann. Math. Stat., v. 32, p. 455-461.

Khinchin, A. I. (1957), Mathematical foundations of information theory: Dover Publications, New York.

Knisel, W. G. and V. Yevjevich (1967), The statistical measure of hydrologic time series: Proceedings, Internat. Hydrology Symp., Colorado State Univ., Ft. Collins.

Korn, G. A. (1966), Random process simulation and measurements: McGraw-Hill, New York.

Kullback, S. (1959), Information theory and statistics: John Wiley and Sons, New York.

Lahti, B. P. (1968), An introduction to random signals and communication theory: Scranton, New York.

McMillan, H. (1953), The basic theorems of information theory: Ann. Math. Stat., v. 24, p. 196-219.

McMullen, C. W. (1968) Communication theory principles: McMillan, New York.

Matalas, N. C. and W. B. Langbein (1962), The information content in the mean: Journ. Geophys. Research, v. 67, no. 9, p. 3441-3448.

Mejia, J. M. (1971), On the generation of multivariate sequences exhibiting the Hurst phenomenon and some flood frequency analyses: Ph.D. Dissertation, Colorado State Univ., Ft. Collins.

Moran, P. A. P. (1959), The theory of storage: Methuen, London.

Moss, M. E. (1970), Optimum operating procedure for a river gaging station established to provide data for design of a water supply project: Water Resources Research, v. 6, no. 4, p. 1051-1061.

Nordin, C. F. and D. M. Rosbjerg (1970), Applications of crossing theory in hydrology: Internat. Assoc. Scientific Hydrology, Bulletin, v. 15, no. 1. p. 27-43.

Nordin, C. F. (1971), Statistical properties of dune profiles: U. S. Geol. Survey, Prof. Paper 562-F.

Nyquist, H. (1924), Certain factors affecting telegraph speed: Bell System Techn. Journ., v. 3, no. 2, p. 324-346.

Parzen, E. (1963), On spectral analysis with missing observations and amplitude modulation: Sankhya Ser. A, v. 25, p. 383-392.

Quimpo, R. G. (1967), Stochastic model of daily river flow sequences: Hydrology Paper no. 18, Colorado State Univ., Ft. Collins.

Quimpo, R. G. (1969), Reduction of serially correlated hydrologic data: Internat. Assoc. Scientific Hydrology, Bulletin, v. 14, no. 4, p. 111-118.

Quimpo, R. G., and J. Yang (1970), Sampling considerations in stream discharge and temperature measurements: Water Resources Research, v. 6, no. 6, p. 1771-1774.

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

Raiffa, H. (1968), Decision analysis, Introductory lectures on choices under uncertainty: Addison-Wesley, Reading, Mass.

Rao, C. P. (1965), Linear statistical inference and its applications: John Wiley and Sons, New York.

Reiher, B. S. and C. S. Huzzen (1967), Some comments on the effective sample size of second order Markov processes: Internat. Assoc. Scientific Hydrology, Bulletin, v. 12, no. 4, p. 63-74.

Rice, S. O. (1945), Mathematical analysis of random noise: Bell System Techn. Journ., vv. 24, p. 46-156.

Riordan, J. (1951), Telephone traffic time averages: Bell System Tech. Journ., v. 30, p. 1129-1144.

Rodriquez-Iturbe, I. (1964), Applications of the theory of runs to hydrology: Water Resources Research, v. 5, no. 6, p. 1422-1426.

Savage, L. J. (1950), The foundations of statistics: John Wiley and Sons, New York.

Shannon, C. E. (1948), A mathematical theory of communications: Bell Systems Techn. Jour., v. 27, p. 379-423 and 623-656.

Sheppard, W. F. (1898), On the calculation of the most probable values of frequency-constants for data arranged according to equidistant divisions of a scale: Proc. London Math. Soc., v. 29, p. 353-357.

Siddiqui, M. M. (1962), Some statistical theory for the analysis of radio propagation data: Journ. of Research of Nat. Bureau of Standards, v. 66D, no. 5, p. 571-580.

Theil, H. (1967), Economics and information theory: Rand McNally, New York.

Tick, L. J. and P. Shaman (1966), Sampling rates and appearance of stationary Gaussian processes: Technometrics, v. 8, no. 1, p. 91-106.

Todorovic, P. (1970), On some problems involving random number of random variables: Ann. Math. Stat. v. 41, no. 3, p. 1059-1063.

Watts, D. G. (1962), A general theory of amplitude quantization with application to correlation determination: Proc. Institution of Electr. Engineers, Part C109, p. 209-218.

Widrow, B. (1956), A study of rough amplitude quantization by means of Nyquist sampling: Institute of Radio Engineers, Trans. on Circuit Theory, v. 3, p. 266-276.

Yevjevich, V. M. (1972), Probability and statistics in hydrology: Water Resources Publications, Ft. Collins.

Zelenhasic, E. (1970), Theoretical probability distribution for flood peaks: Hydrology paper no. 42, Colorado State Univ., Ft. Collins.

APPENDIX A

The bias, variance, and expected information loss for estimates of selected stochastic parameter of Davidson River near Brevard, N. Carolina, are given below as functions of the sampling interval Δt for both discrete point sampling and averaging sampling. The basic equations for the computations are found in the subchapters given in table headings. Graphical presentation of the information losses is given in Chapter 8.

TABLE A.1 Estimation of Mean μ
$\mu = 0.0$

Point Sampling, Subchapter 5.2			
Δt hrs.	Bias	Variance $*10^4$	Inf. Loss $*10^4$
2	0	3.77	3.01
4	0	3.78	3.03
6	0	3.78	3.03
12	0	3.84	3.07
24	0	4.03	3.23

TABLE A.2 Estimation of Variance σ^2
$\sigma^2 = 1.0$

Point Sampling, Subchapter 5.3			
Δt hrs.	Bias $*10^4$	Variance $*10^4$	Inf. Loss $*10^4$
2	3.77	4.96	4.02
4	3.78	5.03	4.07
6	3.78	5.09	4.12
12	3.84	5.68	4.59
24	4.03	8.06	6.50

Average Sampling, Subchapter 6.4			
Δt hrs.	Bias $*10^2$	Variance $*10^4$	Inf. Loss $*10^4$
	Average Total		
2	0 3.77	4.96	.04
4	- 4.26 - 4.22	4.91	3.89
6	- 7.91 - 7.97	4.83	7.87
12	-17.79 -17.75	4.46	17.75
24	-35.91 -35.87	3.66	35.88

TABLE A.3 Estimation of Autocovariance $\gamma(12)$; lag = 12 hours $\gamma(12) = .452$

Point Sampling, Subchapter 5.3				
Δt hrs.	Bias $*10^4$ Average	Bias $*10^4$ Total	Variance $*10^4$	Inf. Loss $*10^4$
2		2.85	3.00	2.44
4		2.87	3.03	2.46
6		2.87	3.05	2.47
12		2.93	3.25	2.64
Average Sampling, Subchapter 6.4				
Δt hrs.	Bias $*10^4$ Average	Bias $*10^4$ Total	Variance $*10^4$	Inf. Loss $*10^4$
2	0	2.85	3.00	2.44
4	32.46	35.32	2.99	8.06
6	41.39	44.24	2.97	11.26
12	89.58	92.42	2.87	40.31

TABLE A.4 Estimation of Autocovariance $\gamma(24)$; lag = 24 hours $\gamma(24) = .032$

Point Sampling, Subchapter 5.3				
Δt hrs.	Bias $*10^4$ Average	Bias $*10^4$ Total	Variance $*10^4$	Inf. Loss $*10^4$
2		3.77	2.48	2.06
4		3.78	2.51	2.08
6		3.78	2.54	2.11
12		3.84	2.84	2.34
24		4.03	4.03	3.29
Average Sampling, Subchapter 6.4				
Δt hrs.	Bias $*10^4$ Average	Bias $*10^4$ Total	Variance $*10^4$	Inf. Loss $*10^4$
2		3.77	2.48	2.06
4	79.20	82.92	2.45	35.16
6	193.29	196.95	2.41	157.42
12	579.03	582.54	2.24	582.48
24	1421.47	1424.64	1.94	1424.64

TABLE A.5 Estimation of the second derivative of the autocovariance function $\gamma''(0)$ $\gamma''(0) = -.231$

Point Sampling, Subchapter 5.4				
Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
2		-.0026	.0220	.0177
4		.1005	.0015	.0997
6		.1439	.0003	.1439
12		.1895	.000017	.1895
24		.2078	.000001	.2078
Average Sampling, Subchapter 6.5				
Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
2	0	-.0026	.0220	.0177
4	.1275	.1723	.00144	.1723
6	.1617	.2031	.00028	.2031
12	.1990	.2211	.000015	.2211
24	.2164	.2283	.000001	.2283

TABLE A.6 Estimation of probabilities of extremes, Exceedance level = 0.0, P(max X(t)<0.0) = .080

Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
Point Sampling, Subchapter 5.5				
2		.020	.00304	.0081
4		.051	.00069	.0480
6		.084	.00034	.0843
12		.150	.00008	.1508
24		.200	.00001	.2008
Average Sampling, Subchapter 6.6				
2		.020	.00304	.0081
4	.0667	.1375	.00424	.1331
6	.1034	.1980	.00294	.1979
12	.1734	.2672	.00068	.2672
24	.2368	.3363	.00015	.3363

TABLE A.6 Estimation of probabilities of extremes - (Continued) Exceedance level = 1.0 standard deviation P(max X(t)<1.0) = .434

Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
Point Sampling, Subchapter 5.5				
2		.0226	.0096	.0116
4		.0818	.0014	.0773
6		.1281	.0005	.1281
12		.2021	.00008	.2021
24		.2489	.000007	.2489
Average Sampling, Subchapter 6.6				
2		.0226	.0096	.0116
4	.1165	.1922	.0045	.1915
6	.1701	.2595	.0022	.2596
12	.2640	.3303	.0003	.3363
24	.3611	.4187	.00004	.4187

TABLE A.6 Estimation of probabilities of extremes - (continued)
Exceedance level = 2.0 standard deviation
P(max X(t)<2.0) = .861

Point Sampling, Subchapter 5.5				
Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
2		.0057	.00130	.00174
4		.0286	.00015	.0280
6		.0437	.00005	.0437
12		.0656	.000006	.0656
24		.0783	.000000	.0783
Average Sampling, Subchapter 6.6				
Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
2		.0057	.00130	.00174
4	.0457	.0664	.00033	.0663
6	.0652	.0869	.00012	.0869
12	.0962	.1096	.00001	.1096
24	.1231	.1290	.00000	.1290

TABLE A.6 Estimation of probabilities of extremes - (continued)
Exceedance level = 3.0 standard deviation
P(max S(t)<3.0) = .988

Point Sampling, Subchapter 5.5				
Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
2		.00046	.000011	.00006
4		.00259	.000001	.00254
6		.00394	.000000	.00394
12		.00585	.000000	.00585
24		.00694	.000000	.00634
Average Sampling, Subchapter 6.6				
Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
2		.00046	.000011	.00006
4	.0048	.0064	.000002	.00640
6	.0068	.0083	.000001	.00830
12	.0096	.0102	.000000	.01024
24	.0112	.0113	.000000	.01131

TABLE A.7 Estimation of mean number of runs pr. day, Crossing level = 0.0
$\bar{N}_{0.0}$ = .917 runs per day

Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
Point Sampling, Subchapter 5.6				
2		-0.0392	.091	.0762
4		-0.2338	.010	.2292
6		-0.3558	.003	.3559
12		-0.5289	.0003	.5289
24		-0.6277	.0003	.6277
Average Sampling, Subchapter 6.7				
2		-.0392	.091	.0762
4	-.3037	-.4789	.023	.4781
6	-.4154	-.6137	.009	.6137
12	-.5771	-.7328	.0014	.7328
24	-.6888	-.8249	.0002	.8249

TABLE A.7 Estimation of mean number of runs - (continued)
Crossing level = 1.0 standard deviation
$\bar{N}_{1.0}$ = .556 runs per day

Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
Point Sampling, Subchapter 5.6				
2		-.0237	.0335	.0290
4		-.1418	.00386	.1388
6		-.2158	.00116	.2159
12		-.3208	.00014	.3208
24		-.3807	.00001	.3807
Average Sampling, Subchapter 6.7				
2		-.0237	.0335	.0290
4	-.1924	-.2963	.0083	.2960
6	-.2648	-.3799	.0033	.3799
12	-.3712	-.4559	.0004	.4559
24	-.4516	-.5140	.00005	.5140

TABLE A.7 Estimation of mean number of runs - (continued)
Crossing level = 2.0 standard deviations
$\overline{N}_{2.0}$ = .124 runs per day

Point Sampling, Subchapter 5.6				
Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
2		-.0058	.001667	.00187
4		-.0316	.000192	.03095
6		-.0481	.000058	.04816
12		-.0716	.000007	.07159
24		-.0849	.000001	.08495
Average Sampling, Subchapter 6.7				
Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
2		-.0053	.001667	.00187
4	-.0482	-.0698	.000365	.0698
6	-.0669	-.0895	.000127	.0895
12	-.0943	-.1079	.000010	.1079
24	-.1141	-.1201	.000000	.1201

TABLE A.7 Estimation of mean number of runs - (continued)
Crossing level = 3.0 standard deviations
$\overline{N}_{3.0}$ = .010 runs per day

Point Sampling, Subchapter 5.6				
Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
2		-.000436	.000011	.000054
4		-.002598	.00001	.002540
6		-.003954	0	.005876
12		-.005876	0	.005876
24		-.006973		.006973
Average Sampling, Subchapter 6.7				
Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
2		-.000436	.000011	.000054
4	-.0046	-.006204	.000002	.006204
6	-.0064	-.00789	. 1	.007898
12	-.0087	-.009416	. 0	.009416
24	-.0099	-.0101	0	.010107

TABLE A.8 Estimation of mean run length
Crossing level = 2.0 standard deviations
$\bar{L}_{2.0}$ = 2.20. (time unit = 2 hrs)

Point Sampling, Subchapter 5.6				
Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
2		.3010	.7058	.6139
4		.8214	.2046	.7959
6		1.4406	.1318	1.4405
12		3.0294	.0694	3.0294
24		4.7989	.0171	4.7989

Average Sampling, Subchapter 6.7				
Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
2		.3010	.7058	.6139
4	1.036	2.7437	2.8831	3.0777
6	1.699	4.7718	4.9851	5.0228
12	3.308	8.3479	4.5995	8.3490
24	5.233	16.684	9.2680	16.6840

TABLE A.8 Estimation of mean run length - (continued)
Crossing level = 3.0 standard deviations
$\bar{L}_{3.0}$ = 1.59. (time unit = 2 hrs)

Point Sampling, Subchapter 5.6				
Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
2		.2176	.3688	.3376
4		.5938	.1069	.5690
6		1.0413	.0688	1.0412
12		2.1898	.0362	2.1898
24		3.4689	.0089	3.4689

Average Sampling, Subchapter 6.7				
Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
2		.2176	.3688	.3376
4	.7439	1.9756	1.5000	2.0915
6	1.2166	3.4289	2.5838	3.5283
12	2.3526	5.9607	2.3771	5.9610
24	3.6716	11.7784	4.6449	11.7784

TABLE A.9 Estimation of the mean run sum
Crossing level = 2.0 standard deviation
$\overline{S}_{2.0}$ = .820. (time unit = 2 hrs)

Point Sampling, Subchapter 5.6				
Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
2		.1122	.0981	.1051
		.3063	.0285	.2895
6		.5373	.0183	.5372
12		1.1299	.0096	1.1298
24		1.7898	.00238	1.7898

Average Sampling, Subchapter 6.7				
Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
2		.1122	.0981	.1051
4	.3729	.9875	.3734	.9637
6	.5925	1.6641	.6063	1.6593
12	1.0510	2.6526	.4644	2.6525
24	1.3507	4.3067	.6175	4.3067

TABLE A.9 Estimation of the mean run sum - (continued)
Crossing level = 3.0 standard deviations
$\overline{S}_{3.0}$ = .450. (time unit = 2 hrs)

Point Sampling, Subchapter 5.6				
Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
2		.0615	.0295	.0395
4		.1678	.0086	.1576
6		.2944	.0055	.2944
12		.6192	.0029	.6192
24		.9808	.0007	.9808

Average Sampling, Subchapter 6.7				
Δt hrs.	Bias		Variance	Inf. Loss
	Average	Total		
2		.0615	.0295	.0394
4	.2021	.5639	.1108	.5037
6	.3195	.9004	.1782	.8854
12	.5591	1.4165	.1331	1.4165
24	.6966	2.2347	.1672	2.2347

KEY WORDS: Discretization, information loss, time interval, average sampling, quantization, stream flow series.

ABSTRACT: Three discretization procedures are considered for quantitative evaluation of loss of information when parameters of continuous stochastic processes are estimated on the basis of discrete sampled data: 1) Discrete point sampling, where the process is sampled at periodic time intervals as a series of instantaneous values, 2) Average sampling, where the process is sampled as a series of average values, 3) Quantization of the variable, where its values are pooled into class intervals. The decision theoretical concept of "expected information loss", based on a linear loss function, is used as the measure of information content. For each discretization procedure general expressions for expected information loss in estimating mean, variance & autocovariance are found as functions of the discretization interval, the length of the sampling

KEY WORDS: Discretization, information loss, time interval, average sampling, quantization, stream flow series.

ABSTRACT: Three discretization procedures are considered for quantitative evaluation of loss of information when parameters of continuous stochastic processes are estimated on the basis of discrete sampled data: 1) Discrete point sampling, where the process is sampled at periodic time intervals as a series of instantaneous values, 2) Average sampling, where the process is sampled as a series of average values, 3) Quantization of the variable, where its values are pooled into class intervals. The decision theoretical concept of "expected information loss", based on a linear loss function, is used as the measure of information content. For each discretization procedure general expressions for expected information loss in estimating mean, variance & autocovariance are found as functions of the discretization interval, the length of the sampling

KEY WORDS: Discretization, information loss, time interval, average sampling, quantization, stream flow series.

ABSTRACT: Three discretization procedures are considered for quantitative evaluation of loss of information when parameters of continuous stochastic processes are estimated on the basis of discrete sampled data: 1) Discrete point sampling, where the process is sampled at periodic time intervals as a series of instantaneous values, 2) Average sampling, where the process is sampled as a series of average values, 3) Quantization of the variable, where its values are pooled into class intervals. The decision theoretical concept of "expected information loss", based on a linear loss function, is used as the measure of information content. For each discretization procedure general expressions for expected information loss in estimating mean, variance & autocovariance are found as functions of the discretization interval, the length of the sampling

KEY WORDS: Discretization, information loss, time interval, average sampling, quantization, stream flow series.

ABSTRACT: Three discretization procedures are considered for quantitative evaluation of loss of information when parameters of continuous stochastic processes are estimated on the basis of discrete sampled data: 1) Discrete point sampling, where the process is sampled at periodic time intervals as a series of instantaneous values, 2) Average sampling, where the process is sampled as a series of average values, 3) Quantization of the variable, where its values are pooled into class intervals. The decision theoretical concept of "expected information loss", based on a linear loss function, is used as the measure of information content. For each discretization procedure general expressions for expected information loss in estimating mean, variance & autocovariance are found as functions of the discretization interval, the length of the sampling

period, and the mean, variance and autocovariance of the continuous process. A stream flow series is analyzed to show the applicability and potential of the approach. The loss due to quantization is found to be neglible. A small sampling interval is essential to prevent a large information loss about extremes and runs, but is of less importance for estimation of mean and variance. Average sampling introduces significant losses of information due to the biasing effect inherent in the sampling procedure. With the exception of the mean, a sample of instantaneous values contain more information about the parameters investigated than a sample of average values, taken over the same sampling interval.

Ref: Loss of Information by Discretizing Hydrologic Series - Mogens Dyhr-Nielsen
Hydrology Paper #54

period, and the mean, variance and autocovariance of the continuous process. A stream flow series is analyzed to show the applicability and potential of the approach. The loss due to quantization is found to be neglible. A small sampling interval is essential to prevent a large information loss about extremes and runs, but is of less importance for estimation of mean and variance. Average sampling introduces significant losses of information due to the biasing effect inherent in the sampling procedure. With the exception of the mean, a sample of instantaneous values contain more information about the parameters investigated than a sample of average values, taken over the same sampling interval.

Ref: Loss of Information by Discretizing Hydrologic Series - Mogens Dyhr-Nielsen
Hydrology Paper #54

period, and the mean, variance and autocovariance of the continuous process. A stream flow series is analyzed to show the applicability and potential of the approach. The loss due to quantization is found to be neglible. A small sampling interval is essential to prevent a large information loss about extremes and runs, but is of less importance for estimation of mean and variance. Average sampling introduces significant losses of information due to the biasing effect inherent in the sampling procedure. With the exception of the mean, a sample of instantaneous values contain more information about the parameters investigated than a sample of average values, taken over the same sampling interval.

Ref: Loss of Information by Discretizing Hydrologic Series - Mogens Dyhr-Nielsen
Hydrology Paper #54

period, and the mean, variance and autocovariance of the continuous process. A stream flow series is analyzed to show the applicability and potential of the approach. The loss due to quantization is found to be neglible. A small sampling interval is essential to prevent a large information loss about extremes and runs, but is of less importance for estimation of mean and variance. Average sampling introduces significant losses of information due to the biasing effect inherent in the sampling procedure. With the exception of the mean, a sample of instantaneous values contain more information about the parameters investigated than a sample of average values, taken over the same sampling interval.

Ref: Loss of Information by Discretizing Hydrologic Series - Mogens Dyhr-Nielsen
Hydrology Paper #54