Reprinted from *Estuarine and Coastal Marine Science* (1978) **6**, 217–224

Notes and Discussions

Environmental Survey Design: A Time Series Approach[a]

Woollcott Smith

Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, U.S.A.

Received 14 July 1976 and in revised form 13 December 1976

Keywords: surveys, environmental surveillance, time series, statistical models, icthyoplankton, Narragansett Bay

In general, the goal of an environmental sampling program is to estimate some parameter of the population being studied. However, in most natural systems, that parameter is changing over time. A survey design must allocate sampling effort between sampling at a single time point and sampling over time. We have used results from sampling theory for stochastic processes to describe an efficient sampling program for estimating the mean of a time-varying parameter of the population. These theoretical results are applied to a 3-year ichthyoplankton sampling program in Mount Hope Bay, Rhode Island.

Introduction

In the design of environmental surveys, decisions about the number of samples to be taken and the number of times the survey is to be repeated can substantially affect the quality of information obtained from the survey and the survey's cost. Saila *et al.* (1976), Kelley & McManus (1969) and others, have discussed elementary sampling theory methods (Sokal & Rohlf, 1969) for allocating sampling and replicate sampling effort to minimize the errors in estimating the population mean. In these discussions it is assumed that the parameter to be estimated is fixed; that is, that the process does not vary with time. In reality the goal of most environmental surveys is to evaluate the long-term behavior of a population parameter over time. There is a rather large statistical literature on sampling time varying processes (Scott & Smith, 1974; Blight & Scott, 1973; Cochran, 1963), but to the best of our knowledge none of these ideas have appeared in the biological or environmental literature.

The survey design depends on both the process sampled and the goals of the survey. We can not, in this note, discuss all the possible alternatives. Instead, in the spirit of the papers that investigated the sampling design for a single time point, we will develop first a simple sampling model for a process that varies over time. The optimum design for estimating the mean of a time varying parameter of the process is described. Finally, the design procedure is applied to an ichthyoplankton survey in Mount Hope Bay, Rhode Island.

Sampling model for a time varying system

Sampling models will vary considerably from one application to another. In this section we will present a rather general model; in later sections we will apply the model to the special case of the ichthyoplankton survey in Mount Hope Bay.

[a]Woods Hole Oceanographic Institution Contribution No. 3804

An implied assumption in a baseline survey is that there is a parameter of the population being investigated that is relatively stable over time. Let $\theta(t)$ denote the value of that parameter at time t. In a particular application $\theta(t)$ might denote the concentration of pollutant at a time t or a measure of difference in plankton densities in two areas, etc. In any case over time θ is continually perturbed by events which are difficult if not impossible to predict. From the point of view of the planner of an environmental survey, $\theta(t)$ must be considered to be a realization of a random process. In many cases we want to find the average value of that realization over a fixed time interval (o, T), so we need to design a survey that will efficiently estimate the quantity

$$\bar{\theta} = \frac{1}{T}\int_0^T \theta(t)dt. \tag{1}$$

In places where fixed monitoring instruments can continuously sample over time the sampling design problem is much simplified. However, in many cases, particularly in chemical and biological sampling, samples must be taken at discrete points in time, often at considerable expense. At each survey time point we have only an estimate of the true value of $\theta(t)$; that is, at discrete time points $t_1, t_2 \ldots t_n$ we observe

$$\hat{\theta}_i = \theta(t_i) + \varepsilon_i, \quad i = 1,2 \ldots n. \tag{2}$$

We assume the ε_i's are independent random errors due to sampling and measurement errors at time t_i. We assume that ε_i has mean o and sampling variance, σ_s^2 defined by

$$\sigma_s^2 = \mathrm{Var}[\varepsilon_i]. \tag{3}$$

The size of σ_s^2 will depend on the sampling effort expended at each time point.

If we know little about the time series properties of the process $\theta(t)$ a relatively robust sampling and estimation procedure for $\bar{\theta}$ is the following. Let N denote the number of time points when sampling will occur. Place the sampling time at the midpoints of each of the N equal intervals that divide the time interval (o,T). That is

$$t_i = \frac{1}{2}\frac{T}{N} + \frac{T}{N}(i-1), \quad i = 1,2 \ldots N. \tag{4}$$

The estimate of $\bar{\theta}$ is then

$$\hat{\bar{\theta}} = \frac{1}{N}\sum_{i=1}^{N} \hat{\theta}_i. \tag{5}$$

The mean squared error of this estimate is

$$\begin{aligned} \sigma_{\bar{\theta}}^2(N) &= E\left[(\hat{\bar{\theta}} - \bar{\theta})^2\right] \\ &= E\left[\frac{1}{N}\sum_{i=1}^{N}\left(\theta(t_i) - \bar{\theta}\right)^2\right] + \frac{\sigma_s^2}{N} \\ &= \sigma^2(N) + \frac{\sigma_s^2}{N}. \end{aligned} \tag{6}$$

$\sigma^2(N)$, the first term of the right hand side of this equation, is the mean squared error due to sampling only at N discrete time points. The second term on the right hand side is the mean squared error due to sampling and measurement error at a single time point. For the

time interval (0,1) Tabilla (1975) has derived an expression for $\sigma^2(N)$ when $\theta(t)$ is a stationary stochastic process with covariance function

$$C(x) = E\{[\theta(t)-\bar{\theta}][\bar{\theta}(t+x)-\bar{\theta}]\}.$$

The expression is

$$\sigma^2(N) = \frac{1}{N}\sum_{i=-N}^{i=+N}\left(1-\frac{|i|}{N}\right)C\left(\frac{i}{N}\right)+2\int_0^1(1-x)C(x)\,dx-\frac{2}{N}\sum_{i=1}^{N}\int_0^1 C(t_i-x)\,dx. \quad (7)$$

For systematic sampling, Quenouille (1949) and Cochran (1946) have derived similar results.

In most of the cases the covariance function, $C(x)$, is unknown. However, we can make a simplifying assumption that the process $\theta(t)$ is Markovian, which in turn implies that

$$C(x) = C_0 e^{-\beta x}. \quad (8)$$

The Markov process $\theta(t)$ is then defined by two parameters: C_0, which is the variance of the process, and β, which can be thought of as the rate at which perturbations in the process decay over time. In a later section we will discuss how to obtain an estimate of C_0 and β from previous sampling programs. Substituting equation (8) into (7) and after some work we obtain for the interval $(0,T)$

$$\sigma^2(N)=C_0\left[\frac{1}{N}-\frac{2}{\beta T}(1+U/\beta T)+\frac{2}{V}\left(e^{-\beta T/N}\left(1-\frac{U}{V}\right)+\frac{2}{\beta T}e^{-\beta T/2N}\right)\right], \quad (9)$$

where the constants U and V are defined by

$$U = 1-e^{-\beta T}$$

and

$$V = N(1-e^{-\beta T/N}).$$

Substituting equation (9) into (6) we see that the survey design problem is to find the σ_s^2 and N that minimize equation (6), subject to certain limits on the cost of the entire survey

Allocating sampling effort

In the preceding section we proposed an estimate, $\bar{\theta}$, for the time averaged mean of a continuous process, and found the variance of the estimator. When we assume that $\theta(t)$ is a Markov process, $\sigma_{\bar{\theta}}^2(N)$ is dependent on C_0 and β (the variance and decay rate of the Markov process) on N [the number of sampling times in the interval $(0,T)$] and on σ_s^2 (the variance of the sampling error at any time point). One can increase the accuracy of a sampling program either by increasing the number of sampling times (that is, by increasing N) or by increasing the accuracy of the estimate of $\theta(t)$ at each time point (that is, by decreasing σ_s^2). Both alternatives have an associated cost.

The actual cost of a survey is a nonlinear function of σ_s and N. The form of this function will depend on the particular application. In this note we assume that there is a fixed cost, g, for including a single time point in the survey, and that for a single time point the cost of estimating θ_i with variance θ_s^2 is $f(\sigma_s^2)$. The total cost, S, of the survey is then

$$S = Ng+Nf(\sigma_s^2) \quad (10)$$

Formally stated, the problem is to find the N that minimizes $\sigma_{\bar{\theta}}^2(N)$ subject to the constraint that the cost not exceed some fixed amount, say S_0. There does not appear to be a simple closed form solution to this minimization problem. However, for each problem one can simply vary N and search for the minimum by evaluating each solution on the computer. The example in the next section will illustrate this procedure.

Finally in the special case where $\theta(t)$ is the mean of a population and a random sampling program is conducted, the variance of the single time point estimate is

$$\sigma_s^2 = \frac{\sigma^2}{m}, \tag{11}$$

where σ^2 denotes the population variance and m denotes the number of samples taken at each time point. If the cost of analysing a single sample is h, we have

$$f(\sigma_s^2) = m \cdot h \simeq \frac{\sigma^2}{\sigma_s^2} \cdot h. \tag{12}$$

For other single time point sampling designs, the cost function will be more complicated.

Design of an ichthyoplankton survey

We have discussed in general terms the problem of sampling a time varying process $\theta(t)$ to estimate the mean of that process in the interval $(0,T)$. Equations (2), (7), and (10) describe the trade-off between spending more sampling effort to estimate $\theta(t)$ at a fixed point in time and the cost of sampling at additional time points. The above results are more or less useful depending on (1) the basic question that the survey is to answer, (2) the appropriateness of the Markov model to the process investigated, and (3) our ability to estimate the statistical parameters of the model β, σ^2 and C_0.

In this section we will use data from a 3-year study of Mount Hope Bay, Rhode Island, conducted by Marine Research, Inc. for New England Power Company to illustrate that this kind of survey can fit our model reasonably well and that at least rough estimates of the parameters of the model can be obtained from available data.

Figure 1 gives a map of Mount Hope Bay showing the 23 ichthyoplankton stations that were sampled at weekly intervals during the spring and summer spawning months. The sampling periods extended from the middle of March to the end of August in the years 1973, 1974, 1975. The results of these surveys are reported in Marine Research progress reports to New England Power Company which operates a 1600 megawatt steam plant at Brayton Point. These reports are available from New England Power Company, Westboro, Massachusetts.

In this hypothetical example we will assume that a future survey is proposed to estimate the relative differences in abundance of total ichthyoplankton populations between the upper and lower half of Mount Hope Bay. Stations 1–12 are in the upper half of Mount Hope Bay and stations 13–23 are in the lower half. We need to find a parameter of the population that reflects the relative abundance between the lower and upper half of the Bay, and to estimate the value of that parameter averaged over the entire spawning period.

Let $X_{t,i,j}$ denote the total ichthyoplankton densities at time t for the jth station in sector i: $i = 1$ and 2, the upper and lower sectors of Mount Hope Bay, respectively. Assuming a multiplicative model we can write

$$X_{t,i,j} = \mu_t \, a_{t,i} \, \rho_{t,i,j}, \tag{13}$$

where μ_t is the overall density at time t, $a_{t,i}$ denotes the proportional increase or decrease in sector i, and $\rho_{t,i,j}$ denotes the proportional random error due to sampling at station j in sector i at time t. A parameter of this model that reflects the relative difference between the two sectors is the log ratio between the relative densities in sectors 1 and 2,

$$\theta(t) = \text{Log}\,(a_{t,1}/a_{t,2}). \tag{14}$$

An estimate of the parameter $\theta(t)$ is

$$\hat{\theta}(t) = \bar{Y}_{t,1} - \bar{Y}_{t,2}, \tag{15}$$

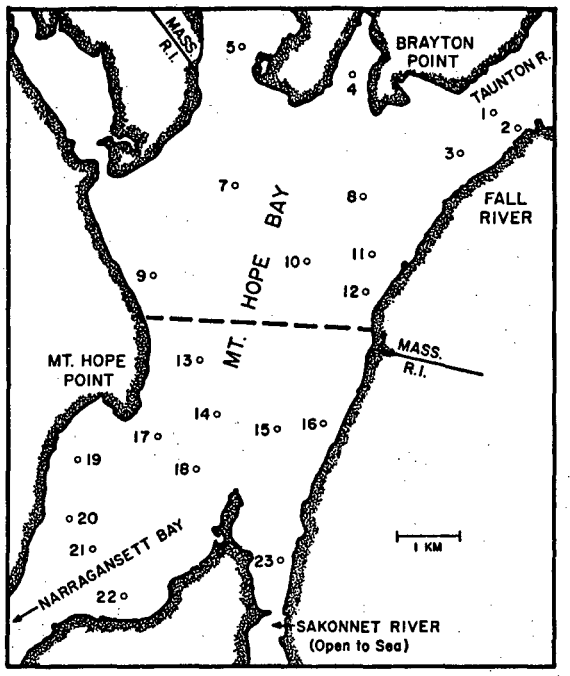

Figure 1. Sampling plan for the Mount Hope Bay ichthyoplankton study. The 23 stations were sampled at weekly intervals during the spring and summer from 1973–1975. Sector 1 is above the dashed line, and sector 2 is below. (Center of chart is at about 41°42′N. Lat.; 71°12′E. Long.)

where

$$\bar{Y}_{t,i}=\frac{1}{m_i}\sum_{j-1}^{m} Y_{t,i,j}$$

m_i denotes the number of stations in sectors i, and $Y_{t,i,j}$ is the log-transform of $X_{t,i,j}$. Figure 2 gives a plot of the estimates of $\hat{\theta}(t)$ over the three survey periods. If we further simplify the model by assuming a random sampling scheme for both sectors and that log $(\rho_{t,i,j})$ has mean 0 and variance σ^2, then the unbiased estimate for the sampling variance at a single time point is

$$\sigma^2=\frac{1}{N(m_1+m_2-2)}\sum_{t=1}^{N}\sum_{i=1}^{2}\sum_{j=1}^{m_i}\left(Y_{t,i,i}-\bar{Y}_{t,i}\right)^2. \tag{16}$$

The variance of $\hat{\theta}_i$ at a single time point is

$$\sigma_s^2=\left(\frac{1}{m_1}+\frac{1}{m_2}\right)\sigma^2. \tag{17}$$

From the past survey we can obtain estimates for σ_s^2, C_0 and β. An estimate of the sampling variance of $\hat{\theta}_i$ is given by equations (16) and (17). The variance of $\hat{\theta}_i$ over all time is the sum of the independent effects of variation in the $\theta(t)$ process and sampling error,

$$\text{Var}\,(\theta_i)=\sigma_s^2+C_0. \tag{18}$$

An estimate of C_0 is then

$$\hat{C}_0=\frac{1}{N-1}\sum_{i=1}^{N}\left(\theta_i-\hat{\bar{\theta}}\right)^2-\hat{\sigma}_s^2, \tag{19}$$

The covariance of $\hat{\theta}_t$ with $\hat{\theta}_{t+1}$ is

$$C(T/N) = C_0 e^{-\beta T/N}, \tag{20}$$

and thus an estimate of β is

$$\hat{\beta} = \text{Log}\,(\hat{C}_0/\hat{C}_1)/\frac{T}{N}, \tag{21}$$

where $\hat{C}_1$ is the covariance of $\hat{\theta}_t$ with $\hat{\theta}_{t+1}$,

$$\hat{C}_1 = \frac{1}{N-1}\sum_{t=1}^{N-1}(\hat{\theta}_t -)\hat{\bar{\theta}}\,(\hat{\theta}_{t+1} - \bar{\theta})\ . \tag{22}$$

For the data given in Figure 2 the estimates of the parameters were $\hat{\sigma}^2 = 0{\cdot}59$, $\hat{C}_0 = 0{\cdot}27$, $\hat{\beta} = 0{\cdot}86$.

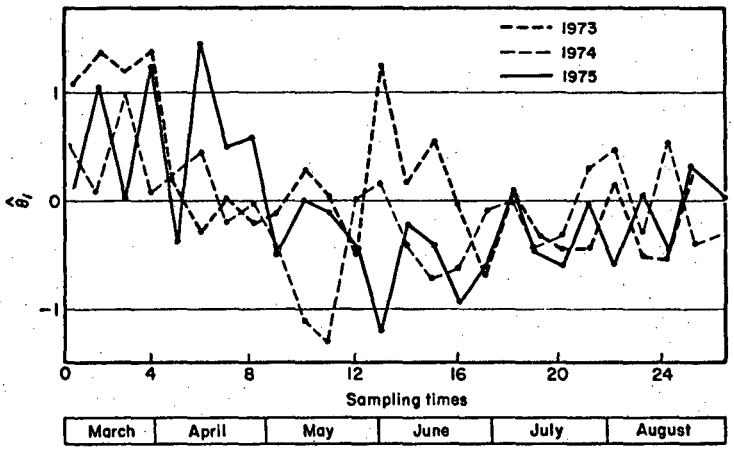

Figure 2. The estimated log ratio of total ichthyoplankton densities between the upper and lower half of Mount Hope Bay for the 1973, 1974 and 1975 spring and summer months.

To apply the results we need information on the cost of sampling and the total maximum cost of the survey. In this hypothetical example we might set the total cost of the survey at \$50,000, the cost of analysing a single sample at \$50, and the cost of hiring a boat, crew and sampling personnel for one day's sampling at \$750. We will also assume that the number of stations in each sector is equal, $m_1 = m_2$.

The trade-off between the number of time points and the number of samples within a time point can be investigated by tabulating the possible combinations of N and m that produce surveys with a total cost less than \$50,000. Figure 3 gives the results of these computations on a log–log plot. The number of sampling periods are plotted against the mean squared error of the estimated time averaged mean, equations (6), (9), and (17). We have fixed σ^2 and C_0 at 0·59 and 0·27, respectively, the estimates from the data. With weeks as the time unit, β is varied to illustrate how the sampling design will change depending on the rate at which changes in the $\theta(t)$ occur. For large β ($\beta \gg 10$) the process varies rapidly over time, and results will be improved by more frequent sampling. In our example, the minimum is about 40 times in the 26-week period with four samples in each sector. With $\beta = +0{\cdot}86$ (the estimate from the 3 years' data) the process varies rather rapidly over the sampling period and the hypothetical minimum is achieved by sampling 30 times during

the spring–summer spawning period. This is very close to the actual number of 26 time points used in Marine Research, Inc. survey. With $\beta = 0{\cdot}1$ and $\beta = 0{\cdot}01$, we see that the optimum number of time points decreases because with a slowly varying process it becomes less important to sample frequently. When $\beta \ll 0{\cdot}001$ or nearly 0, the optimum strategy is to sample at only one time.

An important feature of these results is that they are relatively insensitive to small changes in β and that there is a fairly wide choice of sampling frequencies that yield mean squared errors that are close to the optimum. This is fortunate since at the design stage of an environmental survey, we have only a rough knowledge of the stochastic properties of the $\theta(t)$ process. For practical reasons the sampling period must be in even multiples of days. This will usually not be at the theoretical minimum, but since the function $\sigma^2_{\bar{\theta}}(N)$ is nearly level in the neighborhood of the minimum, one will not lose much efficiency by choosing a sampling period that conforms with the calendar.

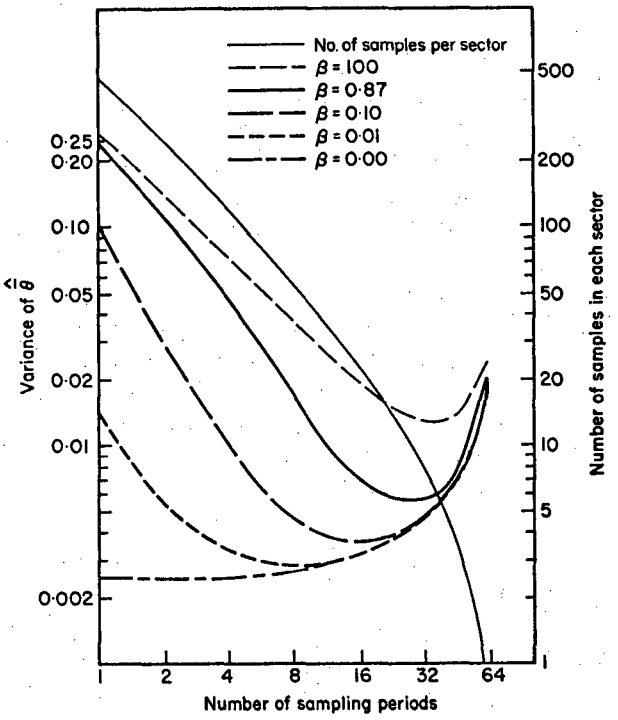

Figure 3. The expected mean squared error for different sampling plans with a fixed cost of \$50 000.00. The fixed cost determines the maximum number of samples in a sector (solid line) given the number of sampling periods.

The sampling theory discussed here does not include the effect of aliasing. Aliasing occurs when processes with a natural periodic component due to tidal cycles, daily fluctuations, etc., are sampled. The Markov processes discussed in this note do not exhibit this kind of periodic variation. If there is a large periodic component in the process sampled, the error in estimating the time averaged mean will be dependent on the relationship between the sampling frequency and the frequency of the periodic component.

Conclusions

In environmental surveys carried out over time, a critical decision is the level of effort to expend in surveys at a single time point and the number of times a survey is to be repeated over time. Knowledge about the variability of the system and the rate of change of the system can be used to find survey designs that will yield better estimates of the time averaged parameter of the population.

We have not discussed what parameters of a population or ecosystem are important and whether the time averaged values of these parameters are the most useful indicators of the state of the system. Certainly in a survey for mean production, total inputs of environmental pollutants, concentration of CO_2 in the atmosphere, etc., the mean value over a time interval is an important quantity. However, for other problems, mean value may well not be important. For instance, the maximum of an environmental variable may be more important than its mean value over time. This is certainly the case in studies of thermal pollution from power plants or in studies of smog in urban environments.

The survey problem discussed here is a rather specific example of a large and varied group of survey design problems. Before we can evaluate a survey design the goals of the survey must be carefully stated. Surveys designed to efficiently answer one set of questions may be of little use in answering a different set of equations.

Acknowledgements

This paper benefited from many conversations with George C. Matthiessen, Donald W. Bourne and Paul Souza of Marine Research, Inc. Melvin Rosenfeld, Loren Haury, Clay Sassaman and Earl Hays provided helpful suggestions on earlier drafts of the manuscript. This work was supported by NOAA Sea Grant 04-6-158—44016.

References

Blight, B. J. W. & Scott, A. J. 1973 A stochastic model for repeated surveys. *Journal of the Royal Statistical Society Ser. B* **35**, 61–66.

Cochran, W. G. 1946 Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics* **17**, 164–177.

Cochran, W. G. 1963 *Sampling Techniques.* 2nd edition, J. Wiley & Sons, New York, 413 pp.

Kelley, J. C. & McManus, D. A. 1969 Optimizing sediment sampling plans. *Marine Geology* **7**, 465–471.

Quenouille, M. H. 1949 Problems in plane sampling. *Annals of Mathematical Statistics* **20**, 355–375.

Saila, S. B., Pikanowski, R. A. & Vaughan, D. S. 1976 Optimum allocation strategies for sampling benthos in the New York Bight. *Estuarine and Coastal Marine Science* **4**, 119–128.

Scott, A. J. & Smith, T. M. F. 1974 Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association* **69**, 674–678.

Sokal, R. R. & Rohlf, F. J. 1969 *Biometry: The Principals and Practice of Statistics in Biological Research.* W. H. Freeman and Company, San Francisco, 376 pp.

Tubilla, A. 1975 Error convergence rates for estimates of multidimensional integrals of random functions. *Department of Statistics, Stanford University Technical Report* No. 72.

Printed in Great Britain by Henry Ling Ltd., at the Dorset Press, Dorchester, Dorset