Ecological Monographs, 61(1), 1991, pp. 53–73

DESIGNING ENVIRONMENTAL FIELD STUDIES[1]

L. L. EBERHARDT AND J. M. THOMAS
Pacific Northwest Laboratory, P.O. Box 999, Richland, Washington 99352 USA

Abstract. Field experiments in ecological and environmental research usually do not meet the criteria for modern experimental design. Subsampling is often mistakenly substituted for true replication, and sample sizes are too small for adequate power in tests of significance. In many cases, field-study objectives may be better served by various kinds of sampling procedures, even though the resulting inferences will be weaker than those obtainable through controlled experimentation.

The present paper provides a classification and description of methods for designing environmental studies, with emphasis on techniques as yet little used in ecology. Eight categories of techniques for field studies are defined in terms of the nature of control exerted by the observer, by the presence or absence of a perturbation, and by the domain of study. The first two categories include classical experimental approaches, replicated and unreplicated. Sampling for modelling provides efficient designs for estimating parameters in a specified model. Intervention analysis measures the effect of a known perturbation in a time series. Observational studies contrast selected groups from a population, while analytical sampling provides comparisons over the entire population. Descriptive survey sampling estimates means or totals over an entire population, while sampling for pattern deals with spatial patterns over a selected region.

We propose that the statistical concept of a "superpopulation" may be useful in ecology, and that it may be desirable to approach community and ecosystem studies in a sampling framework, with experimentation used for a fairly narrow range of subsidiary investigations. Much more attention to processes for drawing inferences about cause and effect is needed, in any case.

Key words: analytical sampling; experimental design; field experiments; inferences; observational studies; pattern; populations; pseudoreplication; sampling design; sampling for modelling; Type I and II errors.

INTRODUCTION

Much of what we know as the scientific method is based on the idea of experimental investigation of an hypothesis. Beginning in the 1920s, statistical methodology was developed for the design and analysis of such experiments, so that the bulk of current research reports in many fields involve statistical analyses of the data. However, a strictly experimental approach to field studies is both difficult and expensive to achieve in the environmental sciences, particularly if worthwhile statistical analyses are to be incorporated. In an especially compelling demonstration, various court decisions forced detailed field studies under the National Environmental Policy Act. Many "impact studies" were subsequently conducted. Most of these confused subsampling and treatment replication (Eberhardt 1976).

The basic problem in impact studies is that evaluation of the environmental impact of a single installation of, say, a nuclear power plant on a river, cannot very well be formulated in the context of the classical agricultural experimental design, since there is only one "treatment"—the particular power-generating station. In a wider context, good experimental control may make it impossible to study the essential phenomena in an ecological system. If the system is more than the sum of its parts, ultimate understanding requires observation of the intact, functioning whole. Many fields of scientific endeavor have the same sorts of difficulties, including astronomy, economics, medicine, and sociology.

Recently Hurlbert (1984) reviewed a sizable number of ecological studies, and again pointed out that subsampling is often mistakenly assumed to constitute replication. He also provided valuable details of the frequency of experiments without replication in ecological field research. Hurlbert classed all such studies as being either "manipulative experiments" or "mensurative experiments." The manipulative class included situations in which the investigator controls circumstances of the study. Most scientists regard these as "experiments." In such investigations, the scientist assigns different treatments to experimental units in a specified manner, usually including untreated "controls."

The second class involved only passive observation of some process not under the investigator's control. As Hurlbert noted, this category is essentially concerned with sampling. His term "mensurative exper-

[1] Manuscript received 6 March 1989; revised 5 December 1989; accepted 14 May 1990.

iment" has the unfortunate connotation of an experimental approach, suggesting that the same analytical approaches can be utilized. In many respects, the same formal mathematical procedures might be employed, but the two approaches differ markedly in the relative strengths of inferences as to cause and effect. If "mensurative experiments" are instead explicitly and properly defined as sampling, then it is feasible to utilize directly a substantial body of existing statistical methodology. A very useful review and discussion of many of these issues as applied to ornithology is that of James and McCulloch (1985), while Kamil (1988) discusses experimental designs in that field.

Techniques for efficient observation have been developed in a number of fields that may be unknown by, or at best very much peripheral to the interests and experience of, most ecologists. For the most part these techniques amount to observation by sampling in space or time. Gains in efficiency are obtained by controlling the allocation of observational efforts in space and time. Biases may be guarded against in the designs, or at least better identified by using a specific design. The main drawback is the weakening of the strong inferences made possible by controlled experimentation. Often, good designs can at least identify the most likely ways that inferences based on observation can go wrong. The objective of this paper is to provide a classification and description of methods for designing environmental field studies. Several of the topics have as yet had little use in ecology thus far, and are consequently treated here in considerable detail.

Because inferences based on experiments without replications or on passive observation of some population or uncontrolled process are so difficult and uncertain, it is doubtful that we can usefully adhere to a particular inferential scheme. Romesburg (1981) described several formal inferential processes in an interesting critique of wildlife research, and proposed that wider use should be made of the hypothetico–deductive method, in which an hypothesis is formulated and then tested by specific experiments that test deductions from the hypothesis.

Unfortunately, natural systems appear to be very "noisy" in the sense of stochastic (chance) fluctuations, and environmental research techniques are subject to substantial "measurement errors," i.e., they rarely measure anything exactly and consistently. In such circumstances it seems desirable to adhere to the more flexible viewpoint proposed by Tukey (1960), in which a long series of successive studies each yield a "decision" (based on statistical tests), but a "conclusion" (a scientific law, perhaps) ultimately depends on reassessment of this whole series of individual results, some of which may then be rejected. Such an outcome is generally unattainable under the rules of strict logic, even if the Type I error (false rejections) of statistical tests in individual hypothesis testing can actually be maintained near the assumed rate.

A Classification of Methods

The essential distinction is whether or not the investigator can arrange for some event to occur at a particular time and place, i.e., a "treatment." The major error made in many contemporary analyses is that of confusing control of events (treatments) with control of the observational process. Circumstances in which "replication" denotes the ability to repeat a treatment should be distinguished from those in which it means taking repeated observations. It is thus important to treat those cases where only observation is possible as sampling schemes. An initial dichotomy then is between (1) conducting a controlled experiment and (2) observing some uncontrolled process by sampling.

Eight different categories serve to differentiate the various possible approaches, and might roughly be arranged on a gradient from strict emphasis on cause-and-effect relationships to nearly pure description. However, a classification on the basis of control exerted by the observer and kinds of events considered may be more useful (Fig. 1). Experiments with and without replication cover the classical experimental approach. "Sampling for modeling" (Eberhardt 1978*a*) stems from research in industrial experimentation, based on a seminal paper by Box and Lucas (1959). The focus in industrial research is on efficient estimation of parameters in a specific non-linear model, such as those used for reaction kinetics in chemistry, that may be used to guide pilot-plant operations on an experimental basis. Under such circumstances, the technique seems appropriately classified under the heading of events controlled by the observer. In other situations it may be useful in observing some uncontrolled process (but the time of initiation should be known at least approximately). Studies of the environmental effects of acid rain provide an example (recent reviews of statistical issues in such studies appeared in *The American Statistician* **39**:243–273).

Intervention analysis pertains largely to the use of time-series methods to study the effect of a distinct perturbation of some kind (Box and Tiao 1975). Examples include the impacts of operating a new power generating station, or changes in atmospheric pollution engendered by opening a new freeway or altering engines to reduce emissions. In many respects such events might be regarded as experiments without treatment replication, but it is useful to distinguish this class by virtue of the inability of the investigator to control events. Also, studies in this class are often retrospective, rather than being designed before the event takes place.

The four categories based on observing a process stemming from circumstances where a distinct perturbation is not evident all depend on sampling, and may be characterized by the way samples are distributed (allocated) over prospective sampling units in the population as a whole. The best known technique is survey

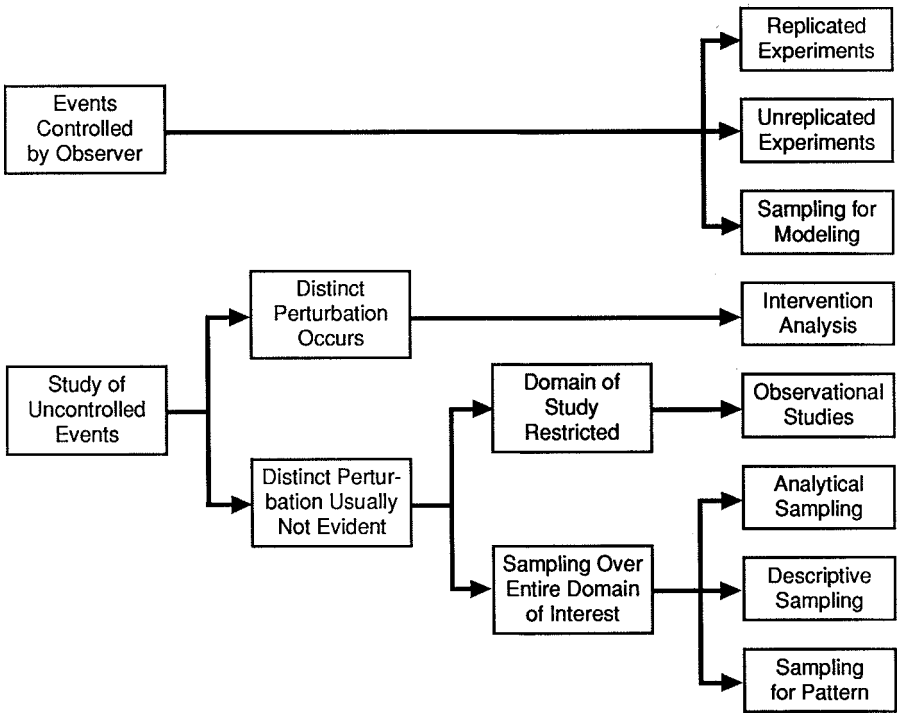

FIG. 1. A classification of the methods considered in this study.

(descriptive) sampling, and is mainly concerned with estimation of means or totals. A dozen or more texts are available, with the most widely used reference in environmental work being that of W. G. Cochran (1977: 4). He distinguished two methods:

> *Sample surveys can be classified broadly into two types—descriptive and analytical. In a descriptive survey the objective is simply to obtain certain information about large groups: for example, the numbers of men, women, and children who view a television program. In an analytical survey, comparisons are made among different subgroups of the population, in order to discover whether differences exist among them and to form or verify hypotheses about the reasons for these differences.*

Observational studies may be distinguished from analytical sampling mainly by the deliberate selection of contrasting portions of populations for study. The name comes from another book by Cochran (1983), who there suggested that analytical surveys have broader and more exploratory objectives. "Sampling for pattern" (Eberhardt 1978*a*) has emerged from the practical need to assess the extent and value of a body of ore, or the volume and extent of an oil field. Research on methods and results tends to be published in the geological literature, so that "geostatistics" may be used in titles of books and papers. In many applications, sampling locations may be uncontrolled by the observer (e.g., drilling of oil wells), and the methodology may be most useful as a way to reduce biases resulting from haphazard sampling.

The balance of this paper will be concerned with further details of the eight methods of Fig. 1 as they relate to existing and potential applications in environmental studies. A synopsis of the methods is:

1) Experiments with replication—strong inferences; preferred approach when feasible;

2) Experiments without replication—cost or circumstances prohibit replication;

3) Sampling for modelling—efficient experimentation for parameter estimation in specified non-linear models;

4) Intervention analysis—retrospective assessments of time-series data;

5) Observational studies—deliberate selection of contrasting groups in lieu of experimentation;

6) Analytical sampling—inferences from sampling over entire population of interest;

7) Descriptive sampling—efficient estimation of means and totals;

8) Sampling for pattern—description of spatial pattern; interpolation to reduce bias from haphazard sampling.

REPLICATED EXPERIMENTS

Strong inferences depend on controlled experiments. Confirmation that two experimental outcomes are indeed different depends on randomization and replication to provide a measure of variability in units treated alike. Since these topics and many related aspects

are thoroughly treated in a wide range of textbooks, we will only attempt to touch on a few issues of special relevance to environmental studies here.

In field settings experiments with replication may come at a high cost. If an investigator has only a few suitable sites (such as ponds), then the use of replicates immediately halves the number of treatments that can be used in a given experiment. If it is established that an entire lake may be needed for an experiment (as advocated by Schindler 1974), then there simply may not be enough suitable lakes, and certainly costs will go up markedly. In other circumstances replication results in a dilution of the field effort, and always poses the temptation to use smaller experimental units than may be wise. In many cases "boundary" effects constitute serious hazards to successful experimentation, so that small experimental units cannot be utilized. For example, Krebs et al. (1973:37) stated that "we conclude that fencing a *Microtus* population destroys the regulatory machinery which normally prevents overgrazing and starvation."

The essential issue is whether an experiment with few replications can be trusted to detect treatment effects that may be considered important. Statisticians approach the problem by defining two types of error relative to a null hypothesis. Type I error is defined as the probability of rejecting a hypothesis of "no effect" when it is in fact true, while Type II error denotes the probability of failing to reject a null hypothesis that is actually false. At present, most ecologists seem to automatically select a Type I error rate of $\alpha = .05$, and to ignore Type II errors. Given a better appreciation of the likely magnitude of actual Type II errors, many researchers would likely use a larger Type I rate, since an increase in the level of Type I error automatically lowers Type II error. The price paid for this change is a higher probability of false rejections of the null hypothesis. The implications of such a change on inferences and on planning experiments need careful investigation.

Various efforts have been made to deal with this kind of problem by quantifying or weighting risks or by assessing relative costs of a wrong decision. A well-known example of power calculations considers the different viewpoints of industry representatives and staff of a regulatory agency in considering impacts of some new construction or new pesticide formulation. Methods for assessment and appropriate modifications of the installation or process may be determined in advance. The main variable left to consider is sample size, which depends on the amount of protection that each party requires against errors damaging to its interests:

a) The manufacturer would rather not have the survey results indicate a significant change when a stipulated degree of change really did not take place (Type I error). If corrective action is not very expensive, industry representatives may be willing to tolerate a fairly high probability of such an error, but if extensive construction changes are involved, it can be expected that they will want this rate to be low, perhaps on the order of $\alpha = .01$.

b) On the other hand, the staff of the regulatory agency would rather not fail to recognize a significant impact when it does occur (Type II error). If only a few replicates are used, the probability of a Type II error will be large, and this cannot (or should not) be tolerated by the agency. Again, the choices of actual rates will depend on likely consequences. If the impacts are not expected to be substantial, a rate of 20% or so may be acceptable, but a prospect of a major impact may lead the regulatory agency to want a very low probability of Type II error.

The two types of error thus provide a prospect for quantitative expression of the two opposing sets of concerns. In purely industrial settings it may be feasible to assign dollar values on both sides, and thus to find an optimal approach in terms of revenues. In environmental settings this approach is seldom possible; hence the need for better data on actual instances and for research on improved approaches. The easy answer—to increase the number of treatment replications—may not be feasible because of the high variability encountered in nature (cf. Eberhardt 1978*b*). Efforts to reduce this variability by various devices may be worthwhile, but need to be carefully considered in terms of the prospect that the variability may in fact be an important component of the ecosystem.

Another aspect of experiments with replication that needs more attention in ecology has to do with the distinction between "fixed-effects" and "random-effects" models in the analysis of variance. As Scheffé (1959) pointed out in Chapter 10, the fixed-effects model is relatively insensitive to violation of the underlying assumptions (normality, homogeneity of variance, etc.) On the other hand, the random-effects model is much more dependent on fulfillment of assumptions. Since these two models apply to rather different classes of experiments, it is important to use this distinction as another dimension in the assessment of sensitivity and inferential trade-offs discussed above. The chief issue is that the fixed-effects model is widely used but frequently not justified for the data at hand. It generally applies to a fixed and finite set of entities, and thus supports only rather limited and specific inferences. Many ecological studies require analyses based on a conceptually indefinitely large population, and thus require a random-effects model. We do not propose that this difficulty can be resolved by more research, but believe that a careful review and further analyses of existing data sets is needed to improve ecologists' general understanding of the inferential differences between these models.

An excellent example of the difficulties in ecological field experimentation was presented by Turk (1978), who posed questions about the analysis of a set of data

on an experimental manipulation of predation. He studied the effect of predation by starfish on other species in an intertidal boulder-field community. On two of six 10 × 10 m areas he added large numbers of starfish, removed all of those located on two areas, and left two unaltered areas as controls. Periodic observations were made on each area for a year before the intervention and for a year afterwards (starfish were removed and added to the "treated" areas throughout the second year). The principal variable measured was the abundance of a snail, *Tegula eisini,* on each of ≈30 boulders in each area. Two responses (Zahl 1978 and Van Belle and Zeisig 1978) noted the severe difficulties of analyzing such data by conventional statistical approaches. Two further commentaries on the same data (Finney 1978, 1982) further emphasized the issues and suggested a serious need for further attention.

One published analysis (Zahl 1978) was based on an analysis of covariance, to adjust for time and number of boulders counted in each plot. Serial correlation was neglected. Another analysis (Van Belle and Zeisig 1978) used logarithms of monthly averages in a split-plot design. A time-series analysis was rejected due to insufficient sample sizes. A third effort (Finney 1978) rejected the first approach as requiring more degrees of freedom than were actually available, and the second as employing "split-plot variation irrelevant to the query," and also as manufacturing degrees of freedom from repeated observations on the same unit. Direct averaging was then used to provide an analysis of variance with five degrees of freedom.

In a later, more-detailed analysis (Finney 1982) the question of serial correlation was assessed in more detail, and the underlying philosophy discussed. Our own unpublished analyses were based on the station-pairs method (Eberhardt 1976) and an alternative approach through the analysis of covariance. Our results suggest, as do the previous analyses, that significant differences exist, but are importantly influenced by inconsistent behavior of counts on the control plots.

We have also reviewed the papers identified in Hurlbert's (1984:Table 3) sample. In our view, the unquestionably important issue of "pseudoreplication" is overshadowed by the very small sample sizes used in ecological field experimentation. One only needs to scan recent volumes of any of the ecological and environmental journals to ascertain that the typical field experiment involves only a few independent sampling units, almost always ≤8, often no more than 6, and frequently the minimum possible for a replicated experiment ($2^2 = 4$). Such sample sizes do not give enough degrees of freedom to obtain useful estimates of error mean squares. Hence it is not feasible to make a reliable determination of the likely power of such an experiment (power tables do not even contain entries for such small samples).

As Hayne (1978:8) pointed out, after reviewing 37 papers dealing with small mammal studies: "where such preliminary examination indicates doubt that the available resources can support an experiment powerful enough to reliably detect a reasonable level of response, then probably the experiment should not be run." Very likely much of Hurlbert's (1984:196) concern about "interspersion" of treatment and control plots may become of secondary importance when experiments sufficiently large to have useful power (small Type II error) are conducted.

Unreplicated Experiments and Intervention Analysis

An essential conceptual distinction between unreplicated experiments and intervention analysis may be whether or not to attempt the study. An unreplicated experiment can presumably always be redesigned to use replicates, given sufficient resources. An "intervention" is not subject to control by the investigator, and thus cannot be replicated in the usual sense of the term. In reality, many ecological field studies can be replicated by straining available resources to the limit. Often this means that other aspects of the study may be short-changed: fewer observations may be made, experimental units may be smaller than is desirable, and various controls may not be included.

Many investigators may simply go ahead and do the experiment without replication. As Hurlbert (1984) showed, statistical analyses are then not provided, or are simply erroneous. Since truly definitive single experiments are very rare in any field of endeavor, progress is actually made through sequences of investigations. Lack of replication may then be most damaging in the long run if chance fluctuations send investigators down the wrong sequential path. This prospect most likely is enhanced by the faulty analyses frequently utilized. That is, research can proceed successfully without any statistical analysis if the scientist can somehow accurately use the outcome of one study in planning further work on the subject at hand. However, many unreplicated experiments are analyzed under the mistaken assumption that subsamples can be regarded as replicates. Such a procedure is likely to result in finding "significant" differences when experiments with replication on the same material would not demonstrate statistically significant differences. This is because any two areas or any two communities can always be shown to be different if one observes them in sufficient detail. (Hurlbert [1984:201] gave a hypothetical example of the consequences in terms of Type I error.) When true replication is used, these differences become part of the "error" term, and thus do not confuse the analysis.

One of the major problems in ecological studies is that a concomitant change in a "response variable" may be caused by some extraneous factor. Such a change may often be due to weather conditions. With this problem in mind, Eberhardt (1976) proposed the use of control areas, also observed before and after the

event in a "station-pairs" or "ratio-method" approach. The basic idea is to employ paired (matched) control and "impacted" stations, for which baseline data are obtained in a pre-operational period. Impacts are evaluated in terms of a shift of the ratio of the paired station data following the intervention.

The basic idea was elaborated at some length in a book by Green (1979). He did not, however, consider matched stations, but instead recommended random location of the individual stations in study areas selected inside and outside the likely zone of impact. A logical but unfortunate consequence of this approach is that the statistical analysis then has to depend on a test of interactions in an analysis of variance. With the exception of a test devised for a particular form of interaction (Tukey's test), this procedure cannot be used without true replications (Green's random location of individual stations amounts simply to subsampling). The difficulties in attempting analyses when unbalanced designs and higher-order interactions are involved have been explored in detail by Thomas et al. (1978). Some useful results on effects of spatio-temporal correlations have been presented by Millard et al. (1985).

In operation, the station-pairs method depends on accumulated evidence that a logarithmic transformation does a reasonable job of meeting most of the assumptions basic to the analysis of variance (Eberhardt 1976, Green 1979, Stewart-Oaten et al. 1986). Such a transformation leads to using the difference between the logarithms of response variables on the paired stations ("impacted" and control stations) as the primary datum for analysis. Careful selection of the pairs can thus approximate the role of "good blocking" in agricultural designs. The usual impact-study data can then be accommodated in a two-way table. Columns (A: main effects) represent station pairs; rows (B: main effects) represent sequential observations over time. The table will normally have two parts; one will represent the pre-impact (baseline) data, and the other the data collected after the intervention occurs. If there are no missing data, the minimal model may thus be a two-way analysis of variance with a single observation per cell.

The exact form of the analysis will necessarily depend on the assumptions that can be justified in given circumstances. In the example above, let y_{ij} represent the logarithm of the ratio of observations on the i^{th} station pair at, say, the j^{th} month. Then we might consider the model:

$$y_{ij} = m + A_i + B_j + G_{ij} + e_{ij},$$

where A_i denotes the main effect due to stations, B_i that for months, and G_{ij} an interaction term. One possible approach is to test for significance between months (B_j), and then use the "contrast" (linear combination) between observations made before and after the intervention as the test of major interest. If the monthly observations can be regarded as independent random variables, one might consider a test based on a "true" error term (rather than using the interaction term as denominator of the F ratio). It is essential to do a good deal of checking on the assumptions underlying any test of significance, since the approximation we propose may not be as "robust" as the usual analysis of variance with replication. Four assumptions are usually postulated as a basis for the analysis of variance. As is well known to statisticians, they are not mutually exclusive. Also, as Scheffé (1959:83, 331–368) has pointed out, case-by-case testing of assumptions can have quite unsatisfactory effects on the inferential procedures associated with an analysis of variance. The preferred procedure is one of establishing the validity of the assumptions (or, more practically, meeting them to a reasonable approximation) in advance by extensive testing over large sets of appropriate data. Such data should be relevant to particular studies to be analyzed in a specific case, i.e., based on many similar experiments because individual experiments seldom contain enough suitable data. However, when several similar experiments are examined, one then must assume that the resulting data are from the same statistical distribution.

The four assumptions for the analysis of variance are:

1) Normality of errors. Based on previous investigations (Eberhardt et al. 1976), we believe that a large fraction of ecological data can be adequately normalized by using a logarithmic transformation. The assumption of a lognormal distribution can, of course, theoretically be rejected for many sets of data on the basis that the observations are discrete counts, rather than continuous in nature, as implied by the lognormal distribution. Although this point is likely inconsequential for most of the data of concern, there will be circumstances where the discrete nature of the data must be taken into account. Thus the use of compound distributions (Johnson and Kotz 1969:Chapter 8) may need to be considered.

2) Equality of variance of errors (homoscedasticity). We have largely been using Scheffé's test for homoscedasticity (Scheffé, 1959:83) for the circumstances we encounter. However Scheffé's test is likely to be unsatisfactory with small samples, so there is further evidence for the need for large-scale studies rather than routine tests. Unfortunately, there are not many appropriate individual sets of ecological data with large samples. Consequently there is a need to infer suitability of a procedure by assessing its behavior over a "large sample of small samples." Possibly such an investigation would be strengthened by use of an approximation to Scheffé's test, developed in earlier research (Eberhardt et al. 1976).

3) Statistical independence of errors. The assumption of statistical independence is sometimes defined as zero correlation among errors. In field experiments,

crop responses on adjacent lots will be more alike than those further away. Observations by the same person at about the same time will also be more similar. These correlations are accounted for by randomization of plots before treatments are applied or by determining the order of investigator observations at random. In environmental studies randomization may be difficult. For the station-pairs method discussed above in this section, it was suggested that monthly observations might be regarded as independent random variables. The resulting monthly observations before and after the intervention can be tested for a statistically significant serial correlation (see Gilbert [1987], sections 4.5.2 and 11.13, for methods of computing and testing serial correlations). Thus, this important assumption may be evaluated prior to the analysis of variance.

4) Additivity. The problem of dealing with interactions is crucial to the statistical evaluation of environmental and ecological studies. We believe that testing for additivity is one of the more difficult problems, in that there are so few "balanced" sets of data available for the circumstances that are essential for a workable design. Many data sets must be accumulated and assessed before much can be predicted about the likely outcomes of a study of additivity. We note that our earlier suggestion that a logarithmic transformation will most likely normalize many ecological data sets implies a multiplicative structure in the original data. In discussing the "cutting edges in biometry" Federer (1984) has suggested that, concerning model construction and selection, "One good paper (Box and Cox 1964) has appeared."

Repeated measures

In any moderately complex study a common practice is to make observations over time. When the study is completed there then arises the question of what do do with these data. In classical agricultural experimentation the focus usually is on the effect of treatments on harvests, so that statistical analyses could be directed to final yields, and any additional measurements made during the course of the experiment were relegated to a subsidiary status. When the end point is not necessarily of primary concern, methods of analysis may be less certain, and possibly controversial. Studies of this kind have been variously labelled as "split-plot" designs, and more recently have begun to be called "repeated measures" designs. Aitken (1981) discussed a simple example, while Ware (1981) described a method called "tracking" and its connection with growth-curve models. Gurevitch and Chester (1986) proposed an approach for a class of ecological experiments.

Applications of repeated-measures techniques can be conducted with replications, and thus might logically fall in the previous section of this report. An important caveat is that the degrees of freedom "experimental error" may again be very few in number, again raising the issue of whether such experiments will have useful statistical power. The topic is discussed here mainly because the pattern of time trends in "treated" and "control" units may be very useful in more-or-less qualitative evaluations of possible differences. Cook and Campbell (1979) describe a variety of applications and some quantitative tools that may be helpful in such analyses.

We believe Hurlbert's (1984:201) discussion of "replicability" as a red herring needs to be reconsidered in the light of such comparisons. In discussions with some of the scientists whose use of the term was criticized by Hurlbert, it appeared that they had in mind the need to demonstrate that replicate control microcosms (and similar entities) would in fact follow nearly the same pattern over time. Such a demonstration would seem to be an important precondition for evaluations of differences in time trends between treated and untreated units in an actual experiment.

Box and Tiao (1975) proposed the use of time-series methods on observational data that are usually serially dependent and often non-stationary. A long series of such observations would be made both before and after an event occurs (an "intervention" takes place). These data would then be fitted to a time-series model. A statistical test may be applied to determine whether a significant difference in adjusted level of the observations exists before and after the intervention. A series of exponential weights that diminish with time before and after the intervention would be used in the test. Unfortunately, the long sequences of data needed for time-series analyses are seldom obtained in ecological contexts.

Sampling for Modelling

Box and Lucas (1959) considered a general model where a response variable is a known function of k variables, and of p parameters. In its general form the Box and Lucas scheme permits study of a multivariate system, with variables like time, temperature, and concentration controlling the level of some product through a known functional relationship with fixed (but unknown) constants (parameters). An example is the experimental operation of a pilot plant in which a chemical process is being used to produce some product. In environmental work the Box and Lucas scheme is applicable to nonlinear distance from source models (e.g., chemical concentrations in small mammals, soil, or air as a function of distance from a stack) or concentrations obtained at a single location as a function of time since cessation of emissions or spills of some substance. Thus, relevant fields of application include radioecology, ecotoxicology, and environmental chemistry. When mechanistic models cannot be postulated, ecologists can use deterministic time series (see *Repeated measures* subsection, above), multivariate regression, or response-surface models (see *Response-surface models,* below). As an example in a specific environmental sampling context, we are often concerned with models for

such things as concentrations of radionuclides or hazardous chemicals as a function of time or depth of soil. An appropriate model will then be:

$$y = f(t; B_1, B_2, \ldots B_p),$$

where t denotes time (but may alternatively represent depth or distance from a source), and B_1 to B_p represent p parameters in the model.

For a specific example, consider a two-compartment model of radionuclide release from a source and its subsequent retention in some receiving system, which might be a defined region of soil, or some element of the biota. Assume that the input to this system decays away exponentially with time, and that the receiving system loses a constant fraction of its contents per unit time (through decay and/or "washout" or excretion). The appropriate equation for quantity in the receiving system at time t is then:

$$y(t) = \left[\frac{B_1 y_0}{(B_1 - B_2)}\right][e^{-B_2 t} - e^{-B_1 t}]$$

We note that this equation is also essentially a "reaction-rate" model, so that the analogy to chemical kinetics continues to be relevant. In the equation, y_0 denotes the initial quantity in the "donor" compartment, which is transferred to the receiving compartment at rate B_1. The receiving compartment loses the material at rate B_2. In environmental circumstances it is usually necessary to regard y_0 as a third parameter. However, in reaction kinetics $y(t)$ is usually expressed as a fraction of the known initial quantity of reactant (y_0); thus only two parameters are shown in the equation and estimated in industrial research. Ecological applications of similar models have been given by Eberhardt (1978*a*) and Blau (1975).

The objective of the study is to select the optimal times to observe (sample) the process in order to estimate the parameters for the equation. Normally there will be a defined time period in which the observations are to be made, either inherent in the conditions under which the study is conducted, or determined by the observer. The time interval can conveniently be represented by $t = t_{\min}$ and $t = t_{\max}$, with $t_{\min}$ frequently (but not necessarily) the time at which the process is initiated. The optimal times for observing the process are obtained in the Box and Lucas (1959) approach essentially by minimizing a "confidence envelope" about the estimated parameters, in much the same sense as one considers a confidence ellipse for the parameters of an ordinary regression model. A number of complications are introduced by the fact that most of the models of interest are non-linear and by the limited number of optimum sampling points dictated by the theory (one point for each parameter). We agree with the authors of several papers discussed below that additional sampling points are needed.

In the Box and Lucas (1959) scheme the actual process for selecting optimal times calls for finding partial derivatives (f_{ru}) of the function, f (the model), with respect to the r^{th} parameter and the u^{th} sampling time. Suppose that there are to be N sampling times. We then represent the partial derivatives in an $N \times p$ matrix, $\mathbf{F}$ (N sampling times for each of p parameters). The optimal sampling times are those that maximize the determinant of the matrix product $\mathbf{F}'\mathbf{F}$ (i.e., the transpose of $\mathbf{F}$ multiplied times $\mathbf{F}$), or, equivalently, minimize the determinant of the inverse of the same matrix product. To obtain a numerical answer, it is of course necessary to specify approximate parameter values.

An initial reaction to the Box and Lucas scheme is likely to be a comment that requiring advance knowledge of the parameters that one is seeking to estimate is not sensible. Two points thus need to be emphasized. One is that the models to be considered are all non-linear in the parameters. Realistic study of such models requires some advance knowledge of the parameter values to be expected. Otherwise, the rapid changes associated with non-linearity are likely to leave the observer with virtually no useful results. Unless there is some advance knowledge of the system's behavior, and hence approximate estimates of parameter values, the action may either be virtually completed or not yet started at the time of sampling.

The second saving feature is that the curves denoting optima are fairly flat-topped, so that an approximate advance knowledge of relevant parameters suffices to produce useful results. Finding the optimal sampling times in an actual problem can be a rather difficult analytical exercise, and frequently requires computerized searches. Given normal, additive errors, an exact solution is available for a linear model. Non-linear models necessarily contain the parameter values in partial derivatives; thus only approximate solutions are possible. The optimum design is to select a set of sampling points that minimizes the determinant of $(\mathbf{F}'\mathbf{F})^{-1}$, the inverse of the matrix product.This provides an approximate minimum joint confidence interval (confidence ellipsoid) for the set of parameter estimates.

Errors of measurement

Box and Lucas (1959) considered only additive normal errors, whereas the distributions encountered in environmental studies are usually markedly skewed, and may be assumed to be approximately lognormal. It is very important to consider the effects of this difference in searching for optimal sampling times. For a simple illustration, consider a simple exponential model, the "decay law":

$$y(t) = y_0 \exp(-Bt) + e_i$$

where e_i now denotes errors of measurement, and is assumed to be an additive, normally distributed random variable, as assumed by Box and Lucas (1959: 85). The optimum sampling times are then at $t_{\min}$ and

at a time when ≈37% of the initial value remains (i.e., ≈63% of the substance has decayed), as shown by Box and Lucas (1959:85).

If we now suppose that the errors are lognormally distributed and multiplicative, the usual procedure will be to transform the observations by taking logarithms. This gives a simple linear regression model for which the optimum sampling times are known to be t_{min} and t_{max}. Clearly the two assumptions about error structure give different results. Most of the available literature considers only additive normal errors, so that it is necessary to reconsider and revise the results and recommendations available there for use in environmental studies. It is also necessary to consider some further complications in practice. In the second case (linearized) model, t_{max} cannot be too large or the measurement errors will be excessive, due to decay or dilution of the substance being studied.

Additive and multiplicative parameters

One of the rather surprising outcomes of the Box and Lucas (1959) scheme is that the number of optimal sampling times (N) is equal to the number of parameters to be estimated (p). This means that the $N \times p$ matrix partial derivatives is actually a square $p \times p$ matrix. The search for optimal sampling times depends on maximizing the absolute value of the determinant of this square matrix. Under this condition it is fairly simple to show that the effects of "additive" parameters are such that an advance estimate of their values is not necessary. Hence, in a model such as:

$$y(t) = B_1 + f(t; B_2, B_3)$$

only the estimates of B_2 and B_3 are required in searching for optimum sampling times. Hence when a logarithmic transformation to achieve additive errors is used, the model is changed. The usual result is that some parameters become additive, and the search for optimal sampling times is often simplified.

Discriminating among models

Although the fact that the optimal set of sampling times equals the number of parameters is convenient in many respects, few practical-minded investigators will want to settle for such a limited number of sampling times. There are several reasons for such a viewpoint. Perhaps the most important is that we seldom can be absolutely certain of the correct model for the phenomenon under study (usually this limitation is more important in environmental studies then it might be in chemical kinetics). Another related aspect is that a wider set of points may be more useful in testing the fit of the model and in estimating variances. The main recommendation evolving from relevant research over the last two decades has been to include some sampling points determined as optimal for an alternative model, along with other points chosen to provide an efficient test for discriminating among alternative models. Further research concerned with models appropriate for ecological and environmental contexts will need to consider this feature along with those of the preceding two subsections.

Response-surface models

A good deal of statistical work has been done on response-surface models. Reviews by Hill and Hunter (1966), Mead and Pike (1975), and Myers et al. (1989) summarize some of this work, describe response-surface methodology, and give examples of applications. Recent books by Khuri and Cornell (1987) and Box and Draper (1987) may also be consulted. The book by Khuri and Cornell is written for the more statistically sophisticated reader. Various aspects of those studies may be of interest here. It is important to note, however, that response-surface models are most useful in searching for a maximum or a minimum. The focus is usually on finding an optimum set of treatment values (usually based on fitting polynomial models) to give maximum yield from some process. In the present context, a response-surface model is used to find the set of times that locates a maximum of some curve (or perhaps some other feature). The Box and Lucas (1959) methodology, on the other hand, seeks to obtain optimal estimates of parameters in the model. Thus the objectives are not the same. We believe that response-surface methods may be useful in sampling for pattern, in common with trend-surface methods.

References on sampling for modelling

Unfortunately there are few detailed sources on sampling for modelling. We thus include here a selection of published papers. Two textbooks (Bard 1974, Beck and Arnold 1977) have fairly extensive discussions on design of experiments with nonlinear models. Chapter 8 of Beck and Arnold (1977) is particularly useful because it provides an accounting of a variety of criteria for optimal experiments. The focus is, however, chiefly on engineering applications. In a biological context, the book edited by Endrenyi (1981) is especially important due to the emphasis on kinetic models. Brief treatments of a range of relevant issues in modelling and model-fitting are provided, including a number of applications of the methodology described here. Another text, by Box et al. (1978) contains somewhat less material on non-linear models than those listed above, but is worth examining because a philosophy and approach to experimentation and modelling is discussed in some detail. Because the authors of this book have been extensively involved over many years in developing and extending the Box and Lucas (1959) scheme, their general views are valuable in applications to a new field, as proposed here.

Some references to the journal literature relative to sampling from non-linear processes provide an impression of the material available. A warning may be

in order here. Virtually all of the literature that we have examined is directed to experimental studies, mostly in industrial research. The reader who expects to find references to sampling applications will thus be disappointed. We know of only a few ecological or environmental samples of use of the methodology (cf. Blau 1975, Eberhardt 1978*a*). A few related studies in pharmacology were reported in Endrenyi (1981).

As noted earlier, the pioneering paper in this area was that of Box and Lucas (1959). The next most important reference is likely to be the review by Cochran (1973). He tracted the subject of experiments for nonlinear functions from some investigations by R. A. Fisher up to the 1970s, by which time much of the pioneering work had been accomplished, and noted that research had largely been concentrated in two areas: (1) optimal estimation of the parameters, under an asymptotic criterion, and (2) tests of the adequacy of a model and discrimination among models, plus model building. He reviewed these areas briefly, providing a useful guide to the preceding work.

One of the apparent limitations of the Box and Lucas scheme is the outcome that the number of sampling times is equal to the number of parameters. This point was examined in some detail by Atkinson and Hunter (1968), who confirmed its validity and recommended that replications of the same set of points be used. They illustrated their recommendation with the model used by Box and Lucas (1959) and some others. Box (1970) also considered the replication aspect, and reported results of some further experience with the scheme. Box (1971) provided an approach to an important practical problem, i.e., what to do when the experimenter is mainly interested in a subset of parameters. Hill and Hunter (1974) also considered this question, and provided a design criterion for experiments concerned mainly with a subset of the parameters.

The limited scope of optimal sampling points provided by the Box and Lucas scheme can be extended by considering alternative models for the phenomenon under study, so that two more or sets of design points can be produced. Hill, et al. (1968) approached this prospect by a sequential process in which model discrimination was first emphasized, with a gradual shift to emphasis on estimating parameters for the "best" model as determined by initial experimentation. A good deal of attention has naturally gone into the question of discriminating among candidate models, due to its general importance in scientific and engineering experimentation. Hunter and Reiner (1965) proposed sequential selection of experimental points at which two models differ the most. Box and Hill (1967) considered sets of conditions for the same purpose. Atkinson (1969) focused on significant differences in fits of the candidate models to the observations, and later (Atkinson 1972) studied the use of extensions of a given model to test fits. These and many other references on model building provide valuable ideas for developing mechanisms to choose sets of points extending the basic Box and Lucas scheme.

Analytical Sampling and Observational Studies

Experimentation on intact, functioning ecosystems is not really practicable in the classical sense of experimental design. Instead, parts of systems or "model ecosystems," such as microcosms of various kinds, are used. Most of the inferences about ecosystems based on this kind of approach are perhaps better described as conjectures. Any real confirmation will depend on somehow combining the experimental data with observations on intact ecosystems. An area of basic research in ecology that is, so far as we can tell, virtually unexplored in any formal sense is the development of approaches to observation (by sampling) of some phenomenon with experimentation on limited parts of the system in which the process operates.

In reality, most progress in ecological and environmental research has been made by combining field observations with controlled experimentation. However, the observational stages of this process are not designed in any rigorous way. Many ecologists now appreciate the need for a carefully laid out, formal design for experimentation, but few carry that concept to the logical conclusion, which is that concurrent field investigations should also be conducted in the framework of a sampling design. We believe that this may be one of the most important areas for further research. To deal realistically with ecosystems, one should put a major emphasis on the observational process, and design that sampling effort at least as carefully as the experimentation suggested by assessments of the intact ecosystem. We believe that this approach may provide a more suitable basis for evaluating one of the uncertainties in current ecological dogma, namely, that we lack suitable definitions for ecosystems. One commonly used definition proposes that an ecosystem consists of a community (or perhaps several communities) and the corresponding abiotic environment.

When one attempts to examine actual communities, however, it soon becomes evident that no two examples are ever quite alike. Many small, but significant, differences between similar communities may be due to chance and to the vagaries of weather, climate, and substrate. There is thus a possible close analogy with the statistical concept of a superpopulation, in which some underlying process or processes serve to generate the particular population being observed, and theoretically could generate an indefinitely large array of similar populations on the same substrate. One might thus suppose that the abiotic environment, along with weather and climate, could conceptually generate a superpopulation of communities.

A theoretical framework for identifying the structure within which superpopulations of communities could be generated might well be that supplied by application

of hierarchy theory to ecology as described by, for example, O'Neill (1988, 1989). A less abstract and complex example is the set of populations that can be generated by Monte Carlo simulations of simple birth and death processes applied to some initial population. For example, Knight and Eberhardt (1984:Fig. 1) generated a sequence of 10 000 annual populations of the Yellowstone grizzly bear (*Ursus arctos*) over 30 yr, of which only one trajectory over time might actually be observed for the Yellowstone area.

The practical significance of the superpopulation concept is that it is essentially a basis for much of analytical sampling. Hence our thesis that experiments should be designed to fit in a matrix of sampling studies of intact ecosystems appears to fit into an established theoretical framework. If further elaboration appears to confirm this supposition, it should be possible to begin to construct some useful guidelines. One evident avenue is that of providing a framework for "perturbation" experiments, which seem to be increasingly popular in some areas of ecology.

An example of such perturbation is the predator-manipulation experiment conducted by Turk (1978) and discussed above (see *Replicated experiments*). The outcomes of such an experiment are extremely interesting in terms of implications about the way communities are regulated. However, they can only improve our knowledge of communities if placed in a framework of knowledge about actual communities. This information must come from sampling studies of such communities, which presumably will produce data on the levels and dynamics of the actual abundance of predator and prey across many similar communities.

Some parallels to the situation discussed here exist in agricultural research. Many decades of field-plot research have been plagued by the knowledge that no one plot experiment can be typical of the full range of agricultural use of a given crop. There have thus been a number of investigations of the issues and problems in combining experiments done at a number of locations over several years. While it may be useful to examine this literature for clues in designing ecological research, we note that the goals may be quite different—understanding ecological systems as contrasted with maximizing a yield. More recent agricultural work on mixtures of crops (cf. Federer 1984) may come closer to ecological research.

The same kinds of arguments can be advanced in response to the practical requirements of pollution studies (Eberhardt 1975*a*). Due to the large scale of exposure and the complex mixtures of polluting substances, such studies essentially must depend on observations, by sampling, of intact, real-world systems.

References on analytical sampling

Analytical sampling has not received much attention in the statistical literature, probably because it falls between the two fields of statistical analysis and survey sampling. Some of the sampling texts do, however, have short sections on analytical sampling. Cochran (1977) provided several brief comments on specific features of sample surveys conducted for analytical purposes. The most extensive is a short section on strata as domains of study, i.e., surveys designed to make comparisons among strata selected for analytical purposes. He supplied the formulae for allocating samples to strata in this case.

In his pioneering book on sampling, Deming (1950) made essentially the same distinction between descriptive and analytical sampling, and provided many examples. He also made the interesting point that, for analytical purposes, even a complete tally of all of the units in particular domains of study will not negate the use of statistical methods for analysis, whereas in a descriptive study, a complete tally eliminates the need for any further analysis. The reason, of course, is that the analytical survey is conducted with the notion of trying to infer from observations why the population takes on its particular characteristics. Consequently, a "superpopulation" idea dominates for analytical purposes. We suppose that the particular population under study arises from an infinite (or nearly so) array of populations (the superpopulation) that might have resulted from the processes creating the population actually observed. This is the reason that the finite population correction, used in descriptive sampling, is dropped for analytical purposes.

In a later paper Deming (1975:147) attempted to distinguish between "enumerative studies" and "analytical studies." He stated that the "aim of a statistical study in an enumerative problem is descriptive" and classified analytic surveys as those "in which action will be taken on the process or cause-system that produced the frame studied, the aim being to improve practice in the future." His classification thus can be associated with the primary dichotomy used here (Fig. 1) between the study of uncontrolled events and of events controlled by the observer. His focus is almost exclusively on situations where an action that can be taken changes the process being studied (e.g., manufacturing, public health) so that the inferential problem is much simpler than in ecology, but the paper provides many valuable comments on the use of statistics in the two classes of studies. An earlier, and broader, classification is that of Wold (1956), who described an initial dichotomy of purpose as being between "description" and "explanation," and then further subdivided these categories into non-experimental and experimental areas.

Jessen (1978) also briefly discussed analytical surveys. For design purposes he suggested a classification by factors and levels, similar to those used for experimental designs. One may thus attempt to classify the population being sampled by certain factors that can be subdivided into levels, and then subsample within these categories.

Kish (1965:Chapter 14) provided a good discussion of issues of inference in survey sampling. He suggested that four classes of variables may be present in a given situation: (1) explanatory variables (also known as the experimental variables), (2) some extraneous variables that can be controlled, either in selection of units or by an auxiliary measurement, (3) other extraneous variables that are uncontrolled and confounded with the explanatory variables, and (4) the remaining extraneous, uncontrolled variables that can be treated as randomized errors. Often, the difference between an experiment and a sample survey depends largely on (3), since those variables can usually be forced into class (4) by the randomization employed in experimental designs. Kish also noted that much of the statistical analysis of experiments is limited to inferences about the population encompassed by the experiment, which may be a much smaller population than that for which the investigator wishes to make inferences. There is thus an argument for somehow integrating the experimental and the sampling approaches into an overall course of investigation. As previously noted, we believe that this may be a valuable approach for consideration in ecological research. Experimental field studies ought to be designed as part of a framework that encompasses extensive studies, by sampling, of the system of overall concern.

Yates (1981:Chapter 6) provided the most detailed assessment of analytical surveys of any of the references that we have found. He remarked, however (p. 321), "It must be emphasized . . . that the chapter is not intended to be an exhaustive treatise on the analysis of investigational surveys: this would require much more space than is available here." Yates effectively discussed the distinction between an experiment and a survey. His chapter is primarily concerned with the use of the classical statistical analytical procedures on survey data from sociological or agricultural surveys, and includes several detailed examples.

The difference between analytical sampling and descriptive sampling, from an ecological point of view, can be clarified through a common example. In many situations an important aspect of field research lies in estimating the abundance of one or more species of plants or animals in several subregions (strata) of a large area. The usual sample-survey methods provide estimates, with standard errors, of abundance, thus "describing" the populations of concern. In an analytical survey, one attempts to determine whether there are significant differences in local abundance, and to ascertain why such differences exist.

Observational studies

Cochran (1972) pointed out that analytical surveys are used to sample a population of interest (the "target" population) in order to conduct statistical analyses of the relations between variables of interest. Observational studies are narrower in scope, and seek to compare the effects of some processes by contrasting selected groups of individuals, or other entities, subjected to different levels (including zero) of the processes. Selecting the groups nearly always leads to dealing with a more limited population (the "sampled" population). All of the more comprehensive references to approaches and techniques are largely concerned with human populations.

Observational studies may thus be described and analyzed in much the same way as experiments, but they lack the element of control by the investigator as to which individual gets what treatment. Once a population of exposed individuals is identified, various elements of control can be used to select the sample for study, but the actual assignment of a treatment is not feasible. Very often an analytical survey will serve to turn up apparent relationships that may then be further examined with an observational study. The early studies of the use of tobacco ultimately led to more narrowly defined studies of the role of smoking in development of cancer and heart disease, for example. In the latter case, obesity and high blood pressure were also important variables suggested by the initial data, and then further investigated in observational studies.

In his book on observational studies, Cochran (1983) suggested three main methods for controlling extraneous effects in an observational study. One is simply refinements in technique that seek to reduce the measurement error in the variable(s) of main interest. In experimental work, this often takes the form of closely controlled ambient conditions (light, temperature, humidity, etc.) and the use of uniform experimental material (e.g., inbred strains of mice or rats). In both situations this kind of control can improve the comparisons but narrow the applicability of the results, since the study conditions may not resemble the "real world" or "target population."

A second approach to control is through blocking and matching, in which we use separate individual replications to keep the extraneous variables as constant as possible within a replication. Thus in agricultural experiments a block may be a relatively small plot with each treatment tested on a portion of this "block." In observational studies an effort may be made to match an individual (or other entity) exposed to the process under study with another unexposed individual in as many of the extraneous variables that may influence the outcomes as possible. A tedious selection process may thus be required, involving locating and examining a large pool of possible "control" individuals.

A third method is to attempt to "remove" such extraneous influences during the statistical analysis of the data. The best known such technique is the analysis of covariance, where regressions on the variables to be controlled are used for an adjustment. Many variants of these basic approaches can be used, with the major examples described quite succinctly by Cochran (1983).

More details and a variety of designs for social and educational research were given by Campbell and Stanley (1963) and Cook and Campbell (1979). Although none of the references pertains directly to ecology or environmental studies, most of the principles and many of the designs should nonetheless be quite useful in planning observational studies in these areas.

An important general principle suggested by Cochran (1972, 1983) is that any effort at analytical or observational sampling is likely to be subject to biases of one sort or another. Experience and judgement become very important in evaluating the prospects for bias, and in attempting to design ways to detect and minimize the effects of bias. It may be particularly helpful if the variable of main interest has a graduated effect over different classifications of auxiliary or extraneous variables. Valuable suggestions along these lines were given by Cochran (1983) and the other references cited above.

Survey Sampling

Since there is a long list of excellent references on survey or descriptive sampling, this section is limited to a brief accounting of the main methods. An easily read, non-technical account is that of Slonim (1960), while non-technical but somewhat more sophisticated views of the main methods are available in books by Stuart (1976) and Williams (1978).

The basic approach in descriptive sampling is that of simple random sampling. Numbers are assigned to each potential sampling unit in the population, and the sample is drawn by utilizing a table of random numbers. One does not have to write down all of the numbers, but can use any convenient scheme to assure that there is a one-to-one correspondence between number and sampling unit. If measurements on the sampling units are unlikely to change over the time period in which the survey is conducted, sampling without replacement is ordinarily utilized, i.e., if a given sampling unit is drawn again it is not surveyed a second time. When large populations are dealt with, the distinction between sampling with or without replacement is not particularly important in terms of estimation and in calculating confidence limits. However, two features of survey sampling should be more widely considered in ecological work. When the population is small, use of a finite population correction factor may be important in calculating variances. Also, if measurements may change with time, use of sampling with replacement needs to be considered.

In the field, random sampling may be much less convenient than systematic sampling, i.e., sampling on a grid or at regular intervals. A somewhat schizoid attitude exists with regard to the use of systematic sampling in survey sampling references. Cochran (1977: 212) remarked that there are circumstances in which the variance of a systematic sample may increase as a larger sample taken, which is, he noted, ". . . a startling departure from good behavior." He provided an excellent brief account of situations in which systematic sampling might be recommended. Hansen et al. (1953: 504) recommended that, ". . . it is safe to use systematic sampling only when one is sufficiently acquainted with the data to be able to demonstrate that periodicities do not exist, or that the interval between the elements of the sample is not a multiple or a submultiple of the period. The risk with systematic sampling from a population with periodicities is particularly serious because the sample itself may give no indication of the periodicity." The danger, of course, is that such sampling from a population showing periodic variation will not be representative of the population as a whole. Yates (1981:43) gave very much the same warning, while Kish (1965:120) suggested, "Because ordinarily we cannot be absolutely certain of having avoided all danger, some statisticians prefer to avoid systematic sampling altogether." But he also remarked that, "In most practical situations after investigating, we can dismiss the dangers both of a monotonic trend and of a periodic fluctuation coinciding with the selection interval."

An important question is whether or not adjacent units of the population are strongly correlated with respect to the variable being measured. If they are, then sampling units that fall close together simply repeat the same information, and a means to avoid such circumstances is highly desirable. The issue becomes much more clear-cut when the main objective is sampling for pattern, as discussed in the following section.

More complex sampling schemes are ordinarily used in descriptive sampling, either to increase efficiency (i.e., obtain narrower confidence limits at the same cost) or to deal with inadequacies in the sampling "frame." In some cases it is not feasible to assign a number to each element of the population, and the sampling scheme needs to be designed so that probabilities of selection are nonetheless known. A simple example would be using households, rather than individuals, as a sampling unit, inasmuch as lists of individuals may not be available, but dwelling places can nonetheless be readily sampled. The analogy with litters or packs of animals is clear. Frequently, this approach goes through several stages, since a population of city blocks or hectares in rural or wildlife areas is easily defined and thus would provide a convenient sampling frame. The sampling process may thus be done in stages, with square miles as primary units, litters or packs (or households) as secondary units, and individuals as the actual units on which measurements are obtained. Such subsampling schemes can be quite complex, and may require an elaborate analysis depending on the design utilized.

A less-complicated approach for dealing with situations where natural sampling units come in groups is cluster sampling. If each of some class of convenient sampling unit can be subdivided into the same number

of smaller units (elements), one can draw a sample of the larger (primary) units, subsample each, estimate an independent total for each of the primary units, and then conduct the balance of the analysis as though no subsampling occurred. Specific analytical attention to elements of the subsample is needed only if one wishes to investigate efficiency of the subsampling rates (i.e., what is the best subsampling fraction?).

In other circumstances, it may be desirable or necessary to enumerate all of the elements of each of the sampled clusters, or various subsampling fractions may be used, depending on characteristics of the cluster. A useful, and interesting, scheme for dealing with clusters is sampling with probability proportional to size (PPS sampling). An important example of this would be in dealing with a population of areas having quite different sizes. One can take a simple random sample of points to locate units (if a randomly located point falls in a unit, then that unit is included in the sample), thus automatically yielding selection probabilities proportional to size. Sometimes such a selection process is accomplished inadvertently, a process dubbed "size-biased sampling." Patil and Rao (1978) gave some examples, and Eberhardt (1978*c*) proposed such a model for certain transect methods.

In ecological work the most commonly used refinement in sampling techniques is stratified sampling. Usually the investigator has some a priori knowledge of the difference between various parts of the population to be sampled. He can then classify each unit of the population into one of several strata. In most cases 3–6 strata are used, and one ordinarily does not have to number each unit individually. Depending on how successfully the units are sorted into homogeneous strata, this scheme can be very efficient indeed. A frequent finding in environmental sampling is that the bell-shaped normal distribution does not apply, and the frequency distributions are often skewed, following approximately a lognormal model. One then tends to find that the coefficient of variation is relatively constant (Eberhardt 1978*b* gave examples). Under such circumstances, the problem of efficiently allocating samples to strata has a very simple and convenient solution, described by Eberhardt et al. (1976).

When units of the population can be assigned a value that is correlated with the variable of interest, this auxiliary variable may be used in ratio sampling. One of the better known examples in survey sampling depends on using data from the most recent decennial census to provide a current estimate of some characteristic. Thus if one wanted to know the current population of some class of metropolitan areas in, say, 1984, enumerating a relatively small sample of such areas can be used to create a ratio of current population to 1980 population for those same areas. Multiplying this ratio times the known 1980 total for the selected class of areas gives what is usually a very accurate estimate of current population. A simple example in ecology arises when a sample of study areas of different sizes has been enumerated for some characteristic. A simple average will give a biased estimate of the total, but a nearly unbiased estimate can be obtained by multiplying the ratio of counts to sample areas times the total area from which the samples were drawn.

Unfortunately, in environmental studies one seldom has a known total for the auxiliary variable over the entire population. If, however, the auxiliary variable can be measured quite inexpensively (compared to the variable of primary concern), then an approach known as double sampling may be worthwhile. One can take a large sample of measurements of the auxiliary variable, and a much smaller subsample of this group would then be used to measure the primary variable. Such an approach may be very valuable when some expensive laboratory determination has to be made on, say, vegetation. Often a simple field measurement will be sufficiently well correlated with the lab assay to permit use of double sampling. A substantial potential area of replication rests on the fact that many metabolic processes in animals are proportional to a fractional power of body mass. Hence if we wish to assay fish for some contaminant, the logarithm of mass is a suitable auxiliary variable (Eberhardt 1975*b*).

One difficulty with double sampling is that the method depends on an approximation that is quite satisfactory under the appropriate conditions (cf. Cochran 1977:Chapter 12), but in ecological work it often will not be feasible to obtain sufficiently large samples to meet these criteria. However, some "monte carlo" simulations (Eberhardt and Simmons 1987) indicated that these conditions can be relaxed considerably, so that quite small samples of the primary variable may yield very useful results. Hence we believe that the technique offers a way to combat one of the major shortcomings in contemporary field work—inadequate sampling.

Sampling for Pattern

There are many ways to summarize a spatial pattern. Their suitability depends on the purpose(s) of the research and the assumptions about the process(es) underlying the particular pattern under study. In the environmental sciences we may be required to deal with a single specific area, and be given data on a particular set of samples previously taken over that area. The objective would be to provide some kind of a summary of the pattern indicated by the samples, usually as a map. The issues that need to be considered depend on the uses to be made of the map.

When the data are simply those that happen to be available, it is likely that there will be areas with very few points and others with clusters of points. An average across all of the data may give a seriously biased result, i.e., it may differ substantially from the true (but unknown) mean value for the entire region of interest. In this case, an alternative would be to use the data to fit some kind of surface that represents the values as-

sociated with the available data points, to interpolate between data points. A mean value for this surface may then provide a better estimate of the true value. Also, such a map will provide the user with a much better "feel" for the nature of the dispersion of a particular variable.

In the more satisfactory circumstances where sampling locations can be selected according to a design, it is possible and usually desirable to avoid clustering. An important decision may then be whether to resort to systematic sampling. An intermediate situation also needs to be considered, in which a set of samples has already been taken in an haphazard fashion, but can be supplemented by further sampling. In this case, some of the newer methods, broadly described as "kriging," may be especially useful (Journel and Huijbregts 1978, Clark 1982).

There are various ways to construct a map from a given set of data:

1. *Trend surfaces.*—These are usually generated by a two-dimensional process analogous to curve-fitting by least squares. The most popular underlying model is that of a polynomial, and the major shortcoming of these polynomial fits is a tendency for strange behaviors at the edges of the fitted surfaces. Trend surfaces are most useful to give a broad general impression of the surface. The resulting fits may be used as components of the more elaborate stochastic process model (kriging) described below. Care in using polynomials is essential, since the higher-degree models can yield some highly pathological results.

2. *Moving averages.*—An alternative approach is just to smooth the observations by using moving averages. These are constructed by replacing a given value by the average containing it and some neighboring points, starting at one end and moving through the series of points. Such schemes may not work well in the presence of clusters of points, since the values in a cluster may exert too great an influence on points in sparsely populated surrounding areas. This shortcoming may be attacked by using some kind of weighting scheme, so that the influence of a point depends on its distance from the location being mapped. Another useful refinement takes into account only the nearest points in a given quadrant.

3. *Spatial splines.*—In sampling along a line, curves may be fitted by using smooth curves to interpolate in the intervals between data points. Various constraints can be put on these interpolatory curves near the data points to prevent erratic behaviors; e.g., the curves may be required to pass through the data points and should do so smoothly. In sampling a two-dimensional area, the problem is much more difficult because the convenient ordering of the points along a line is no longer available. Therefore we do not consider splines further here.

4. *Stochastic process models.*—This approach is based on the assumption that the "true" surface being studied is only one example of an infinite number of possible outcomes from some underlying process. This permits considering two components. One is an underlying large-scale trend that would be present in any realization of the process. The other is a localized or small-scale fluctuation. It should be noted that trends and fluctuations often are only a matter of scale. If one considers small areas within a larger region, the fluctuations observed on a larger area may look like "trends" with smaller scale "fluctuations" of their own in the subarea. A relevant model is

$$Z(x, y) = m(x, y) + e(x, y)$$

where (x, y) denote spatial coordinates of a sampling point, $Z(x, y)$ gives the value associated with a given point (e.g., concentration of a unit mass of soil), $m(x, y)$ represents the trend, and $e(x, y)$ denotes the fluctuations at that point. A critical feature here is that $e(x, y)$ is not a random error in the sense of the usual statistical model. Observations taken a small distance apart will be very nearly the same and highly correlated. Both the difficulties and the advantages of this approach depend very much on the assumption of a correlation or covariance function to represent the similarity of nearby observations.

With the model given above one can elect to use an estimate of the trend, $m(x, y)$, to show overall behavior of the main feature of the process, or to use trend plus fluctuation as a way to interpolate between data points. There is thus more flexibility than is available with the previous methods. However, this comes at a price comprised of both complexity and additional assumptions. The simple versions of the model assume that the covariance function and the trend are both known. Neither assumption is realistic in many practical problems. Another difficulty is that the main approaches in use require the fitted functions to pass through the data points exactly. This is, of course, essential if it is assumed that the variable of interest is measured without error. However, in many field situations it will not be true. An example occurs in sampling soil for radioactivity. Ordinarily, only a small mass of soil can be assayed, so that the samples taken in the field are subsampled in the laboratory, and no two subsamples of the original sample can be expected to give the same result. Thus, one should not require that the fitted surface should pass through every data point.

"Kriging" is essentially a method of estimation that assumes that the underlying model belongs to the class of stochastic process models described at the start of item 4, above. There are a number of variants at the basic method, and various special developments. Undoubtedly, more variations on the basic themes will soon surface, and others will have escaped our attention. The existence of the several varieties of kriging, along with the somewhat elaborate mathematics involved in the original development (initially published in French), have led to some confusion.

A comment by David (1976:3) that "kriging . . . can be expressed as plain regression" is quite correct for some elementary versions. A linear model is assumed, and the main difference from ordinary multiple regression is that weighting is used. Weighted multiple regression is, of course, not a new development. The difficulties come in at several points, which we only briefly touch on here. A more detailed assessment is needed to ascertain the ways in which the methodology can be useful for ecological research.

Perhaps the most essential issue that needs to be addressed is whether the right questions are being asked about the data. Mining engineers and petroleum geologists are normally concerned with estimating a quantity of some kind, usually that of the economically recoverable resource in a region. The appropriate methodology is one aimed at efficiently estimating a total or a mean. In field ecology and many environmental studies, the key questions are often not concerned with totals or averages. It is thus quite possible that good answers to the wrong questions may be provided by a particular methodology. Hence we believe that it is essential to investigate carefully the several versions of kriging (and related methodology) before recommending their use in particular environmental contexts.

A second issue concerns the underlying model assumptions that must be made for environmental applications. These assumptions need to be considered in several stages. A first, fundamental stage is the choice of the "process" model. The equation given under item 4 above can conveniently be illustrated by contemplating construction of a topographic map. First, one accurately determines elevations at a series of locations ("control points"). In many situations the topography is such that there will be an overall trend in the general "lay of the land" on which various local "fluctuations" are superimposed (hills, streams, etc.). If one constructs a contour map from such data, the contours on the map should coincide exactly with the observed data points. As was noted earlier, this coincidence is not necessarily true in many environmental studies. For example, soil sampling for radioactivity will seldom give the same result at adjacent points, or for repeated samples from any very small area. Consequently, the underlying model will often have three components:

$$Z(x, y) = m(x, y) + f(x, y) + e(x, y)$$

where $m(x, y)$ again represents the trend, $f(x, y)$ represents a fluctuation (i.e., a real local variation in concentration), and $e(x, y)$ is now an "error" term in the usual statistical sense. The fitted model will not, and should not, coincide with the observed data points. A computer program that forces such a fit may be quite inappropriate for the data. A realistic application, then, may require three components in the underlying model—an overall trend, identifiable local fluctuations, and an error component. Haslett and Rafferty (1989) give an interesting example of analyses of this kind applied to meteorological data.

An immediate further difficulty is the well-established fact that concentrations of many substances usually follow a skewed frequency distribution. Since virtually all of the theory for estimation methodology (e.g., kriging) assumes normally distributed errors, a transformation of the data is required. Usually a logarithmic transformation is appropriate. A non-linear model may be needed to account for the trend, and may possibly be transformable to a linear model, so that two situations calling for transformations may be encountered. There can be no assurance that they will be mutually compatible.

The next issue requiring attention is the correlation between nearby locations in the region being studied. A basic assumption in kriging is the so-called "weak stationarity." Two aspects need to be considered. The first is "first-order" stationarity, which assumes that the expected value of the variable of interest is constant over the region being studied. This is usually stated by considering a (vector) distance (h) between two points and indicating that the expected value of the differences is zero, i.e., that:

$$E\{Z(x, y) - Z[(x, y) + h]\} = 0.$$

Since this will not be true in many practical cases, we may have to postulate the trend component, $m(x, y)$, of the model above and arrange to "remove" that trend to obtain first-order stationarity.

A second component involves a "second-order" relationship, which is represented by the expectation of the squared difference of the variable of interest measured at two locations. This is the "intrinsic hypothesis" of kriging:

$$E(\{Z(x, y) - Z[(x, y) + h]\}^2) = 2g(h)$$

i.e., that the expectation of squared difference depends only on a function of the distance between two locations. Plots of this function, $g(h)$, are called the semivariogram, essentially a graph showing the correlation between values recorded at points located a distance h apart. The expression above can conveniently be thought of as a "spatial variance," i.e., it measures the variance between variables that are associated with definite locations. These are the "regionalized variables" (Journel and Huijbregts 1978:10-11) of kriging (in most of statistical theory, random variables are used in abstraction from any physical location).

To use kriging, one must arrive at a functional form for $g(h)$, and then estimate the constants involved in that functional form. An exponential model is frequently assumed, and one of the constants will often represent the error term, $e(x, y)$, above (actually the error variance). In applications it may be difficult to obtain enough suitable data to actually fit a variogram to data from the region under study. A poorly estimated or arbitrary function may thus be involved in the es-

timation process. In a final step in the process, a series of weights (adding to unity) are obtained through a Lagrangian multiplier operation on a system of equations based on variogram values for the observed locations. These weights can then be used in a moving-average process to generate values of the variable of interest over the entire study region, along with variances.

It seems quite evident that there are a number of places where the estimation process can go astray, because we are uncertain as to how well the underlying assumptions are met on real environmental data. Hence it appears that an essential step is to investigate the needs for kriging or related processes in field ecology before attempting to probe too deeply into their varying methodologies. In those instances where field data are not extremely expensive, it may be much better to collect some additional data then to begin to use a complex process with inadequate data. When additional observations are very expensive or impossible to obtain, then the effort required to find the "best" methodology may be justified.

Although applications to ecological and environmental needs are beginning to appear, mostly in reports of various kinds rather than refereed journals (however, see Bromenshenk et al. 1985, Thomas et al. 1986), little attention has yet been paid to the appropriateness of the approach in these fields. We believe that the method does have a considerable potential for environmental work, particularly because it offers an approach to situations in which data have been collected haphazardly. Often such data either cannot be replaced, or the cost of doing so is prohibitive. Kriging has considerable appeal as a way to reduce the prospects of biases that are inevitable in such data. Alternatively, it may be a valuable tool in determining the most effective ways to collect more observations to "fill in the gaps" in an existing spatial data set. At least two features of kriging need to be studied in detail before applications in ecological and environmental studies should be considered. One is the matter of the semivariogram, which is essentially a correlogram, and the key component of the kriging model. Basically, it dictates how two sampling points in a given region are assumed to be correlated. Thus, to use the method, one must both postulate a functional form (a model) for the variogram and determine values for the parameters in that functional form. Consequently, there is a need to investigate both the appropriate functions and the requirements for reliable determination of parameters on relevant data sets.

The second problem has to do with the fact that the usual rather complex fitting procedure used in point-kriging data forces the fitted surface to coincide exactly with the observed values at those spatial points where the data was collected. In many of the original applications for which the method was developed this is a useful and realistic requirement. However, it is not a sensible procedure in many ecological and environmental applications, where measurement errors must be associated with the data points. Customarily, ecologists and others tend to use plots or quadrats on which to enumerate the abundance of a given species, or the local concentration of a pollutant or other substance. Discrete objects in nature, such as plants, tend not to be randomly distributed, but to be grouped or clustered in some subregions and to be relatively rare in others. Sampling with a fixed unit of area (a plot or quadrat) will then be an inefficient approach if one wishes mainly to study patterns. The alternatives are basically those of varying the plot size locally or utilizing "distance sampling." To our knowledge, little research has been done on schemes for continuously varying plot size over the region of study. C. L. Bacheler (*personal communication*) varied plot sizes in studies conducted in New Zealand, but depended largely on an empirical justification of the results. We believe that it is feasible to develop an approach based on double-sampling theory that will permit use of variable-sized plots in sampling for pattern.

Distance sampling is another method that has received a good deal of attention from a theoretical point of view over selected decades (for reviews, see Matern 1960, Bartlett 1975, and Ripley 1981). Virtually all of these efforts have been concerned with either estimating average densities or detecting significant deviations from a random pattern. If a goal is to measure patterns, then most of the available results will not be too helpful. However, we believe that the basic technique will nonetheless be very useful, since it offers a prospect of highly efficient sampling. Plots measure local density. If there is a pattern, local density will vary, and individual plots of fixed size will thus give highly variable results. Distance sampling consists of measuring the distance from a sampling point to the nearest, or n^{th}-nearest, object of interest. Hence, it supplies a direct measure of local density; thus, it may be the best approach to studying patterns. We believe distance methods should be studied as a tool for measuring patterns, not density (as has been done in the past).

A useful starting place may be some of the techniques for testing for significant deviations from randomness, since tests often provide a useful index of the phenomenon being tested. One such index (Eberhardt 1967) was termed "Eberhardt's index" (Bartlett 1975) or "Eberhardt's statistic" (Hines and Hines 1979). Further progress may depend on testing field data collected by distance measurements and "variable plots" in some of the presently used methods for assessing pattern, such as kriging.

Discussion

We believe that the major issue identified in this study is the inadequate sample sizes used in contemporary field experimentation. The problem of "pseudoreplication" is real and important, and will have to

be dealt with in two ways. One need is educational, as discussed by Hurlbert (1984). We need to eliminate the erroneous assumption that subsampling constitutes true replication. The other need is to recognize situations, such as the power-plant "impact" studies, where true experiments are not feasible.

In some respects, this second aspect of the pseudoreplication question may also be a part of the major issue of inadequate replication. Very likely the first element in efforts to resolve the issue is one of attempting to achieve widespread recognition of the problem. In some circumstances adequate replication may be feasible. For example, Eberhardt (1988) has noted that most wildlife management agencies do have an opportunity to achieve adequate replication by way of a large number of distinct management units that could, at least conceptually, provide sizable numbers of replicates for experiments built into management operations.

The broader question may be approached by recalling the concept of a "target" population, as contrasted with the population actually sampled. Apart from some applied research, the results of most experimentation are extrapolated to a much larger population. In ecology, we seek better understanding of populations, communities, and ecosystems by experimentation on a small fraction of some larger system. As noted above (see *Introduction*) there is a strong likelihood that isolating components for an experiment will either block out or destroy some of the interactions that may comprise essential features of the larger system. Certainly we cannot say much about how well the population sampled by most experiments represents the "target" population.

Consequently a crucial need may be to reconsider overall goals in field research. Should we, in some sense, revert to descriptive ecology? Should the overall design be that of observing the larger system by sampling, with observational studies to contrast certain components, and small-scale experiments as components of the overall design? We believe that the several classes of methodology described here offer the tools for such an approach.

An overall philosophy and conceptual framework will undoubtedly need widespread discussion. Various relevant debates have been going on for some years. For example, Conner and Simberloff (1986) have espoused the use of the framework of the modern statistical approach—null hypotheses, Type I and II errors, etc.—in discussing drawing conclusions from non-experimental observations. While we think the techniques discussed here will be helpful by supplying a more formal framework for such efforts, the deeper issue is that of the logic of cause and effect in complex systems. As we have noted above (see *Analytical sampling and observational studies: observational studies*), the mechanics of statistical analyses may be essentially the same in an experimental and a non-experimental situation. The real problems come in drawing inferences from the results of the analyses.

Holland (1986) discussed "statistics and causal inference" by placing emphasis on measuring the effects of causes rather than on the more traditional course of discussing possible causes of observed effects. Unfortunately, as he notes, this approach is best served by the control provided by experimentation, since one can then identify causes with treatments, and manipulate treatments to reduce the prospect that unobserved causes may be involved in outcomes. Nonetheless, the model used provides a relatively simple framework that can help identify the distinctions between inferences based on controlled experiments and those made when only the time and place of observation can be controlled. Inferences based on sampling are much harder to support, and a great deal of attention is needed on this topic.

While sampling for modelling has been discussed in this paper, little has been said here about modelling as presently practiced in ecology. Although most of the present uses of sampling for modelling are concerned with the kinetics of trace substances, the methodology should also be useful for various other modelling purposes. We believe, however, that the wider subject of ecological and environmental modelling has not fulfilled many of the claims put forth in earlier years. As noted by Eberhardt (1977), modelling became "systems ecology," which might better be labelled "systems design" since it has proven virtually impossible to show that the models really represent ecological systems.

In many senses modelling has been substituted for adequate treatments of the inferential questions in a largely observational setting. We continue to think that modelling is best employed as an aid to understanding observational data, with the promises of "predictive power" as yet not substantiated (cf. Eberhardt 1977). The matter of "understanding the data" has much to do with drawing inferences about cause and effect, and thus seems to us to be the crucial philosophical question for field ecology today. Hence we do not wish to be understood as being opposed to modelling, but instead think that its directions and purposes need more attention.

From the more immediate point of view of planning field investigations, it should be noted that the several classes of sampling described here offer important gains in efficiency, i.e., they provide more information per dollar expended in the field. Eberhardt and Gilbert (1980) used data from field studies of soil contamination to show the sharp difference in allocation of samples to strata or domains of study required for three different sampling schemes (descriptive sampling, sampling for pattern, and analytical sampling). The difficulty in realizing such gains is largely that of setting on a single set of objectives for a study. Fig. 2 suggests one approach to distinguishing the objectives of the four classes of sampling. Essentially, descriptive sam-

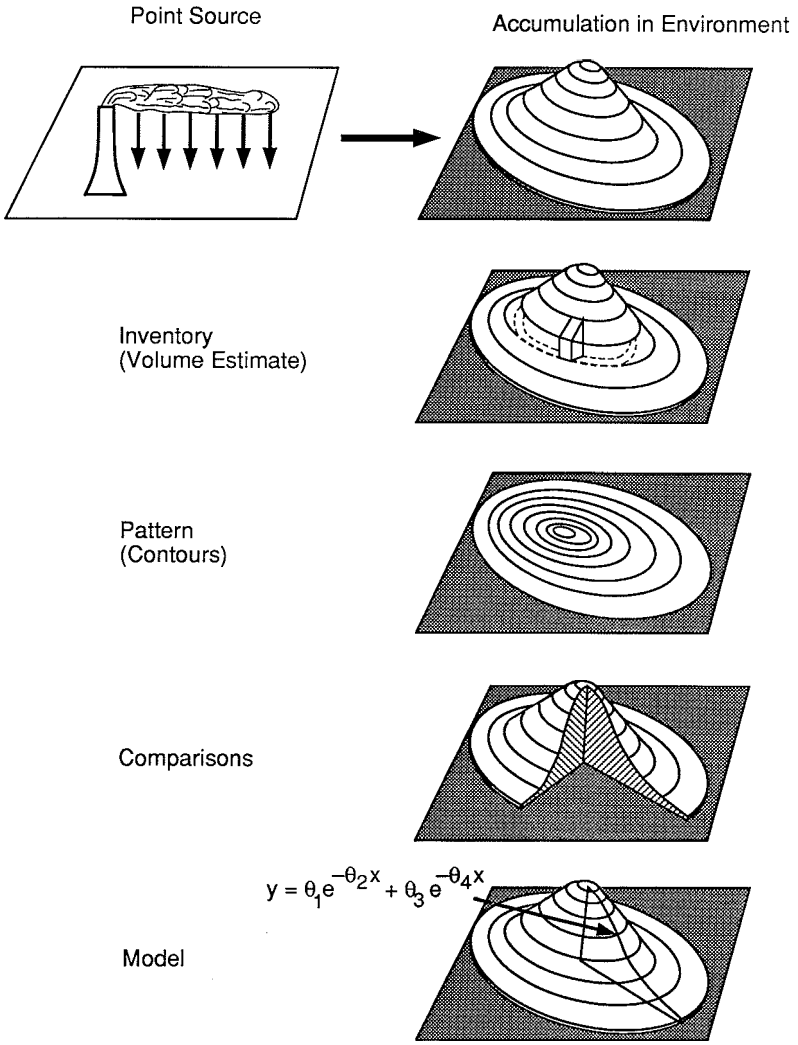

FIG. 2. An illustration of the four categories of sampling. The diagrams at the right are hypothetical concentrations in soil around a point source, here suggested as a smokestack. In practice, it is likely that the elliptical concentration pattern might be more strongly skewed in the direction of the prevailing winds.

pling can be described as an "inventory" method, sampling for pattern produces a contour map, while analytical sampling provides a basis for comparisons. Sampling for modelling stands somewhat apart due to the focus on parameter estimation for a specific model.

ACKNOWLEDGMENTS

This research was supported by the Environmental Research Division of the Office of Health and Environmental Research of the United States Department of Energy under contract DE-AC06-76-RLO1830. Helpful comments from S. H. Hurlbert, D. Meeter, and three anonymous referees are gratefully acknowledged.

LITERATURE CITED

Aitken, M. 1981. Regression models for repeated measurements. Biometrics **37**:831–832.

Atkinson, A. C. 1969. A test for discriminating between models. Biometrika **56**:337–347.

———. 1972. Planning experiments to detect inadequate regression models. Biometrika **59**:275–293.

Atkinson, A. C., and W. G. Hunter. 1968. The design of experiments for parameter estimation, Technometrics **10**: 271–289.

Bard, Y. 1974. Nonlinear parameter estimation. Academic Press, New York, New York, USA.

Bartlett, M. S. 1975. The statistical analysis of spatial pattern. Chapman and Hall, London, England.

Beck, J. V., and K. J. Arnold. 1977. Parameter estimation. John Wiley & Sons, New York, New York, USA.

Blau, G. E. 1975. Mathematical model building with an application to determine the distribution of Dursban insecticide added to a simulated ecosystem. Advances in Ecological Research **9**:133–163.

Box, G. E. P., and D. R. Cox. 1964. An analysis of transformations. Journal of the Royal Statistical Society B**26**: 211–243.

Box, G. E. P., and N. R. Draper. 1987. Empirical model-building and response surfaces. John Wiley & Sons, New York, New York, USA.

Box, G. E. P., and W. J. Hill. 1967. Discrimination among mechanistic models. Technometrics **9**:57–71.

Box, G. E. P., W. J. Hunter, and J. S. Hunter. 1978. Statistics for experimenters: an introduction to design, data analysis, and model building. John Wiley & Sons, New York, New York, USA.

Box, G. E. P., and H. L. Lucas. 1959. Design of experiments in non-linear situations. Biometrika **46**:77–90.

Box, G. E. P., and G. C. Tiao. 1975. Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical Association **70**: 70–79.

Box, M. J. 1970. Some experiences with nonlinear experimental design criteria. Technometrics **12**:569–589.

———. 1971. An experimental design criterion for precise estimation of a subset of the parameters in a nonlinear model. Biometrika **58**:149–153.

Bromenshenk, J. J., S. R. Carlson, J. C. Simpson, and J. M. Thomas. 1985. Pollution monitoring of Puget Sound with honey bees. Science **227**:632–634.

Campbell, D. T., and J. C. Stanley. 1963. Experimental and quasi-experimental designs for research. Houghton-Mifflin, Boston, Massachusetts, USA.

Clark, I. 1982. Practical geostatistics. Applied Science, London, England.

Cochran, W. G. 1972. Observational studies. Pages 77–90 *in* T. A. Bancroft, editor. Statistical papers in honor of G. W. Snedecor. Iowa State University Press, Ames, Iowa, USA.

———. 1973. Experiments for nonlinear functions. Journal of the American Statistical Association **68**:771–781.

———. 1977. Sampling techniques. Third edition. John Wiley & Sons, New York, New York, USA.

———. 1983. Planning and analysis of observational studies. John Wiley & Sons, New York, New York, USA.

Conner, E. F., and D. Simberloff. 1986. Competition, scientific method, and null models in ecology. American Scientist **74**:155–162.

Cook, T. D., and D. T. Campbell. 1979. Quasi-experimentation: designs and analysis issues for field settings. Houghton-Mifflin, Boston, Massachusetts, USA.

David, M. 1976. The practice of kriging. Pages 31–48 *in* M. Guarascio, M. David, and C. Huijbregts, editors. Advanced geostatistics in the mining industry. D. Reidel, Dordrecht, The Netherlands.

Deming, W. E. 1950. Some theory of sampling. John Wiley & Sons, New York, New York, USA.

———. 1975. On probability as a basis for action. The American Statistician **29**:146–152.

Eberhardt, L. L. 1967. Some developments in "distance sampling." Biometrics **23**:207–216.

———. 1975*a*. Some problems in measuring ecological effects of chronic low-level pollutants. Pages 565–586 *in* The costs and effects of chronic exposure to low-level pollutants in the environment. Hearings before the subcommittee on the environment and the atmosphere of the Committee on Science and Technology, 94th Congress. United States Government Printing Office, Washington, D.C., USA.

———. 1975*b*. Some methodology for appraising contaminants in aquatic systems. Journal of the Fisheries Research Board of Canada **32**:1852–1859.

———. 1976. Quantitative ecology and impact assessment. Journal of Environmental Management **42**:1–31.

———. 1977. Applied systems ecology: models, data, and statistical methods. Pages 43–55 *in* G. S. Innis, editor. New directions in the analysis of ecological systems. Part 1. The Society for Computer Simulation, La Jolla, California, USA.

———. 1978*a*. Designing ecological studies of trace substances. Pages 8–33 *in* D. C. Adriano and I. L. Brisbin, Jr., editors. Environmental chemistry and cycling processes. CONF-760429. United States Department of Energy, Washington, D.C., USA.

———. 1978*b*. Appraising variability in population studies. Journal of Wildlife Management **42**:207–238.

———. 1978*c*. Transect methods for population studies. Journal of Wildlife Management **42**:1–31.

———. 1988. Testing hypotheses about populations. Journal of Wildlife Management **52**:50–56.

Eberhardt, L. L., and R. O. Gilbert. 1980. Statistics and sampling in transuranic studies. Pages 173–186 *in* W. C. Hanson, editor. Transuranic elements in the environment. DOE/TIC22800. Technical Information Center, United States Department of Energy, Washington, D.C., USA.

Eberhardt, L. L., R. O. Gilbert, H. L. Hollister, and J. M. Thomas. 1976. Sampling for contaminants in ecological systems. Environmental Science and Technology **10**:917–925.

Eberhardt, L. L., and M. A. Simmons. 1987. Calibrating indices of population abundance using double sampling. Journal of Wildlife Management **51**:665–675.

Endrenyi, L. (editor). 1981. Kinetic data analysis: design and analysis of enzyme and pharmacokinetic experiments. Plenum, New York, New York, USA.

Federer, W. T. 1984. Cutting edges in biometry. Biometrics **40**:827–839.

Finney, D. J. 1978. Reader reaction: testing the effect of an intervention in sequential ecological data. Biometrics **34**: 706–707.

———. 1982. Intervention and correlated sequences of observations. Biometrics **38**:255–267.

Gilbert, R. O. 1987. Statistical methods for environmental pollution monitoring. Van Nostrand Reinhold, New York, New York, USA.

Green, R. H. 1979. Sampling design and statistical methods for environmental biologists. John Wiley & Sons, New York, New York, USA.

Gurevitch, J., and S. T. Chester, Jr. 1986. Analysis of repeated measures experiments. Ecology **67**:251–255.

Hansen, M. H., W. N. Hurwitz, and W. G. Madow. 1953. Sample survey methods and theory. Volume 1. Methods and applications. John Wiley & Sons, New York, New York, USA.

Haslett, J., and A. E. Rafferty. 1989. Space-time modelling with long-memory dependence: assessing Ireland's wind power resource. Applied Statistics **38**:1–50.

Hayne, D. W. 1978. Experimental designs and statistical analyses. Pages 3–13 *in* D. P. Snyder, editor. Populations of small mammals under natural conditions. Volume 5. Special Publications Series. Pymatuning Laboratory of Ecology, University of Pittsburg, Pittsburg, Pennsylvania, USA.

Hill, W. J., and W. G. Hunter. 1966. A review of response surface methodology: a literature survey. Technometrics **8**: 571–590.

Hill, W. J., and W. G. Hunter. 1974. Design of experiments for subsets of parameters. Technometrics **16**:425–434.

Hill, W. J., W. G. Hunter, and D. W. Wichern. 1968. A joint design criterion for the dual problem of model discrimination and parameter estimation. Technometrics **10**: 145–160.

Hines, W. G. S., and R. J. O'Hara Hines. 1979. The Eberhardt statistic and the detection of nonrandomness of spatial point distributions. Biometrika **66**:73–79.

Holland, P. W. 1986. Statistics and causal inference. Journal of the American Statistical Association **81**:945–960.

Hunter, W. G., and Reiner, A. M. 1965. Designs for discriminating between two rival models. Technometrics **7**: 307–323.

Hurlbert, S. H. 1984. Pseudoreplication and the design of

ecological field experiments. Ecological Monographs **54**:187–211.

James, F. C., and C. E. McCulloch. 1985. Data analysis and the design of experiments in ornithology. Chapter 1 *in* R. F. Johnston, editor. Current ornithology. Volume 2. Plenum, New York, New York, USA,.

Jessen, R. J. 1978. Statistical survey techniques. John Wiley & Sons, New York, New York, USA.

Johnson, N. L., and S. Kotz. 1969. Discrete distributions. Houghton Mifflin, Boston, Massachusetts, USA.

Journel, A. G., and C. J. Huijbregts. 1978. Mining geostatistics. Academic Press, New York, New York, USA.

Kamil, A. C. 1988. Experimental design in ornithology. Chapter 8 *in* R. F. Johnston, editor. Current ornithology. Volume 5. Plenum, New York, New York, USA.

Khuri, A. I., and J. A. Cornell. 1987. Response surfaces: design and analysis. Marcel Dekker, New York, New York, USA.

Kish, L. 1965. Survey sampling. John Wiley & Sons, New York, New York, USA.

Knight, R. R., and L. L. Eberhardt. 1984. Projected future abundance of the Yellowstone grizzly bear. Journal of Wildlife Management **48**:1434–1438.

Krebs, C. J., M. S. Gaines, B. L. Keller, J. H. Myers, and R. H. Tamarin. 1973. Population cycles in small rodents. Science **179**:35–41.

Matern, B. 1960. Spatial variation. Meddelanden fran Statens Skogsforskningsinstitut **49**:1–144.

Mead, R., and D. J. Pike. 1975. A review of response surface methodology from a biometric point of view. Biometrics **31**:803–851.

Millard, S. P., J. R. Yearsley, and D. P. Lettenmaier. 1985. Space–time correlation and its effects on methods for detecting aquatic ecological change. Canadian Journal of Fisheries and Aquatic Science **42**:1391–1400.

Myers, R. H., A. I. Khuri, and W. H. Carter, Jr. 1989. Response surface methodology: 1966–1988. Technometrics **31**:137–157.

O'Neill, R. V. 1988. Hierarchy theory and global change. Chapter 3 *in* T. Rosswall, R. G. Woodmansee, and P. G. Rosser, editors. Scales and global change. John Wiley & Sons, New York, New York, USA.

———. 1989. Perspectives in hierarchy and scale. Chapter 10 *in* J. Roughgarten, R. M. May, and S. A. Levin, editors. Perspectives in ecological theory. Princeton University Press, Princeton, New Jersey, USA.

Patil, G. P., and C. R. Rao. 1978. Weighted distributions and size-biased sampling with applications to wildlife populations and human families. Biometrics **34**:179–189.

Ripley, B. D. 1981. Spatial statistics. John Wiley & Sons, New York, New York, USA.

Romesburg, H. C. 1981. Wildlife science: gaining reliable knowledge. The Journal of Wildlife Management **45**:293–313.

Scheffé, H. 1959. The analysis of variance. John Wiley & Sons, New York, New York, USA.

Schindler, D. W. 1974. Eutrophication and recovery in experimental lakes: implications for lake management. Science **184**:897–898.

Slonim, M. J. 1960. Sampling: a quick, reliable guide to practical statistics. Simon and Schuster, New York, New York, USA.

Stewart-Oaten, A., W. R. Murdoch, and K. R. Parker. 1986. Environmental impact assessment: "pseudoreplication" in time? Ecology **67**:929–940.

Stuart, A. 1976. Basic ideas of scientific sampling. Charles Griffin, London, England.

Thomas, J. M., J. A. Mahaffey, K. L. Gore, and D. G. Watson. 1978. Statistical methods used to assess biological impact at nuclear power plants. Journal of Environmental Management **7**:269–290.

Thomas, J. M., J. R. Skalski, J. F. Cline, M. C. McShane, J. C. Simpson, W. E. Miller, S. A. Peterson, C. A. Callahan, and J. C. Greene. 1986. Characterization of chemical waste site contamination and determination of its extent using bioassays. Environmental Toxicology and Chemistry **5**:471–501.

Tukey, J. W. 1960. Conclusions vs. decisions. Technometrics **2**:423–433.

Turk, T. R. 1978. Testing the effect of an intervention in sequential ecological data. Biometrics **34**:128–129.

Van Bell, G., and T. Zeisig. 1978. Response to a query by J. R. Turk. Biometrics **34**:708.

Ware, J. H. 1981. Tracking: prediction of future values from serial measurements. Biometrics **37**:427–437.

Williams, B. 1978. A sampler on sampling. John Wiley & Sons, New York, New York, USA.

Wold, H. 1956. Causal inference from observational data. Journal of the Royal Statistical Society, Series **A**, **119**, Part 1:28–65.

Yates, F. 1981. Sampling methods for censuses and surveys. MacMillan, New York, New York, USA.

Zahl, S. 1978. Response to a query by J. R. Turk. Biometrics **34**:707–708.