OCS Study
MMS 2003-035

Using Before-After-Control Impact in Environmental Assessment: Purpose, Theoretical Basis, and Practical Problems

Final Technical Summary

Final Study Report

U.S. Department of the Interior
Minerals Management Service
Pacific OCS Region

OCS Study
MMS 2003-035

Using Before-After-Control Impact in Environmental Assessment: Purpose, Theoretical Basis, and Practical Problems

Final Technical Summary

Final Study Report

Authors

Allan Stewart-Oaten
Principal Investigator

Prepared under MMS Cooperative
Agreement Nos. 14-35-0001-30471 & 14-35-0001-30761
by
Southern California Educational Initiative
Marine Science Institute
University of California
Santa Barbara, CA   93106

U.S. Department of the Interior
Minerals Management Service
Pacific OCS Region

Camarillo
May 2003

Disclaimer

This report has been reviewed by the Pacific Outer Continental Shelf Region, Minerals Management Service, U.S. Department of the Interior and approved for publication. The opinions, findings, conclusions, or recommendations in this report are those of the author, and do not necessarily reflect the views and policies of the Minerals Management Service. Mention of trade names or commercial products does not constitute an endorsement or recommendation for use. This report has not been edited for conformity with Minerals Management Service editorial standards.

Availability of Report

Extra copies of the report may be obtained from:
U.S. Dept. of the Interior
Minerals Management Service
Pacific OCS Region
770 Paseo Camarillo
Camarillo, CA 93010
Phone: 805-389-7621

A PDF file of this report is available at:
http://www.coastalresearchcenter.ucsb.edu/SCEI/

Suggested Citation

The suggested citation for this report is:

Stewart-Oaten, Allan. Using Before-After-Control-Impact in Environmental Assessment: Purpose, Theoretical Basis, and Practical Problems. MMS OCS Study 2003-035. Coastal Research Center, Marine Science Institute, University of California, Santa Barbara, California. MMS Cooperative Agreement Numbers 14-35-0001-30471 & 14-35-0001-30761. 116 pages.

Table of Contents

FINAL TECHNICAL SUMMARY

STUDY TITLES:

Study I. **a.** Environmental Assessment: Statistical Description of Variable Effects on Fluctuating Populations

b. Environmental Assessment: Statistical Description of Variable Effects on Fluctuating Populations (Continuation)

Study II. Adding Biology to BACI: Exploring the Use of Functional Groups, Trophic Relationships and Multiple, Ecologically Similar Comparison Sites in Choosing Models and Estimating Effects Impacts Analysis

REPORT TITLE: Using Before-After-Control-Impact in Environmental Assessment: Purpose, Theoretical Basis, and Practical Problems

CONTRACT NUMBERS: 14-35-0001-30471 & 14-35-0001-30761

SPONSORING OCS REGION: Pacific

APPLICABLE PLANNING AREA: Southern California

FISCAL YEAR(S) OF PROJECT FUNDING:

Study I: **a.** FY 91, FY 92, FY 93, FY 94

b. FY 95, FY 96, FY 97

Study II: FY 97, FY 98

COMPLETION DATE OF THE REPORT: March 2001

COST(S):

Study I: **a.** FY 91 - $27,440, FY 92 - $36,163, FY 93 - $31,587, FY 94 - $16,650

b. FY 95 - $22,346, FY 96 - $83,980, FY 97 - $78,674

Study II: FY 97 - $30,148, FY 98 – $47,400, FY 99 – no cost

CUMULATIVE PROJECT COSTS:

Study I: **a.** $111,840

b. $185,000

Study II: $77,548

PROJECT MANAGER: Russell J. Schmitt

AFFILIATION: University of California, Santa Barbara

ADDRESS: Coastal Research Center, Marine Science Institute, University of California, Santa Barbara, CA 93106-6150

PRINCIPAL INVESTIGATOR: [1]Allan Stewart-Oaten

Co-PI (Study II): [2]Stephen C. Schroeter

ADDRESSES: [1]Department of Biological Sciences and Department of Ecology, Evolution, and Marine Biology, University of California, Santa Barbara, CA, 93106-9610; [2]Marine Science Institute, University of California, 2270 Camino Vida Roble, Suite L, Carlsbad, CA 92009

BACKGROUND: Three types of decision in the management of offshore mineral resources are: whether to accept, modify or reject a proposed development (e.g., an oil platform), whether to continue, modify or discontinue an ongoing operation, and whether an operation has caused such damage that penalties or mitigation should be required. The second and third types often depend on monitoring assessment: determining the effects of a development after a period of operation. Many environmental variables, like population abundances, fluctuate naturally over time even without anthropogenic intervention, so assessments need to deal with time series where both serial correlation and systematic (e.g., seasonal) variation are likely.

OBJECTIVES: The main aims of these two programs have been:
(1) to develop ways to estimate or describe effects of an "alteration" of the environment on naturally fluctuating biological variables, using one or more neighboring, similar "control" sites to reduce and estimate the effects of natural temporal variation on these estimates;
(2) to test the methods on real data, from the annual surveys of the Channel Islands National Park Service: specifically, to see whether neighboring, similar sites can be used to predict each other's fluctuations over time, such that the temporal variation and serial correlation of the residuals from the prediction are smaller than those of the original data.

DESCRIPTION: Two related methods were studied. One uses the difference between values observed at the "Impact" site and those observed at the controls. The values at the sites might be transformed, and multiple values from a set of controls can be averaged or otherwise summarized. The other method uses control values as covariates; in effect, it finds equations for predicting Impact values from control values under "before alteration" conditions and compares their predictions with either the predictions of the corresponding equations under "after" conditions or with the actual values of the sites observed after the alteration is in place. In this project, our main aim has been to develop and explain the broad approach, dealing with objections and misunderstandings, and explaining why some other approaches, which do not account for variation over time, do not separate natural variation from human effects.

The Channel Islands data are annual surveys of about 70 species at 13 sites since 1981; a further site was added in 1983 and two more in 1986. Some species have been added and others dropped over the years. There are data on size and recruitment, but we have worked mainly with the abundance data for three groups, "band" , "quad" and "rpc" (the names refer to the sampling method). At each site, all samples use a single 100-metre transect. The positions on this transect are chosen anew each year, by choosing a random point and spacing positions equally from it to each end. The "band" species are sampled by counting in bands (currently 3X20m but this has varied) across the transect at the chosen positions. The "quad" species are sampled by counts in quadrats (currently 2m^2 but this has varied), and the "rpc" species by the fractions of contacts with (currently) 40 points on the boundaries of two concentric ellipses. Our main efforts here have gone into finding ways to summarize the data to check the idea that data from "similar" sites might have similar fluctuations. We have used plots and correlations for combinations of species, pairs and triples of sites, and transformations (raw data, logs, reciprocals, etc.), but there are so many possibilities that we need ways to summarize these summaries.

SIGNIFICANT CONCLUSIONS: 1. An explicit, unambiguous definition of an "effect" is needed for design and interpretation of assessments. A few approaches and critiques use an explicit definition which is flawed. An example is "Impacts are those disturbances that cause mean abundance in a site to change more than is found on average" (Underwood 1992: Journal of Experimental Marine Biology and Ecology, 161, p. 152). The "mean" and "average" are undefined: in fact, the "mean" is implicitly taken to be the mean over the study period (thus not allowing for natural fluctuation over a time period of this length), and the average is over a "population" of sites which must be chosen subjectively - in practice, usually implicitly. Many other studies use implicit definitions with similar flaws. The definition of an effect should refer only to the alteration site, since a given effect is not changed if unaffected sites are naturally very different from this site. If other sites are similar to the alteration site, they can be used to improve estimation of the effect, but not to define it.

2. There may not be a single "best" use of "control" sites. The most direct use is by differences or as covariates (see DESCRIPTION), but this involves a tradeoff between reducing the effects of large, long-term natural fluctuations, assumed to be widespread, and increasing the effect of short term, local variation and sampling error. This can be beneficial even when the apparent error of effect estimates seems to increase, because the apparent error is likely to underestimate the variation due to long-term fluctuations. However, the control site that best reduces these fluctuations may be different for different species. There may be no control site similar enough to do it well. There may be no large, long-term fluctuations, or some that are local but too rare or irregular to be allowed for in a time series model.

STUDY RESULTS: 1. In papers 2, 6 and especially 5 below, we have defined an effect as the difference between the average abundance at the Impact site over some long period, such as the life-length of the alteration, and the average that would have been obtained without the alteration. We then show that, in principle, time series methods can be used to estimate this difference and give an estimate of uncertainty.

2. The Channel Islands data illustrate several of the difficulties in the use of control sites (see CONCLUSIONS). Some pairs of close sites have high correlations for some species, but the relation between correlation and distance is weak, for any obvious meaning of "distance". For example, maps with lines joining sites with correlations > x (for various x) show plenty of lines connecting distant sites, and sites facing north to sites facing south. Sampling error is part of the cause: most of the time series are best fitted as independent observations. Variation over the replicate quadrats, is usually much smaller than the variation over time, but it does not include the error in the fixed 100m transect as an estimate for the site. Another cause may be that the sites were deliberately chosen to give as broad a range of conditions as possible - and thus to be dissimilar from each other. The study continues by comparing series from the first 50m of the 100m transect to the series given by subtracting the second half from

it: the series have very similar variance and correlation properties, though the difference series is fitted by independent observations more often. Two broad alternatives when sampling error is high are (1) to analyze "Impact only" time series, using neighboring sites less formally to rule out alternative explanations for an observed effect, and (2) to use time series models derived from simple population dynamics models.

STUDY PRODUCTS:

Stewart-Oaten, A. 1996a. Goals in environmental monitoring. (In Detecting Ecological Impacts, C. Osenberg and R. J. Schmitt, eds, Academic Press.)

Stewart-Oaten, A. 1996b. Problems in the analysis of environmental monitoring data. 1996. (In Detecting Ecological Impacts, C. Osenberg and R. J. Schmitt, eds, Academic Press.)

Bence, J. R., A. Stewart-Oaten and S. C. Schroeter. 1996. Estimating the size of an effect from a Before-After-Control-Impact-Pairs design: the predictive approach applied to a power plant study. (In Detecting Ecological Impacts, C. Osenberg and R. J. Schmitt, eds, Academic Press.)

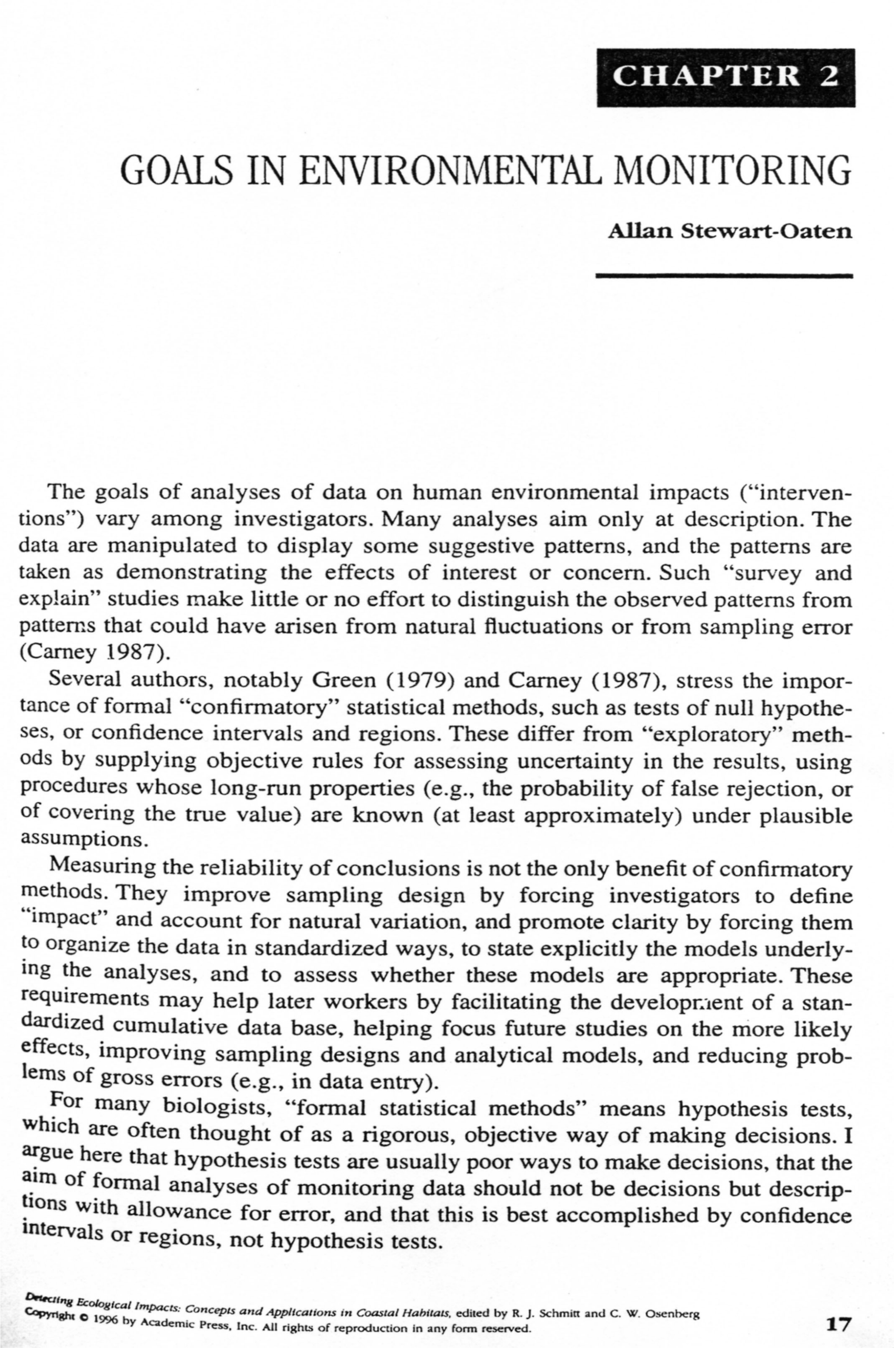
Stewart-Oaten, A. 1996c. Sequential Estimation of log(Abundance). Biometrics 52: 38-49.

Stewart-Oaten, A. and J. R. Bence. 2001. Temporal and Spatial Variation in Environmental Impact Assessment. Ecological Monographs, 71: 305-339.

Stewart-Oaten, A. 2001a. Impact assessment. Encyclopedia of Environmetrics. (Wiley 2001).

Stewart-Oaten, A. 2001b. Pseudoreplication. Encyclopedia of Environmetrics. (Wiley 2001).

FINAL STUDY REPORT



CHAPTER 2

GOALS IN ENVIRONMENTAL MONITORING

Allan Stewart-Oaten

The goals of analyses of data on human environmental impacts ("interventions") vary among investigators. Many analyses aim only at description. The data are manipulated to display some suggestive patterns, and the patterns are taken as demonstrating the effects of interest or concern. Such "survey and explain" studies make little or no effort to distinguish the observed patterns from patterns that could have arisen from natural fluctuations or from sampling error (Carney 1987).

Several authors, notably Green (1979) and Carney (1987), stress the importance of formal "confirmatory" statistical methods, such as tests of null hypotheses, or confidence intervals and regions. These differ from "exploratory" methods by supplying objective rules for assessing uncertainty in the results, using procedures whose long-run properties (e.g., the probability of false rejection, or of covering the true value) are known (at least approximately) under plausible assumptions.

Measuring the reliability of conclusions is not the only benefit of confirmatory methods. They improve sampling design by forcing investigators to define "impact" and account for natural variation, and promote clarity by forcing them to organize the data in standardized ways, to state explicitly the models underlying the analyses, and to assess whether these models are appropriate. These requirements may help later workers by facilitating the development of a standardized cumulative data base, helping focus future studies on the more likely effects, improving sampling designs and analytical models, and reducing problems of gross errors (e.g., in data entry).

For many biologists, "formal statistical methods" means hypothesis tests, which are often thought of as a rigorous, objective way of making decisions. I argue here that hypothesis tests are usually poor ways to make decisions, that the aim of formal analyses of monitoring data should not be decisions but descriptions with allowance for error, and that this is best accomplished by confidence intervals or regions, not hypothesis tests.

Detecting Ecological Impacts: Concepts and Applications in Coastal Habitats, edited by R. J. Schmitt and C. W. Osenberg

A Case for Confidence Intervals

Tukey (1960) makes a useful distinction between conclusions and decisions. Roughly, this is the distinction between choosing what to believe and choosing how to act. Conclusions are statements we accept until unusually strong contradictory evidence appears: they are "subject to future rejection" and are "judged by their long-run effects, by their 'truth', not by specific consequences of specific actions." Further distinctions are made between statistical and experimenter's conclusions (the latter involving more uncertainty, because of the possibility of systematic error) and between qualitative and quantitative conclusions (the former judging the truth of a single assertion about a parameter, the latter presenting a region where its value is believed to lie).

Decisions are choices of actions, determined by our assessment of their probable consequences in a specific situation. They may mimic conclusions—that is, we may choose to act as if a particular conclusion is true—but we make some decisions without any evidence at all, and we may reasonably make decisions implying opposite conclusions (e.g., to carry life insurance and also to save for retirement).

In the subsections below, I argue that (a) hypothesis tests are not well suited to decision making; (b) what is really wanted from biologists, ecologists and biological data analysts in environmental monitoring is not decisions but conclusions (including the allowance to be made for error) which can become part of the basis for decision making, usually by others; (c) the conclusions should not be results of significance tests because these carry too little information, e.g., about effect size, and are confusing; and (d) confidence intervals provide clear and informative conclusions, though they may need to be augmented to allow for uncertainty not considered in the formal framework.

Hypothesis Tests and Statistical Decisions

The inadequacy of statistical hypothesis testing for making decisions was recognized by the founders of the predominant methodology:

> The sum total of the reasons which will weigh with the investigator in accepting or rejecting the hypothesis can very rarely be expressed in numerical terms. All that is possible … is to balance the results of a mathematical summary … against other less precise impressions. … The tests themselves give no final verdict, but as tools help the worker who is using them to form his final decision … (Neyman and Pearson 1928).

Among these less precise impressions are "*a priori* or *a posteriori* considerations," as stressed by Bayesians. A later paper recognizes the costs of wrong decisions:

> If we reject H_0, we may reject it when it is true; if we accept H_0, we may be accepting it when it is false … in some cases it is more important to avoid the first, in others the second. … Is it more serious to convict an innocent man or to acquit a guilty? That will depend on the consequences of the error. … The use of these statistical tools in any given case, in

> determining just how the balance should be struck, must be left to the investigator (Neyman and Pearson 1933).

Despite these cautions, the Neyman-Pearson theory does seem to have been intended as a decision-making system, as evidenced by the use of acceptance sampling, clearly a decision setup, as a paradigm (Neyman and Pearson 1936) and by Neyman's later advocacy of "inductive behavior" (e.g., Neyman 1957, 1962). Together with game theory, it laid the groundwork for the more comprehensive theory of decision making under uncertainty developed by Wald (1950). This posed the problem of choosing from among a set of actions, A, when the gain of choosing act a is $G(a,s)$, where s is the "state of Nature," known only to be a member of a set, S. Data analysis enters if we are allowed to observe a variable, X, whose distribution, $P_{s,}$ depends on s. The task then is to choose a decision function, δ, which selects an action, $\delta(X)$, for each X, so as to maximize the expected gain, $\int G(\delta(X),s)P_s(dX)$. To pose hypothesis testing as a decision problem, S is the set of possible values of the true parameter (e.g., the mean of a population), H_0 asserts that this parameter is in some subset of S, the set of actions is $A = \{$"accept H_0", "reject H_0"$\}$, and the gain may be 0 if we choose correctly but some negative number (depending on the true s) if we do not; X may be the values in a sample from the population whose mean is s.

For all but the simplest problems (e.g., testing a simple hypothesis against a simple alternative), there is no best solution: even with the observed X, the gain depends on the unknown value of s. There are various types of "good" solutions, such as minimax (the solution whose worst result is least bad) and admissible (a solution such that no other solution does at least as well for every s, and better for at least one s). An important type of admissible solution is the Bayesian: a solution whose expected gain, averaged over S according to some distribution p, i.e., $\int\int G(\delta(X),s)P_s(dX)p(ds)$, is largest. If p is known, then Bayes' theorem can be used to rewrite this integral as $\int\int G(\delta(X),s)p(ds|X)P(dX)$, where $p(.|X)$ is the conditional distribution of s given X, and P is the unconditional distribution of X; the "best" $\delta(X)$ can then be chosen to be the action that maximizes the inner integral. Controversy arises here over the determination, or even the existence, of p, the "prior" distribution on the states. It is usually not possible to interpret this as a distribution in the frequentist sense, i.e., as given by the relative frequency of various outcomes in a long run of trials. But this "objectivist" interpretation of probability is not the only one. One alternative is the "personalistic" view, in which probability measures the confidence that an individual has in some proposition. From this basis, Savage (1954) developed and persuasively advocated a Bayesian approach to statistical inference and decision making. Shorter or more gentle accounts of this work are given by Edwards et al. (1963) and Pratt et al. (1965a, 1965b).

Neyman-Pearson inference, Wald's decision theory, and Savage's Bayesianism have all spawned large literatures and many useful insights and techniques. There is no consensus on a "best" approach: many statisticians will mix fields, e.g., using Bayesian methods to make inferences with good Neyman-Pearson properties. A minority would add Fisher's (1937, 1956) fiducial inference to the list. Some feeling for the variety of positions, the main arguments, and perhaps the areas of general agreement can be found in the special issue of *Synthese* (1977).

Tukey (1960) sees the Neyman-Pearson theory of hypothesis testing as a step toward a theory of decision making but "if this view is correct, Wald's decision theory now does much more nearly what tests of hypotheses were intended to do. Indeed, there are three ways in which it does better": focusing on gains or losses (rather than error probabilities), considering a wider range of setups with less stringent assumptions, and showing that there will not be "a single best procedure but rather an assortment of good procedures ... from which judgment and insight ... (perhaps best expressed in the form of an *a priori* distribution) must be used to select the 'best' procedure." However, aspects of the Neyman-Pearson theory, such as the power function and confidence procedures, remain valuable, along with tests of significance, in "conclusion theory".

These views seem broadly accepted by statisticians. They are not unchallenged, but in most disciplines involving data-based decision making (e.g., statistics, economics, business and engineering), the dissenters would give the Neyman-Pearson theory a smaller role, not a larger one. They would apply some form of decision theory, most often Bayesian, to conclusions as well as decisions.

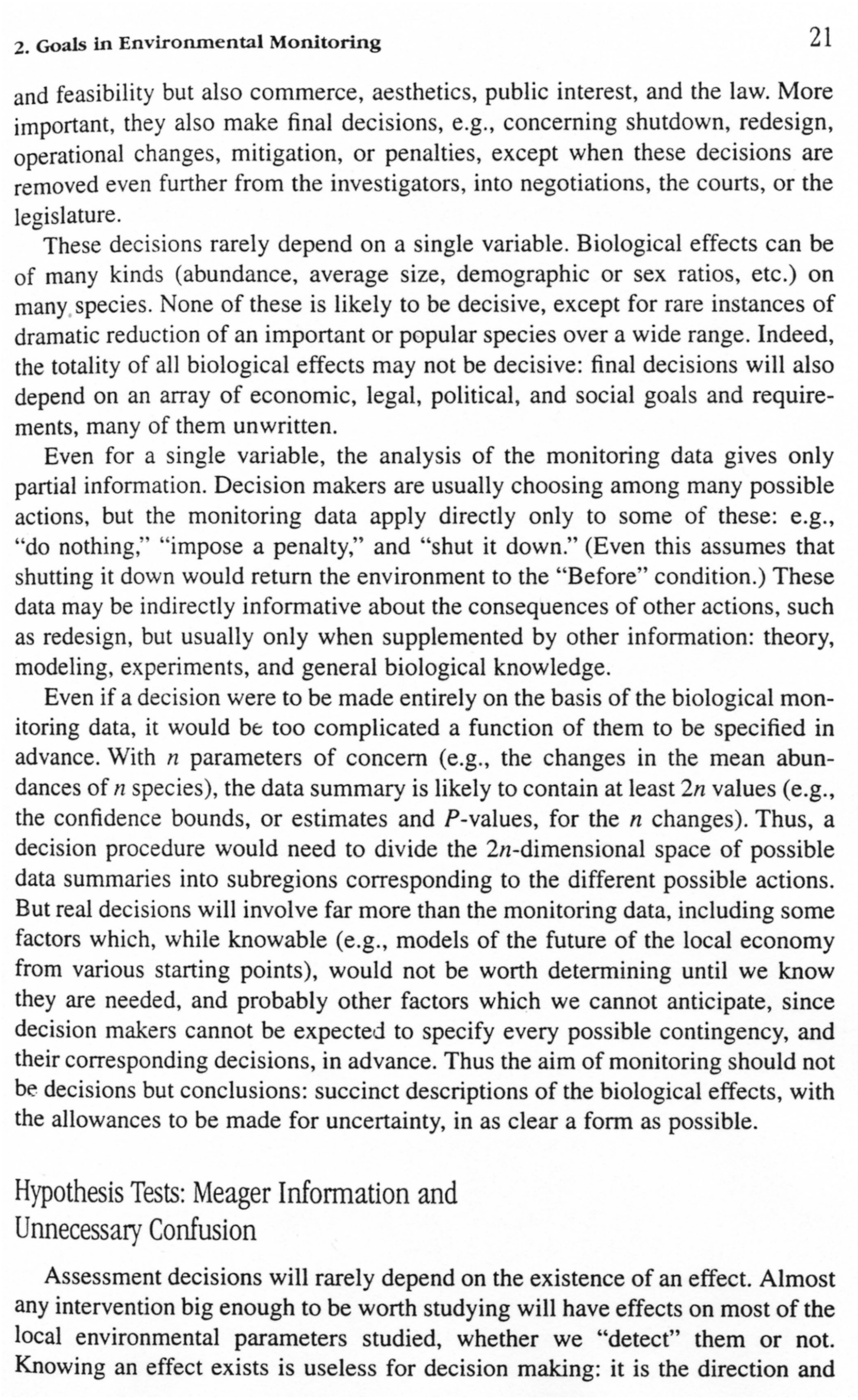
However, biology is a holdout of the "*P*-value culture" (Nelder 1991). Hypothesis testing is presented as a decision problem and treated as the only way to deal quantitatively with uncertainty, whether of conclusions or decisions. Conversely, decision problems are frequently forced, with great effort and ingenuity, into the hypothesis testing mold. The achievements of 50 years of decision theory are not rejected but simply ignored altogether: like trying to use an ax to do fine woodwork, while ignoring the band saw.

Conclusions, Not Decisions

In pure research, we usually need conclusions, not decisions. Environmental monitoring is not pure research, and its ultimate aim is decision making, but here, too, the role of scientific investigators and data analysts is to present conclusions which become part (and only part) of the basis for decision making.

The ultimate decisions are usually not made by the investigator. In pure research, investigators decide what to study and how to study it, but the result is a set of conclusions; others (referees, editors, funding agencies and readers) decide whether to accept the conclusions, or to publish them, or to make or fund further studies because of them. Their main function is to "reduce the spread of the bundle of working hypotheses which are regarded as still consistent with the observations" (Tukey 1960).

In monitoring, investigators often have even less discretion. Managers and review boards often decide what to study, considering not only scientific interest



and feasibility but also commerce, aesthetics, public interest, and the law. More important, they also make final decisions, e.g., concerning shutdown, redesign, operational changes, mitigation, or penalties, except when these decisions are removed even further from the investigators, into negotiations, the courts, or the legislature.

These decisions rarely depend on a single variable. Biological effects can be of many kinds (abundance, average size, demographic or sex ratios, etc.) on many species. None of these is likely to be decisive, except for rare instances of dramatic reduction of an important or popular species over a wide range. Indeed, the totality of all biological effects may not be decisive: final decisions will also depend on an array of economic, legal, political, and social goals and requirements, many of them unwritten.

Even for a single variable, the analysis of the monitoring data gives only partial information. Decision makers are usually choosing among many possible actions, but the monitoring data apply directly only to some of these: e.g., "do nothing," "impose a penalty," and "shut it down." (Even this assumes that shutting it down would return the environment to the "Before" condition.) These data may be indirectly informative about the consequences of other actions, such as redesign, but usually only when supplemented by other information: theory, modeling, experiments, and general biological knowledge.

Even if a decision were to be made entirely on the basis of the biological monitoring data, it would be too complicated a function of them to be specified in advance. With n parameters of concern (e.g., the changes in the mean abundances of n species), the data summary is likely to contain at least $2n$ values (e.g., the confidence bounds, or estimates and P-values, for the n changes). Thus, a decision procedure would need to divide the $2n$-dimensional space of possible data summaries into subregions corresponding to the different possible actions. But real decisions will involve far more than the monitoring data, including some factors which, while knowable (e.g., models of the future of the local economy from various starting points), would not be worth determining until we know they are needed, and probably other factors which we cannot anticipate, since decision makers cannot be expected to specify every possible contingency, and their corresponding decisions, in advance. Thus the aim of monitoring should not be decisions but conclusions: succinct descriptions of the biological effects, with the allowances to be made for uncertainty, in as clear a form as possible.

Hypothesis Tests: Meager Information and Unnecessary Confusion

Assessment decisions will rarely depend on the existence of an effect. Almost any intervention big enough to be worth studying will have effects on most of the local environmental parameters studied, whether we "detect" them or not. Knowing an effect exists is useless for decision making: it is the direction and size that matter. Even this question may be too narrow: some effects might be positive under some conditions (e.g., in winter, or when currents flow north) and negative under others (Eberhardt 1976, p. 34, Murdoch et al. 1989, p. 94, Reitzel et al. 1994). Neyman and Pearson's conviction/acquittal analogy is false here: the question is not "is he guilty?" but "what is he guilty of?" and (for decision makers) "what should be the sentence?"

The null hypothesis of "no change" is a straw man, and "detection" of changes is irrelevant. "Nonsignificance" or "failure to detect" an effect means merely that our data or analyses are insufficient to allow us to make an assertion about the change's direction, at a significance level of no demonstrated relevance. It does not mean we have no information: the evidence may point to a large change, but be highly uncertain when taken in isolation. To report it only as "NS" on the basis of the 0.05 cutoff is to engage in self-censorship. "Detection" or a P-value is better but still inadequate; it conveys information about direction, but not about size.

Thus hypothesis testing provides too little information for most decision making. At its best (the P-value), it uses an implausible model of "no effect" to compute the probability of observing data more unfavorable to this model than ours are. An accept/reject "decision" conveys even less. An estimate is far more informative, but a test result and even a power curve (or power evaluated at some arbitrary alternative) adds virtually nothing to it. (That is, nothing directly; from the estimate and the P-value, one can often compute a measure of the estimate's reliability, such as a standard deviation or confidence interval—but this justification applies also to handing over the raw data, unanalyzed.)

In addition, test results are misleading or confusing for many people. Hypothesis testing is awash in jargon, and its logic and the meaning of its results are not simple. Berger and Sellke (1987) argue that the P-value "gives a very misleading impression as to the validity of H_0, from almost any evidentiary viewpoint," mainly by showing how different it is from any reasonable calculation of $\Pr(H_0|x)$, the conditional probability of H_0 given the data x. They justify this by claiming that "Most nonspecialists interpret (the P-value) precisely as $\Pr(H_0|x)$"—an unproven claim, but many statisticians believe it. Perhaps a more striking justification is that Neyman himself once made this error (Good 1984).

A far more damaging misinterpretation is that a "significant" result (or a small P-value) indicates a large, important effect, while a "nonsignificant" effect is nonexistent or unimportant (Yoccoz 1991). This also seems common to "most nonspecialists"—and to specialists not on their toes. Indeed, it is unclear why a test of "no effect" would be proposed for decision making in impact assessment except on the basis of this error.

This confusion can do practical damage. For example, the California Ocean Plan forbids "significant declines in light transmittance," and defines "significant" to mean "statistically significant at the 95% level." (This is interpreted to correspond to a 0.05 level test.) But a large enough monitoring program will eventually find a "significant" change, and the law seems to attribute no relevance to the question of its *biological* significance. A power company or municipality can best satisfy the law by restricting the monitoring program (e.g., on the grounds of expense) so as to ensure a high level of uncertainty and low power of "detection" (Mapstone, Chapter 5).

When many possible effects are studied, multiple testing adds additional confusion. Mead (1988) presents published examples of results, which biologically motivated plots would have made beautifully clear, rendered incomprehensible by multiple testing methods. In assessment, it has been claimed that not only should estimated changes be statistically significant to be taken seriously, but also the *number* of statistically significant changes should itself be statistically significant (Patton 1991).

These artificial complexities are harmful. Managers, nonscientist review boards, and some investigators are very likely to misunderstand the meaning of test results, especially confusing statistical and practical significance. Also, tests focus attention too strongly on only part of the evidence. No standard formal method provides a complete assessment of the reliability and practical significance of a field result. All are affected by model uncertainty, and the multiple testing problem is real, even though the methods are usually unhelpful. Biological understanding is needed to relate the conclusions to each other, to auxiliary experiments and observations, and to mechanisms and processes which are known or plausible consequences either of the intervention or of alternative explanations. Struggling with a variety of tests and their interpretations is not the best use of biologists' time and skills.

Confidence Intervals: Quantitative Conclusions

Final reports must present both descriptions of changes and measures of the uncertainty of these descriptions. One way to do this is to focus on parameter estimation, especially confidence intervals:

> The greatest ultimate importance, among all types of statistical procedures we now know, belongs to *confidence procedures* which, by making interval estimates, attempt to reach as strong conclusions as are reasonable by pointing out ... whole classes (intervals, regions, etc.) of *possible* values, so chosen that there can be high confidence that the 'true' value is *somewhere among them.* Such procedures are clearly quantitative conclusion procedures. They make clear the essential 'smudginess' of experimental knowledge (Tukey 1960).

These descriptions must be easily understood by decision makers who are not trained in statistics. This is especially important if changes are expected to vary with seasons or other environmental conditions, so that estimated changes will have both varying values and varying uncertainties. Confidence intervals satisfy this too. Complex results can be presented clearly, without oversimplifying, to an audience of nonscientist decision makers: parameter estimates and confidence regions not only have obvious relevance to decisions but also are natural candidates for graphing. Even "power" is portrayed, in the length of the interval rather than in a welter of α's, β's, Δ's and noncentral *t* and *F* distributions.

Estimates of direction and size are needed for two reasons. One is obvious: these are the main determinants of whether an effect is harmful. The other has to do with the reliability of conclusions. The assessment of a given field result should consider not only the formal statistical summary of the "internal" evidence of the data directly pertaining to it, but also the "external" evidence of the agreement of the conclusion with our understanding of the mechanisms involved, and with other data or conclusions from the study (e.g., concerning changes in similar species). This is particularly the case when the conclusion concerns not only whether a given change has occurred but also whether the intervention caused it, i.e., whether it is an "effect." For these external judgments, the estimated sizes of changes, when combined with measures of these estimates' reliability, are more useful summaries of the internal evidence than are measures of how strongly it indicates the changes' existence (Hill 1965).

A final advantage of presenting confidence intervals and regions rather than hypothesis tests can only be outlined here. If impact assessment is seen as a statistical decision problem, it is difficult to avoid Bayesian formulations and solutions, at least as an ideal (Pratt et al. 1965a, 1965b). This ideal may be unattainable. It often needs detailed specification and quantification of all aspects of the problem: "states of nature" (possible impacts—but also, ultimately, economic, political, and aesthetic parameters), possible actions, the cost or gain function, and "personal" prior distributions on the states. The last risks having debates over assessment degenerate into arguments about prior distributions and the qualifications of their proponents. At present, it seems safer to present "objective" assessments of uncertainty, based on the field observations, separately from assessments based on compatibility with prior information, other results, and auxiliary experiments. However, if the Bayesian logic is accepted, then a non-Bayesian approach adopted for practical reasons should approximate it as well as possible. In fact, confidence intervals based on approximately Normal point estimates do approximate their Bayesian equivalents (posterior probability or "credible" intervals), at least for the diffuse priors one would expect when many poorly known processes are at work, while hypothesis tests do not approximate theirs (Edwards et al. 1963, Pratt 1965, Lindley 1965, Berger and Sellke 1987, Casella and Berger 1987).

Discussion

I have argued that confidence intervals are preferable to hypothesis tests of "no effect" because they directly assess the main concern (effect size), are easy to understand, display "power" automatically, are more relevant to an overall or causal assessment, and correspond reasonably well to the Bayesian ideal.

Some counter arguments should be noted. Perhaps the weakest is the unfounded claim that "the majority" prefers tests. This is a surprising argument for scientists to use: the majority once preferred phlogiston, and thought Copernicus was wrong. It is a poor reason for using a bad procedure, especially since many users do *not* understand the tests they use.

Tests may be more familiar to regulators, who frequently test whether a pollutant is below a threshold. However, this does address effect size: the threshold is not usually zero but a level judged important. The context is also different. Regulation frequently involves ongoing, routine judgments of many pollution sources on the basis of a single variable. Assessment requires a one-time (or few times) judgment of a single project on the basis of many variables. Even so, with moves toward sale of pollution "credits," hypothesis tests may give way to confidence intervals in regulation, too.

There are cases where the law requires hypothesis tests, as for the California Ocean plan discussed above. The tests must then be done. But these laws are written with statistical advice, some of it bad: they can and should be changed with better advice.

Some developers would like decision rules laid out in advance: they don't want the rules to change after the investments have been made. But we all want things we can't have. Sensible decisions will be complex functions of biological and economic data and models, and of factors wé can't predict, and we mislead clients by promising what we can't deliver. It may be possible to promise something weaker: e.g., regulators could make a list of key species and promise that, provided none of their abundances has decreased by more than 30%, these species will not be used as the basis for some of the more dramatic possible decisions, such as shutdown or radical redesign. (These species might still be the basis for milder decisions, and the more dramatic decisions might still arise because of other factors.) A small part of the decision-making process could then consist of testing "H_0: Species X has not declined by more than 30%," at the 50% level for each species. This has equal risk at the boundary for developers and for Species X and, assuming abundance estimates have approximately symmetric distributions, is easy to carry out and understand since no variance estimates are needed: H_0 is rejected if, and only if, the estimated decline is >30%. (We would still, presumably, need to describe the estimation method in advance.) With n species all reduced by 30%, the chance of no rejections is only $1/2^n$ for independent estimates; but it would increase with smaller reductions and more accurate estimates, and rejection alone would trigger no penalties, only less restricted decision making.

It can be argued that all this involves no change at all: confidence intervals and tests are interchangeable, the confidence interval containing all values not rejected by the test; 95% is just as arbitrary as 0.05; and one could also test subsets of the data, e.g., winter results only, and convert these to confidence intervals. These arguments are not completely true. The most useful tests may be those for goodness-of-fit (Box 1980), which are not usually invertible. (Nor is the standard test for equality of two Binomial probabilities.) A test demonstrating a change in the mean of a transformed variable, such as $\sqrt{}$(Impact site abundance + 0.5) −

$\surd$(Control site abundance + 0.5), is hard to convert into a confidence interval for the change at the Impact site: confidence intervals force one to focus on the more meaningful parameters (Bence et al., Chapter 8). Even when they are mathematically equivalent, tests and confidence intervals do not convey equivalent messages. A *P*-value for "no effect" cannot be converted into a confidence interval for effect size unless the size estimate is given. Many readers will not make the conversion, assuming that the *P*-value summarizes the information, so test results are misleading if confidence intervals (or something similar) are needed. The arbitrariness of 95% is a minor matter: other confidences could be indicated simultaneously on plots. In short, estimates and confidence intervals give the needed information in a usable form; hypothesis tests do not.

Acknowledgments

This chapter owes a great deal to suggestions from J. Bence, M. Keough, R. Schmitt, and especially C. Osenberg. This work was supported in part by the Minerals Management Service, U.S. Department of the Interior, under MMS Agreement No. 14-35-0001-30471 (The Southern California Educational Initiative). The views and interpretation contained in this document are those of the author and should not be interpreted as necessarily representing the official policies, either express or implied, of the U.S. Government.

References

Berger, J. O., and T. Sellke. 1987. Testing a point null hypothesis: The irreconcilability of *P*-values and evidence (with Discussion). Journal of the American Statistical Association **82**:112–139.

Box, G. E. P. 1980. Sampling and Bayes inference in scientific modeling and robustness (with discussion). Journal of the Royal Statistical Society **A143**:383–430.

Carney, R. S. 1987. A review of study designs for the detection of long-term environmental effects of offshore petroleum activities. Pages 651–696 in D. F. Boesch and N. N. Rabalais, editors. Long-term environmental effects of offshore oil and gas development. Elsevier, New York, New York.

Casella, G., and R. L. Berger. 1987. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. Journal of the American Statistical Association **82**:106–111.

Eberhardt, L. L. 1976. Quantitative ecology and impact assessment. Journal of Environmental Management **4**:27–70.

Edwards, W., H. Lindman, and L. J. Savage. 1963. Bayesian statistical inference for psychological research. Psychological Review **70**:193–242.

Fisher, R. A. 1937. The fiducial argument in statistical inference. Annals of Eugenics **6**:391–398.

Fisher, R. A. 1956. Statistical methods and scientific inference. Oliver and Boyd, Edinburgh.

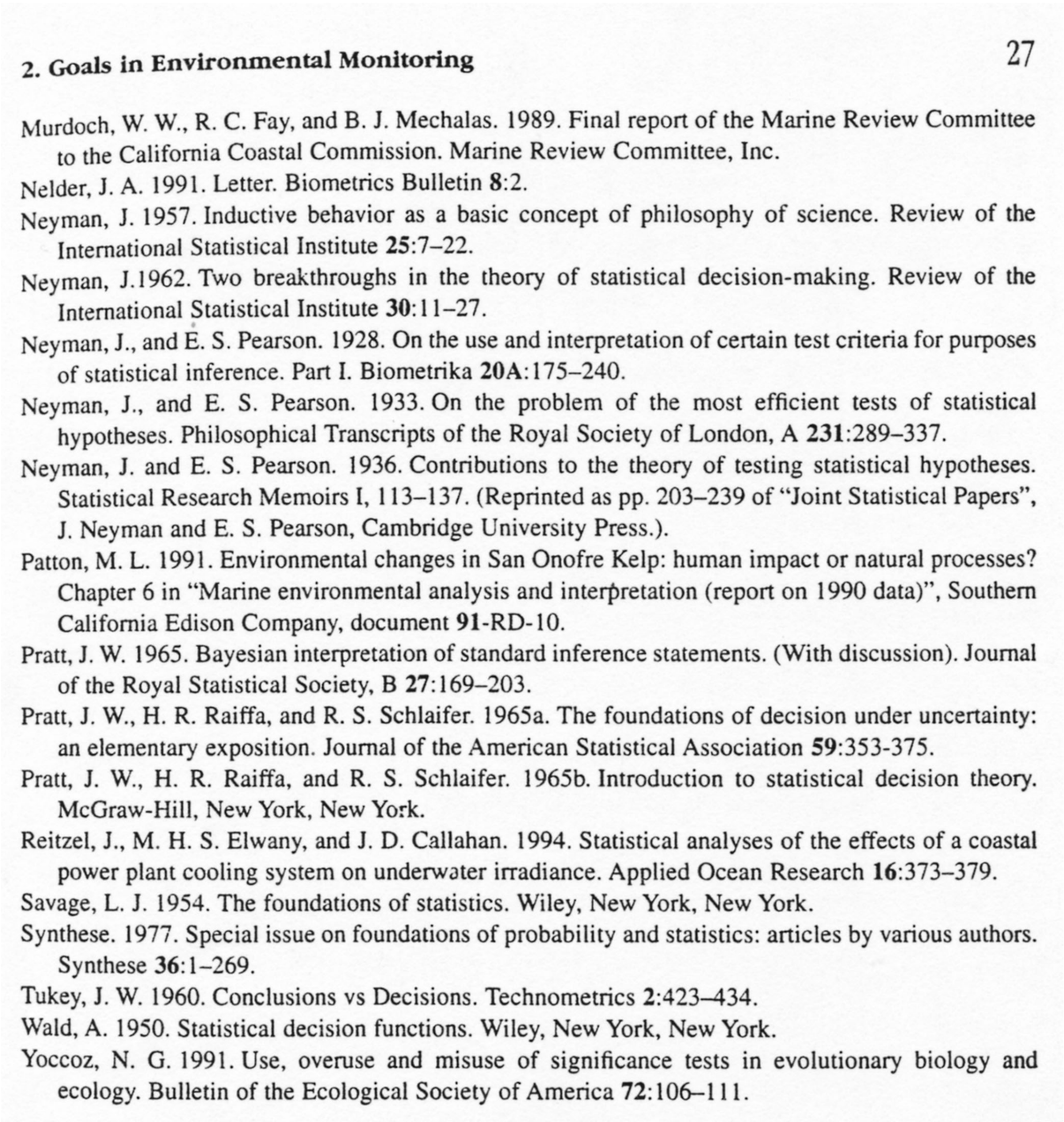
Good, I. J. 1984. An error by Neyman noticed by Dickey. Journal of Statistical Computation and Simulation **20**:159–160.

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

Hill, A. B. 1965. The environment and disease: association or causation. Proceedings of the Royal Society of Medicine **58**:295–300.

Lindley, D. V. 1965. Discussion of Pratt (1965). Journal of the Royal Statistical Society, B **27**:192–193.

Mead, R. 1988. The design of experiments. Cambridge University Press, Cambridge, England.



Murdoch, W. W., R. C. Fay, and B. J. Mechalas. 1989. Final report of the Marine Review Committee to the California Coastal Commission. Marine Review Committee, Inc.

Nelder, J. A. 1991. Letter. Biometrics Bulletin **8**:2.

Neyman, J. 1957. Inductive behavior as a basic concept of philosophy of science. Review of the International Statistical Institute **25**:7–22.

Neyman, J.1962. Two breakthroughs in the theory of statistical decision-making. Review of the International Statistical Institute **30**:11–27.

Neyman, J., and E. S. Pearson. 1928. On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. Biometrika **20A**:175–240.

Neyman, J., and E. S. Pearson. 1933. On the problem of the most efficient tests of statistical hypotheses. Philosophical Transcripts of the Royal Society of London, A **231**:289–337.

Neyman, J. and E. S. Pearson. 1936. Contributions to the theory of testing statistical hypotheses. Statistical Research Memoirs I, 113–137. (Reprinted as pp. 203–239 of "Joint Statistical Papers", J. Neyman and E. S. Pearson, Cambridge University Press.).

Patton, M. L. 1991. Environmental changes in San Onofre Kelp: human impact or natural processes? Chapter 6 in "Marine environmental analysis and interpretation (report on 1990 data)", Southern California Edison Company, document **91**-RD-10.

Pratt, J. W. 1965. Bayesian interpretation of standard inference statements. (With discussion). Journal of the Royal Statistical Society, B **27**:169–203.

Pratt, J. W., H. R. Raiffa, and R. S. Schlaifer. 1965a. The foundations of decision under uncertainty: an elementary exposition. Journal of the American Statistical Association **59**:353-375.

Pratt, J. W., H. R. Raiffa, and R. S. Schlaifer. 1965b. Introduction to statistical decision theory. McGraw-Hill, New York, New York.

Reitzel, J., M. H. S. Elwany, and J. D. Callahan. 1994. Statistical analyses of the effects of a coastal power plant cooling system on underwater irradiance. Applied Ocean Research **16**:373–379.

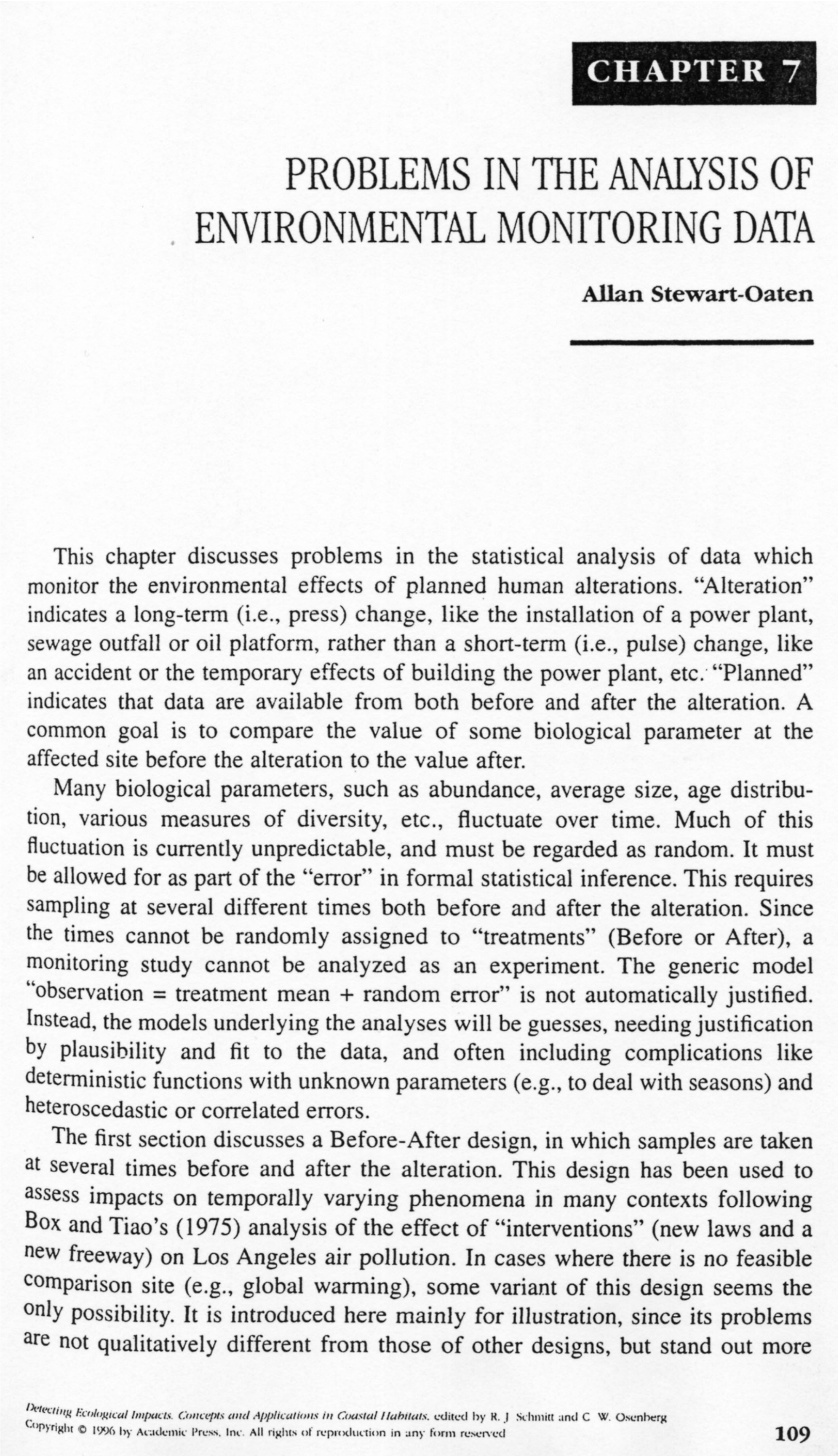
Savage, L. J. 1954. The foundations of statistics. Wiley, New York, New York.

Synthese. 1977. Special issue on foundations of probability and statistics: articles by various authors. Synthese **36**:1–269.

Tukey, J. W. 1960. Conclusions vs Decisions. Technometrics **2**:423–434.

Wald, A. 1950. Statistical decision functions. Wiley, New York, New York.

Yoccoz, N. G. 1991. Use, overuse and misuse of significance tests in evolutionary biology and ecology. Bulletin of the Ecological Society of America **72**:106–111.



CHAPTER 7

PROBLEMS IN THE ANALYSIS OF ENVIRONMENTAL MONITORING DATA

Allan Stewart-Oaten

This chapter discusses problems in the statistical analysis of data which monitor the environmental effects of planned human alterations. “Alteration” indicates a long-term (i.e., press) change, like the installation of a power plant, sewage outfall or oil platform, rather than a short-term (i.e., pulse) change, like an accident or the temporary effects of building the power plant, etc. “Planned” indicates that data are available from both before and after the alteration. A common goal is to compare the value of some biological parameter at the affected site before the alteration to the value after.

Many biological parameters, such as abundance, average size, age distribution, various measures of diversity, etc., fluctuate over time. Much of this fluctuation is currently unpredictable, and must be regarded as random. It must be allowed for as part of the “error” in formal statistical inference. This requires sampling at several different times both before and after the alteration. Since the times cannot be randomly assigned to “treatments” (Before or After), a monitoring study cannot be analyzed as an experiment. The generic model “observation = treatment mean + random error” is not automatically justified. Instead, the models underlying the analyses will be guesses, needing justification by plausibility and fit to the data, and often including complications like deterministic functions with unknown parameters (e.g., to deal with seasons) and heteroscedastic or correlated errors.

The first section discusses a Before-After design, in which samples are taken at several times before and after the alteration. This design has been used to assess impacts on temporally varying phenomena in many contexts following Box and Tiao’s (1975) analysis of the effect of “interventions” (new laws and a new freeway) on Los Angeles air pollution. In cases where there is no feasible comparison site (e.g., global warming), some variant of this design seems the only possibility. It is introduced here mainly for illustration, since its problems are not qualitatively different from those of other designs, but stand out more

Detecting Ecological Impacts. Concepts and Applications in Coastal Habitats, edited by R. J. Schmitt and C. W. Osenberg

clearly. Even to define the parameters describing biological or ecological change between one period and another requires a model of a stochastic process. For some variables of interest, e.g., abundance, such models may involve strong temporal fluctuations. Deterministic fluctuations will lead to bias unless model forms can be guessed approximately correctly. Random fluctuations are likely to have significant long-term serial correlation of unknown structure, which can cause estimates of effects to have large variances and these variances to be badly underestimated from the data.

The second section outlines how the use of a Control site as a covariate may make acceptable effect estimates, and variance estimates, possible. In this "BACIPS" (Before-After-Control-Impact Paired Series) design, data are taken "simultaneously" at one or more Impact sites near the alteration and at one or more Control areas, nearby and similar but far enough from the alteration to be little affected by it, on sequences of sampling occasions Before and After the alteration. This general design has also been called a "multiple time series quasi-experiment" (Campbell and Stanley 1966), "pseudo-experiment" or "pseudodesign" (Eberhardt 1976), "Control-Treatment Pairs (CTP)" (Skalski and McKenzie 1982), and "BACI" (Stewart-Oaten et al. 1986). The idea is that suitable Controls will "track" the Impact sites in some sense. A change in this tracking relationship following the alteration will be evidence for an effect.

However, this analysis depends on how the tracking relationship is modeled. Some alternative models are introduced and briefly discussed. It is argued that analysis using more than one model may be needed, with necessarily rough ways of checking the compatibility of their conclusions. Some additional problems of basing assessment on parameters other than the mean (or median), such as variances, are discussed.

The third section discusses some approaches to causal assessment.

This chapter is guided by Tukey's (1962) dictum: "Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise". Its purpose is not to disparage formal methods, but to argue that the right questions in impact assessment are likely to have only approximate, messy, and possibly multiple answers, requiring informal combinations of formal results.

Before-After Studies

Defining Parameters

Suppose we are interested in how the abundance of a given species has changed following the alteration. Let $N_B(t)$ be the abundance at the Impact site at time t during the period Before the alteration was installed (or began operating), and $N_A(t)$ the abundance at time t during the period After installation. These are *true* abundances, assumed known exactly over a Before period $T_s < t < T_0$ and an After period $T_0 < T_E$, respectively. Sampling error is an additional complication, but is distracting at this point.

We might say that "the" abundance has decreased if $N_A(t)$ is in some sense smaller than $N_B(t)$. But neither of these is a simple number: both are functions of t. Nor can these two functions be compared in the usual way, by asking whether $N_B(t) > N_A(t)$ for each time, t, or by comparing their averages over some period. There is no time for which both functions exist: $N_B(t)$ exists only for $t < T_0$, the installation time, and $N_A(t)$ only for $t > T_0$.

We could define the function $N_A(t)$ to be smaller than $N_B(t)$ if the average of $N_A(t)$ over the After period ($T_0 < t < T_E$) is smaller than the average of $N_B(t)$ over the Before period. One objection to this is that it depends on the periods chosen. If the After period's "Winter" fraction is larger than the Before period's, then we might have $N_A(t) < N_B(t)$ by this definition, even though $N_A(t)$ may have a greater "Winter" average *and* a greater "Summer" average than $N_B(t)$. Matching seasons or using a weighted average might solve this problem (Cochran and Rubin 1974).

A second objection is that both $N_B(t)$ and $N_A(t)$ are determined partly by "random" factors such as births, deaths, and movements by individuals, invasions by predators, competitors or disease, short-term events like storms and upwellings, etc. It may be that $N_A(t) < N_B(t)$ by this definition, only because of random factors which had nothing to do with the alteration. This suggests that judgment should be based on some kind of average or distribution of what *could have* happened, rather than directly on what did happen.

This conclusion especially applies if the consequences of the decision to be made will depend on future abundances, rather than past ones, as in decisions about ceasing operations, modifying designs, or compensatory mitigation. In these cases, the abundances up to the time of the decision are useful mainly as guides to future abundances —i.e., to their probability distributions. Even decisions concerning punishment or reparations for damage already done require a comparison between what has happened following the alteration, i.e., $N_A(t)$ for $T_0 < t < T_E$, and what *would have happened* had the alteration not occurred, i.e., the distribution of possibilities for $N_B(t)$ for the same period.

Time-Series Modeling

To define distributions of what the abundances could have been, or could be in future, we need to regard $N_B(t)$ and $N_A(t)$ as stochastic processes. Such processes can be modeled by giving a formula from which, given past values of a function and also the values of a collection of random variables generated independently in a specified way, all future values of the function could be determined. For example, the abundance of an annual population in a constant environment might be described by $N(t) = r_t N(t-1)/(1 + cN(t-1))$, where the r_t's are independent draws from some specified distribution. Given $N(0)$ and the set of random values $r_1, r_2, \ldots$, the entire process could be calculated.

Given the past values of a process, we can simulate a possible future on a computer, by using a random number generator and applying the formula. This future will depend in part on the chance values generated (e.g., the r_t's). We could then simulate another future, using the same past and formula, but a new set of numbers from the same generator. Each future constitutes a "realization" of the process: each is a function of time. For any time, t, in the future, and any number x, we can determine the probability that a realization generated in this way will have a value $\leq x$ at time t, i.e., $P\{N(t) \leq x\}$. For instance, we could simulate a large number of futures and count the fraction with this property. These probabilities give the distribution of the process at time t. Similarly, for any two times, t_1 and t_2, we can determine $P\{N(t_1) \leq x_1$ and $N(t_2) \leq x_2\}$ for any x_1 and x_2, and thus the joint distribution for these two times. From these distributions, means, variances and covariances can be obtained, all functions of time (or of two times). Similar probabilities can be computed for any finite set of future times; the collection of all such probabilities gives the distribution of the entire process.

Thus we wish to compare the distributions of the Before and After processes, N_B and N_A, generated by possibly different formulae and random number generators, using a single partial realization of each, $N_B(t)$ $(T_S < t < T_0)$ and $N_A(t)$ $(T_0 < t < T_E)$.

Focus on the distribution circumvents the problem that the realizations, $N_B(t)$ and $N_A(t)$, are never observed at the same time, t. Even though the alteration prevents any actual realizations from occurring, there is a distribution of possible realizations of $N_B(t)$ for $t > T_0$, because the distribution depends only on the past, the appropriate formula, and the distributions of the random variables. Thus the distributions, or key parameters like means and variances, could be compared at any time t, even though realizations cannot be. The effect of the alteration (or the change coincident with the alteration) could be defined as the difference between the means of $N_A(t)$ and $N_B(t)$. If no other causes are operating, this is the difference between the mean abundance obtained with the alteration and the mean that *would have been* obtained had the alteration not occurred.

Estimating a Varying Mean

Unfortunately, the distributions, or their parameters, are still functions of time. For example, given a "history" of past values, H, at time t_0 say, the mean of $N_B(t)$ at time t is

$$M_B(t,H,t_0) = E\{N_B(t) \mid H,t_0\}, \qquad (1)$$

the mean of the values at time t of all possible realizations beginning from a history H at time t_0. This is *not* an average over time: t, H and t_0 are all fixed. We are averaging over the possible values of $N_B(t)$, each possible value corresponding to a set of possible values of the random variables involved in the algorithm (e.g., r_1, r_2, ..., in the example above). Similarly we have $M_A(t,H,t_0)$, the variance functions $V_B(t,H,t_0)$ and $V_A(t,H,t_0)$, and the covariance functions $C_B(t_1,t_2,H,t_0) = \text{Cov}\{N_B(t_1),N_B(t_2) \mid H,t_0\}$ and $C_A(t_1,t_2,H,t_0)$. These functions are not known because the algorithms (the formulae and the distributions of the random variables) for generating realizations are not known. They must be partly guessed, guided by biological knowledge and intuition, tractability and flexibility, and partly estimated from the data (i.e., from the observed realizations).

Some judgment is unavoidable. To make any inferences at all, or even to produce meaningful descriptive summaries of the data, some assumptions are necessary so that observations taken at different times can be combined to estimate parameters relevant to all times.

Modeling the Mean Function. One possibility is to model the mean functions, M_B and M_A, as explicit functions of time, which are known except for a small number of time-independent parameters, to be estimated from the data and other information. If time is measured in years, we might assume

$$M_B(t,H,t_0) = \mu_B + \alpha_B \sin 2\pi t + \beta_B \cos 2\pi t + h(t,H,t_0), \qquad (2)$$

where $h \rightarrow 0$ as $t - t_0$ increases (the effects of past history die away). Assuming the process began far in the past, M_B oscillates sinusoidally (seasonally) about a fixed value. The effect of the alteration could be defined in terms of μ_B, α_B, and β_B and the corresponding After parameters, with all six estimated from the data.

An immediate problem is that this functional form may be wrong. The mean of the process may not be smoothly sinusoidal. Our estimates of the alteration's effects will then be biased. Alternative forms are available, e.g., replacing $\alpha_B \sin 2\pi t + \beta_B \cos 2\pi t$ by a polynomial or by separate fixed means for each "season" (defined by the biology, not the calendar), but these may be wrong too.

However, the greater problems are likely to be estimators with very high variances, and underestimation of these variances. These arise because of large, long-lasting fluctuations that are not allowed for in the mean functions, so must be treated as random deviations from them. These fluctuations can be caused by major environmental events, like El Niño or a large storm, or by biological events unrelated to the alteration, like an epidemic or a predator–prey cycle. Such a fluctuation may last through a significant part of the sampling program; if it does, then the observed abundances, $N(t_i)$, will be "serially correlated" (the series is "autocorrelated"): several of them, especially those close in time, will have essentially the same random deviation from the mean, so (i) this deviation does not "cancel out" in the average, and (ii) these observations will not vary much, so the true variance will be underestimated. Thus, effect estimates may be very unreliable (containing a large chance component), but may seem to be reliable.

One way to deal with this problem is to try to model major periodic phenomena deterministically as part of the mean function. Calling something "random" rather than "deterministic" is often rather arbitrary, based on what we are unable to predict in our present state of knowledge more than on what is inherently unpredictable. However, unless the phenomenon is well understood, this seems likely to complicate interpretation without reducing variance, by introducing additional parameters which explain little of the variation but whose estimates are correlated with the estimates of interest, i.e., of the alteration's effects, such as $\mu_A - \mu_B$ in Equation 2.

Modeling the Errors. An alternative is to write a model for the abundances, e.g.,

$$N(t_i) = M(t_i) + \varepsilon_i, \tag{3}$$

where M is the parametric mean function, and the errors, ε_i, may be correlated. The past history, H, being unknown, can be taken as being far in the past: it is usually plausible that, if H is information about $N(t)$ for $t < t_0$, then H will be irrelevant if t_0 is long enough before the first observed time, t_1. The errors could be modeled in a way that takes specific account of their likely mechanisms, but this is hard to do when there are several sources of error whose mechanisms are little known. Generic models, especially linear models (ARMA models, Box and Jenkins 1976) of the form

$$\varepsilon_i = \Sigma_j b_j \varepsilon_{i-j} + a_i + \Sigma c_j a_{i-j}, \tag{4}$$

where the a_i's are uncorrelated random values and the b's and c's are unknown constants, are usually preferred. This class of models can allow for trends or seasonal patterns (so that "$\alpha_B \sin 2\pi t + \beta_B \cos 2\pi t$" can be omitted from Equation 2), by focusing on differences of successive observations, or observations a year apart, and allowing the current error to depend directly on the error a year earlier.

However, these linear models may not fit long-term phenomena like El Niño well: e.g., a recent pattern of declining abundances may indicate the beginning of an El Niño, and thus foretell a continuing decline followed by a long period of low abundance, but other patterns of recent abundances may be uninformative. Patterns of this kind may be better described by models with long-range dependence, e.g., with correlations that decay slower than exponentially over time (Beran 1992).

Attributing seasonal patterns to errors is mainly for series with no clear physical mechanism for seasonality, and better suited to forecasting than to estimation. Without differencing, the forecasts have exponentially declining seasonal patterns. The use of differences implies "homogeneous" behavior, independent of the current level of the process: there is no tendency to return to a mean that is a periodic function of time, like Equation 2—i.e., no "density-dependent" regulation, so a time series made up of yearly averages would be "nonstationary". This seems unlikely for abundances.

Finally, these models imply constant variances and correlations that depend only on the number of intervening samples; this seems doubtful with unevenly spaced observations, or if disturbance is greater at some times of the year, but the analytical consequences may be mild (cf. Stigler 1976).

Using Covariates. The typically small numbers and span of sampling times in impact assessment data will make it difficult to carry out either the mean function or the ARMA approach. Both introduce new parameters to estimate. Even if a simple assumed linear form is correct, variances and covariances may still be underestimated, because the series is so short that the variance of the average of the errors, i.e., $V\{\sum e_i/n\}$, is not negligible compared to the variance of a single observation, i.e., $V\{e_i\}$ (see Priestley 1981, Equation 5.3.12). Above all, even when these problems are minor or resolved, the main achievement of such models will be realistic estimates of the variances of our estimates of the alteration's effects. The models do relatively little to reduce the variances. Even if the correct error model were given to us, the estimated effects would often still have variances too large for practical use.

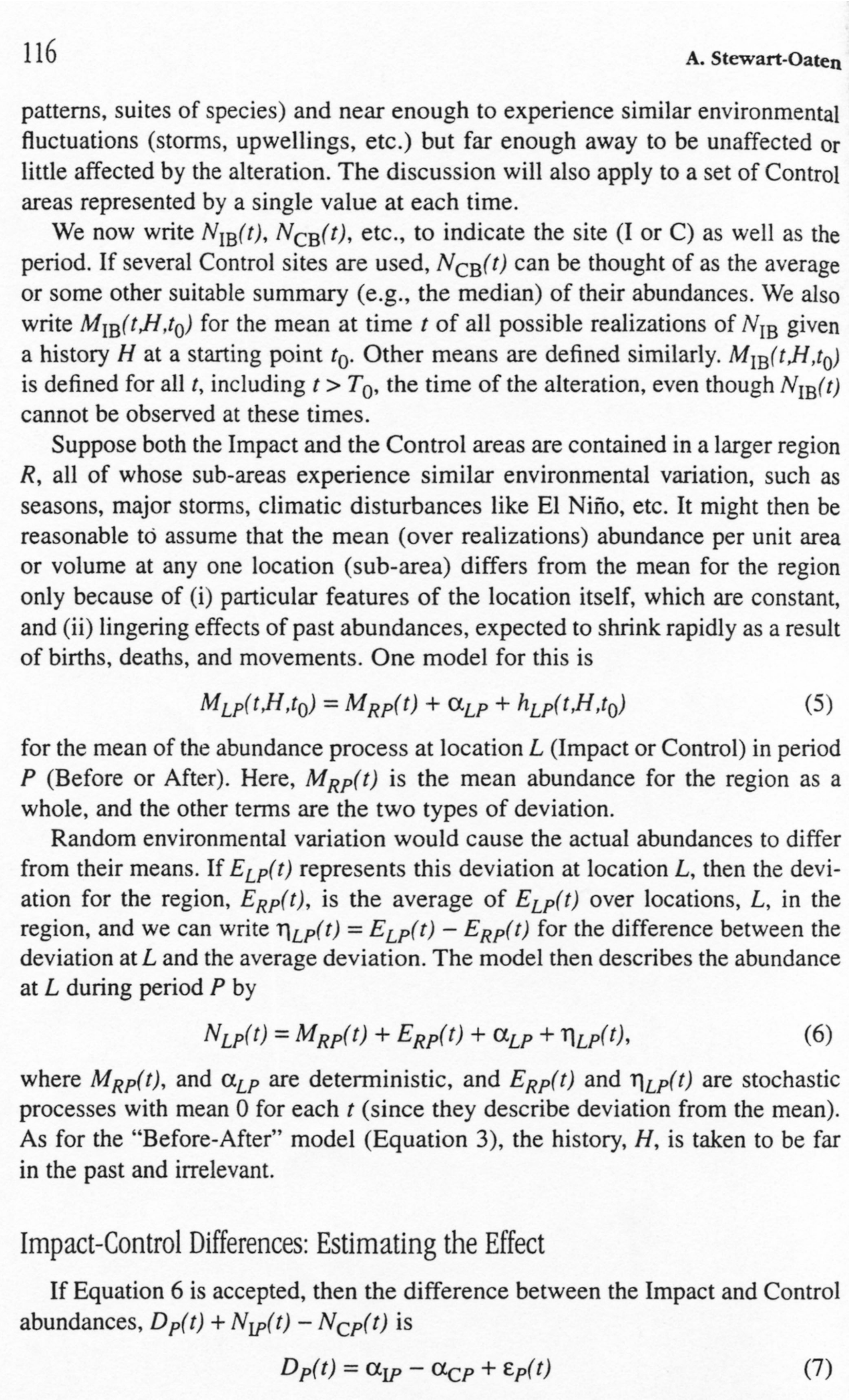
A third approach can potentially both reduce the variances of estimated effects and estimate these variances accurately. This is to include in the model other observable variables which are affected by the natural fluctuations but not affected by the alteration. These "covariates" can be used to estimate the contribution of the natural fluctuations to the abundance. By removing this contribution, we obtain a "corrected" or "adjusted" abundance which has smaller temporal variance and smaller serial correlation than the raw abundance but is equally affected by the alteration. We can estimate what the abundance would have been under "standard" conditions, and estimate the effect of the alteration by the change it would cause under these conditions. This approach can also reduce deterministic bias: e.g., a covariate like water temperature may be a better indicator of seasonal variation than the time of year itself.

This approach also has difficulties. It requires a model for estimating the natural fluctuations on the basis of the covariates. Most commonly, some form of regression of the observed values (the abundances) against the covariates is used, possibly with either or both being transformed first. If the model form is wrong, estimated alteration effects are likely to be biased. It also requires that the covariates be reasonably good indicators of the natural fluctuations. The covariate model will include some additional parameters to be estimated, thus reducing the information available for estimating effects. If these parameters have little explanatory value, the variances of our effect estimates may actually increase, and our estimates of these variances will become more complicated but no more accurate. This can occur either if the covariate is not strongly correlated with the natural fluctuation or if it is observed with substantial error.

Before-After-Control-Impact Paired Series Designs

Impact-Control Differences: Model

In many cases, the most effective covariate for the abundance at the Impact site is likely to be the simultaneous abundance at a Control site. This is not an experimental control, since treatments are not assigned, randomly or otherwise, by the investigator. In this discussion, a Control is an area which is similar to the Impact area in features judged to be important (e.g., depth, topography, current



patterns, suites of species) and near enough to experience similar environmental fluctuations (storms, upwellings, etc.) but far enough away to be unaffected or little affected by the alteration. The discussion will also apply to a set of Control areas represented by a single value at each time.

We now write $N_{\mathrm{IB}}(t)$, $N_{\mathrm{CB}}(t)$, etc., to indicate the site (I or C) as well as the period. If several Control sites are used, $N_{\mathrm{CB}}(t)$ can be thought of as the average or some other suitable summary (e.g., the median) of their abundances. We also write $M_{\mathrm{IB}}(t,H,t_0)$ for the mean at time t of all possible realizations of N_{IB} given a history H at a starting point t_0. Other means are defined similarly. $M_{\mathrm{IB}}(t,H,t_0)$ is defined for all t, including $t > T_0$, the time of the alteration, even though $N_{\mathrm{IB}}(t)$ cannot be observed at these times.

Suppose both the Impact and the Control areas are contained in a larger region R, all of whose sub-areas experience similar environmental variation, such as seasons, major storms, climatic disturbances like El Niño, etc. It might then be reasonable to assume that the mean (over realizations) abundance per unit area or volume at any one location (sub-area) differs from the mean for the region only because of (i) particular features of the location itself, which are constant, and (ii) lingering effects of past abundances, expected to shrink rapidly as a result of births, deaths, and movements. One model for this is

$$M_{LP}(t,H,t_0) = M_{RP}(t) + \alpha_{LP} + h_{LP}(t,H,t_0) \qquad (5)$$

for the mean of the abundance process at location L (Impact or Control) in period P (Before or After). Here, $M_{RP}(t)$ is the mean abundance for the region as a whole, and the other terms are the two types of deviation.

Random environmental variation would cause the actual abundances to differ from their means. If $E_{LP}(t)$ represents this deviation at location L, then the deviation for the region, $E_{RP}(t)$, is the average of $E_{LP}(t)$ over locations, L, in the region, and we can write $\eta_{LP}(t) = E_{LP}(t) - E_{RP}(t)$ for the difference between the deviation at L and the average deviation. The model then describes the abundance at L during period P by

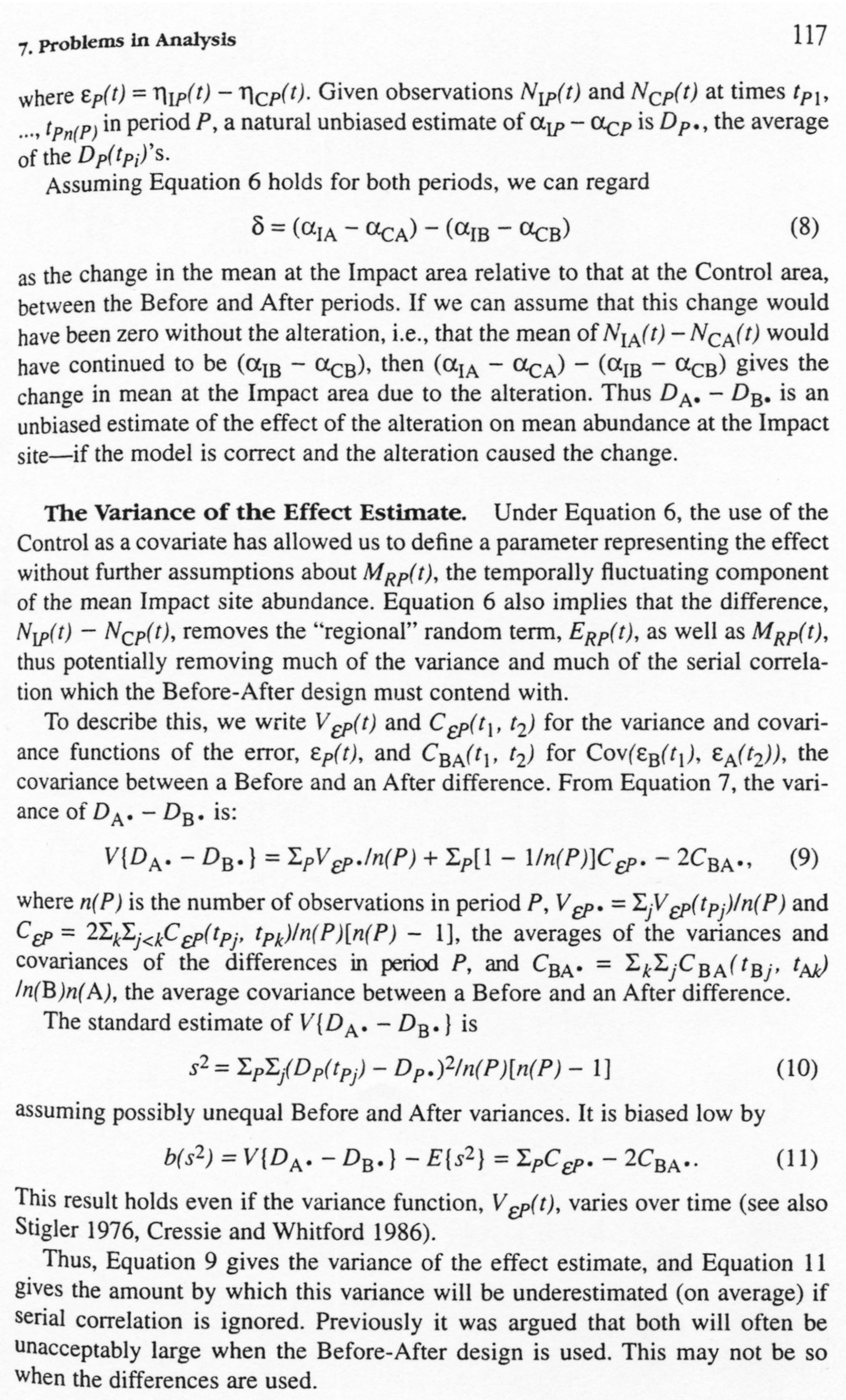
$$N_{LP}(t) = M_{RP}(t) + E_{RP}(t) + \alpha_{LP} + \eta_{LP}(t), \qquad (6)$$

where $M_{RP}(t)$, and α_{LP} are deterministic, and $E_{RP}(t)$ and $\eta_{LP}(t)$ are stochastic processes with mean 0 for each t (since they describe deviation from the mean). As for the "Before-After" model (Equation 3), the history, H, is taken to be far in the past and irrelevant.

Impact-Control Differences: Estimating the Effect

If Equation 6 is accepted, then the difference between the Impact and Control abundances, $D_P(t) + N_{IP}(t) - N_{CP}(t)$ is

$$D_P(t) = \alpha_{IP} - \alpha_{CP} + \varepsilon_P(t) \qquad (7)$$



where $\varepsilon_P(t) = \eta_{\mathrm{I}P}(t) - \eta_{\mathrm{C}P}(t)$. Given observations $N_{\mathrm{I}P}(t)$ and $N_{\mathrm{C}P}(t)$ at times t_{P1}, ..., $t_{Pn(P)}$ in period P, a natural unbiased estimate of $\alpha_{\mathrm{I}P} - \alpha_{\mathrm{C}P}$ is $D_{P\bullet}$, the average of the $D_P(t_{Pi})$'s.

Assuming Equation 6 holds for both periods, we can regard

$$\delta = (\alpha_{\mathrm{IA}} - \alpha_{\mathrm{CA}}) - (\alpha_{\mathrm{IB}} - \alpha_{\mathrm{CB}}) \qquad (8)$$

as the change in the mean at the Impact area relative to that at the Control area, between the Before and After periods. If we can assume that this change would have been zero without the alteration, i.e., that the mean of $N_{\mathrm{IA}}(t) - N_{\mathrm{CA}}(t)$ would have continued to be $(\alpha_{\mathrm{IB}} - \alpha_{\mathrm{CB}})$, then $(\alpha_{\mathrm{IA}} - \alpha_{\mathrm{CA}}) - (\alpha_{\mathrm{IB}} - \alpha_{\mathrm{CB}})$ gives the change in mean at the Impact area due to the alteration. Thus $D_{\mathrm{A}\bullet} - D_{\mathrm{B}\bullet}$ is an unbiased estimate of the effect of the alteration on mean abundance at the Impact site—if the model is correct and the alteration caused the change.

The Variance of the Effect Estimate. Under Equation 6, the use of the Control as a covariate has allowed us to define a parameter representing the effect without further assumptions about $M_{RP}(t)$, the temporally fluctuating component of the mean Impact site abundance. Equation 6 also implies that the difference, $N_{\mathrm{I}P}(t) - N_{\mathrm{C}P}(t)$, removes the "regional" random term, $E_{RP}(t)$, as well as $M_{RP}(t)$, thus potentially removing much of the variance and much of the serial correlation which the Before-After design must contend with.

To describe this, we write $V_{\varepsilon P}(t)$ and $C_{\varepsilon P}(t_1, t_2)$ for the variance and covariance functions of the error, $\varepsilon_P(t)$, and $C_{\mathrm{BA}}(t_1, t_2)$ for $\mathrm{Cov}(\varepsilon_{\mathrm{B}}(t_1), \varepsilon_{\mathrm{A}}(t_2))$, the covariance between a Before and an After difference. From Equation 7, the variance of $D_{\mathrm{A}\bullet} - D_{\mathrm{B}\bullet}$ is:

$$V\{D_{\mathrm{A}\bullet} - D_{\mathrm{B}\bullet}\} = \Sigma_P V_{\varepsilon P\bullet}/n(P) + \Sigma_P[1 - 1/n(P)]C_{\varepsilon P\bullet} - 2C_{\mathrm{BA}\bullet}, \qquad (9)$$

where $n(P)$ is the number of observations in period P, $V_{\varepsilon P\bullet} = \Sigma_j V_{\varepsilon P}(t_{Pj})/n(P)$ and $C_{\varepsilon P} = 2\Sigma_k \Sigma_{j<k} C_{\varepsilon P}(t_{Pj}, t_{Pk})/n(P)[n(P) - 1]$, the averages of the variances and covariances of the differences in period P, and $C_{\mathrm{BA}\bullet} = \Sigma_k \Sigma_j C_{\mathrm{BA}}(t_{\mathrm{B}j}, t_{\mathrm{A}k})/n(\mathrm{B})n(\mathrm{A})$, the average covariance between a Before and an After difference.

The standard estimate of $V\{D_{\mathrm{A}\bullet} - D_{\mathrm{B}\bullet}\}$ is

$$s^2 = \Sigma_P \Sigma_j (D_P(t_{Pj}) - D_{P\bullet})^2/n(P)[n(P) - 1] \qquad (10)$$

assuming possibly unequal Before and After variances. It is biased low by

$$b(s^2) = V\{D_{\mathrm{A}\bullet} - D_{\mathrm{B}\bullet}\} - E\{s^2\} = \Sigma_P C_{\varepsilon P\bullet} - 2C_{\mathrm{BA}\bullet}. \qquad (11)$$

This result holds even if the variance function, $V_{\varepsilon P}(t)$, varies over time (see also Stigler 1976, Cressie and Whitford 1986).

Thus, Equation 9 gives the variance of the effect estimate, and Equation 11 gives the amount by which this variance will be underestimated (on average) if serial correlation is ignored. Previously it was argued that both will often be unacceptably large when the Before-After design is used. This may not be so when the differences are used.

Large fluctuations will often be the result of large-scale disturbances affecting the whole region, so that most of the variance at a site will be due to the "common" regional variation, $E_{RP}(t)$ in Equation 6, which is removed when we take differences. If the Control and Impact sites are not far apart, $\eta_{IP}(t)$ and $\eta_{CP}(t)$, the local deviations from the average regional fluctuation, may be highly correlated, further reducing $V_{\varepsilon P}(t) = V\{\eta_{IP}(t) - \eta_{CP}(t)\}$. These local deviations should be more quickly removed than regional deviations, not only because they are smaller but also because of mixing (of nutrients or planktonic stages) and movement within the region: e.g., a chance increase at a site, unrelated to changes in long-term physical or chemical conditions at the site, should be quickly dissipated to neighboring sites. If so, serial correlation of the differences, $D_P(t)$, will decrease rapidly with time.

Reducing Variance and Bias. Smaller variances and correlations will reduce $V = V\{D_A\bullet - D_B\bullet\}$. The latter will also reduce the bias, $b = b(s^2)$ in Equation 11, both absolutely and relative to the variance. Widely spaced sampling times will reduce b, but also reduce the number of observations, thus increasing V, unless the sampling period is lengthened.

Sampling error will increase $V_{\varepsilon P}(t)$, the variance of the errors, but would not usually affect the covariances. Thus, reducing the sampling error will reduce V, but not b. A confidence interval for δ (Equation 8) should have length about $2t\sqrt{V}$ (where t is from the t distribution). The standard interval, using s^2, has length about $2t\sqrt{[V - b]}$, so is too short by about $2tb/\{\sqrt{V} + \sqrt{[V - b]}\}$. Both the absolute and relative error increase if V decreases but covariances do not.

If the underestimate of variance seems likely to be serious, the autocorrelation of the errors can be allowed for, e.g., by writing an explicit model. If variances within periods are equal, differences at t_j and t_{j+m} in the same period have correlation ρ^m, and differences in different periods are uncorrelated, then $b(s^2) = 2\rho V/(1 - \rho)$. Thus multiplying s^2 by $1 + 2r/(1 - r)$, where r is the first order serial correlation (estimating ρ), may approximate the right adjustment, though only roughly.

Alternative Models

The model of Equation 6 assumes that spatial and temporal variation are additive: i.e., systematic and random large-scale fluctuations, like seasons and storms, are assumed to affect all sites in the region approximately equally, so that they largely cancel in the differences, $N_I(t) - N_C(t)$. We now consider some alternatives. For applications of some of these, see Bence et al. (Chapter 8).

Additivity after Transformation. Spatial and temporal variation may not be additive on the abundance: e.g., they could be additive on the log of the abundance (i.e., multiplicative on the abundance) so Equations 6 and 7 hold for

$\log[N_{LP}(t \mid H, t_0)]$, or additivity may apply to the reciprocal (the area or volume needed to support one individual) or to some other transformation.

One way to approach this problem is to seek the right transformation, e.g., using a power transformation with power chosen by the Tukey test (Tukey 1949, Snedecor and Cochran 1980, p. 283) or a similar method (e.g., Andrews 1971, Berry 1987). While useful, this approach is not trouble free (Smith et al. 1991, 1993). Some transformations require rather arbitrary adjustments when the observed abundance is zero (e.g., due to sampling error). The change in the difference of the means of the transformed variables may not be easy to interpret in terms of the original abundances. There may not be a "right" transformation, e.g., if abundances at Impact are higher than at Control under some conditions or seasons, but lower under others. Even if there is a "right" transformation for the Before period, it may not be right for the After period if the alteration does not act in the same way as other effects.

Ratio Models. Suppose the Impact and Control sites are sub-areas of a larger region subject to mixing and experiencing similar environmental variation, both systematic and random. The total population of the region, $N_R(t)$, fluctuates in response to this variation, and also to local variation at the sub-areas, but is then redistributed by movement, births and deaths. The abundance in a sub-area might then tend to be a roughly fixed proportion of the abundance of the region, the proportion being determined by such factors as water movement (bringing in recruits), usable space, and local survivorship.

If, given $N_R(t)$, Impact and Control abundances are given by independent Poisson variables, $N_{\mathrm{I}}(t)$ and $N_{\mathrm{C}}(t)$, with means $\alpha_{\mathrm{I}} N_R(t)$ and $\alpha_{\mathrm{C}} N_R(t)$, then standard tests and confidence intervals for the ratio, $r_{\mathrm{IC}} = \alpha_{\mathrm{I}}/\alpha_{\mathrm{C}}$, are derived from those for the parameter $p_{\mathrm{I}} = \alpha_{\mathrm{I}}/(\alpha_{\mathrm{I}} + \alpha_{\mathrm{C}})$, the probability of "success" in the Binomial distribution for $N_{\mathrm{I}}(t)$ successes, from $N_{\mathrm{C}}(t) + N_{\mathrm{I}}(t)$ trials (Lehmann 1959, p. 180). If, given the values of the sequence $\{N_R(t_i)\}$, the pairs $\{(N_{\mathrm{C}}(t_i), N_{\mathrm{I}}(t_i))\}$ are independent (over time), then the combined estimate of p_{I} is $\hat{p}_{\mathrm{I}} = 1/[1 + \Sigma N_{\mathrm{C}}(t_i)/\Sigma N_{\mathrm{I}}(t_i)]$. If the totals, $\Sigma N_{\mathrm{C}}(t_i)$ and $\Sigma N_{\mathrm{I}}(t_i)$, are not small, then $\hat{r}_{\mathrm{IC}} = \hat{p}_{\mathrm{I}}/(1 - \hat{p}_{\mathrm{I}}) = \Sigma N_{\mathrm{I}}(t_i)/\Sigma N_{\mathrm{C}}(t_i)$ is approximately Normal with mean r_{IC} and variance $r_{\mathrm{IC}}{}^2/\Sigma[N_{\mathrm{C}}(t_i) + N_{\mathrm{I}}(t_i)]$. We might measure the change at Impact relative to Control by the difference between the Before and After r_{IC}'s, substituting $\hat{r}_{\mathrm{IC}}$ for r_{IC} in the variance formulae for a confidence interval. Eberhart (1976) suggests a similar approach.

Three uncertain assumptions in this model are (i) that $\alpha_{\mathrm{I}}/\alpha_{\mathrm{C}}$ is the same at each time, (ii) that $N_{\mathrm{C}}(t)$ and $N_{\mathrm{I}}(t)$ are Poisson, and (iii) that the $\{(N_{\mathrm{C}}(t_i), N_{\mathrm{I}}(t_i))\}$ pairs are independent given the $N_R(t)$'s. These have the unlikely corollary that, given the sample size, $\Sigma[N_{\mathrm{C}}(t_i) + N_{\mathrm{I}}(t_i)]$, the number of sampling times is irrelevant: a single Before time and a single After time would suffice. More realistic approaches include: (i) treating $p_{\mathrm{I}} = p_{\mathrm{I}}(t)$ as variable in time, e.g., as a stochastic process with mean p_{I}; (ii) assuming the mean of $N_L(t)$ ($L = \mathrm{C}$ or I) to be θ_L which is itself gamma distributed with mean $\alpha_L N_R(t)$, so that $N_C(t)$ and $N_I(t)$ have negative binomial distributions (this leads to a special case of (i), with $p_I(t_i)$ having a Beta distribution, if the two gamma distributions have the same scale parameter); (iii) treating the values $\hat{p}_I(t_i) = N_I(t_i)/[N_C(t_i) + N_I(t_i)]$ as a possibly correlated series, with variances, $V\{\hat{p}_I(t_i)\}$, not necessarily proportional to $1/[N_C(t_i) + N_I(t_i)]$.

A Predictive Model. Perhaps the most common covariate model is

$$N_I(t_i) = \gamma + \beta N_C(t_i) + \varepsilon_i, \qquad (12)$$

where the error, ε_i, is uncorrelated with $N_C(t_i)$. With this model, we could estimate an alteration effect as a change in γ or in β between Before and After. But it seems preferable to estimate or describe the change (i) by presenting both regression lines, thus showing that the effect varies with "environmental conditions", as represented by $N_C(t)$ (e.g., Mathur et al., 1980; Bence et al., Chapter 8), and (ii) presenting as the overall estimate the change in $\gamma + \beta N_C$, where N_C is a "typical" Control value, possibly the average of all observed $N_C(t)$ values, both Before and After. The hope is that most of the variation in $N_I(t)$ is "explained by" variation in $N_C(t)$. Large, low-frequency variation, like El Niño, which can be difficult to model but affects both N_C and N_I, might then not play a significant role, so that the ε_i's can be treated as independent or as obeying a simple generic model, e.g., autoregressive of order 1 ($\varepsilon_i = b\varepsilon_{i-1} + a_i$, where $b < 1$: see Equation 4).

This model seems a potentially useful combination of the ideas of additive and multiplicative differences between the Impact and Control sites, but an attempt to derive it from a rough mechanistic model shows that nuisance variation may not be completely removed this way. Suppose $N_R(t)$, the regional average abundance at time t, has mean $\mu(t)$ and variance $\sigma^2(t)$; that $N_R(t)$, $N_C(t)$ and $N_I(t)$ are jointly Normal; and that, given $N_R(t)$, $N_C(t)$ and $N_I(t)$ have means $\alpha_C N_R(t)$ and $\alpha_I N_R(t)$, variances ϕ_{CC} and ϕ_{II}, and covariance ϕ_{CI}. Then the unconditional distribution of $N_I(t)$ and $N_C(t)$ is Normal with means $\alpha_I\mu(t)$ and $\alpha_C\mu(t)$, variances $\tau_{II} = \phi_{II} + \alpha_I^2\sigma^2(t)$ and $\tau_{CC} = \phi_{CC} + \alpha_C^2\sigma^2(t)$, and covariance $\tau_{CI} = \phi_{CI} + \alpha_C\alpha_I\sigma^2(\mathrm{t})$. Standard manipulations show that, given $N_C(t)$, the distribution of $N_I(t)$ is Normal with mean $E\{N_I(t) \mid N_C(t)\} = (\alpha_I - b\alpha_C)\mu(t) + bN_C(t)$ and variance $V = \tau_{II} - \tau_{CI}^2/\tau_{CC}$, where $b = \tau_{CI}/\tau_{CC}$.

Thus neither $\mu(t)$ nor $\sigma^2(t)$, the "regional" mean and variance, drop out in this version: both the slope and the intercept in Equation 12 are functions of time. This is a form of the "errors in variables" problem (Fuller 1987, Snedecor and Cochran 1980, p. 171): Equation 12 holds with "$\alpha_C N_R(t_i)$" instead of "$N_C(t_i)$"; the latter is an estimate, with error, of the former; when it is substituted, its error becomes part of the "ε_i", which is thus correlated with the "independent" variable, $N_C(t_i)$.

Thus, the "predictive" model of Equation 12 does not follow from this argument. It may not follow from any simple mechanistic argument, though more careful attention to mechanisms and distribution choices may do better. This does not mean it should not be used: statistical analyses are frequently based on generic models chosen more because they are simple, well understood, and have about the right behavior, than because of a mechanistic derivation. If σ^2 is large compared to the ϕs, or if it does not vary much over time, and if $\mu(t)$ can be modeled as a seasonal function, then the modified Equation 12,

$$N_I(t_i) = \gamma + \alpha_1 \sin 2\pi t + \alpha_2 \cos 2\pi t + \beta N_C(t_i) + \varepsilon_i, \tag{13}$$

might be satisfactory. This model would allow the relative advantages of the sites, or the effect of the alteration, to vary with seasons.

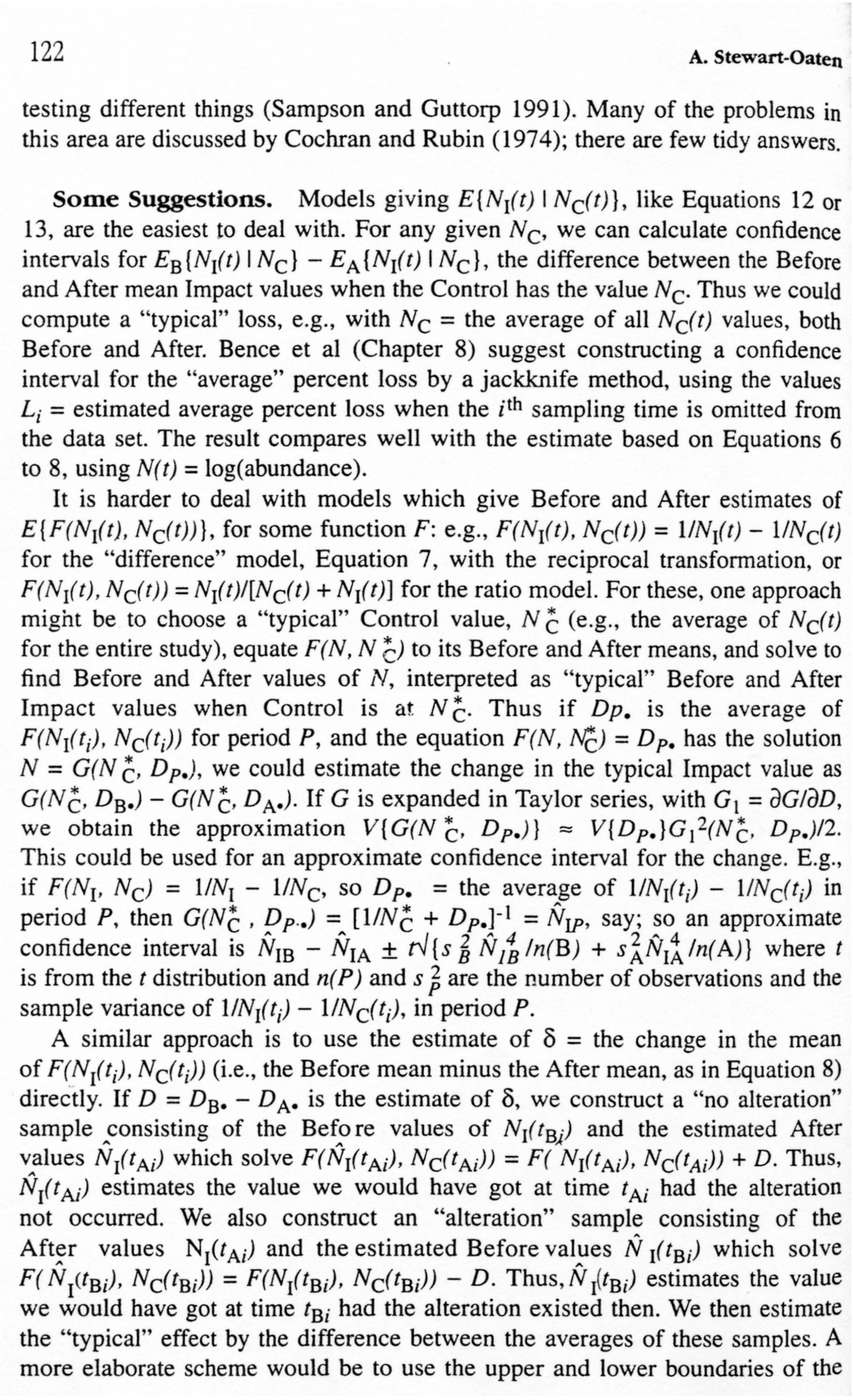
Matching. In some cases, matching could be used to remove deterministic time effects, including cases where the alteration itself has different effects under different conditions. For example, data could be analyzed separately for winter and summer or for periods of upcoast and downcoast currents (Reitzel et al. 1994).

Model Uncertainty

The previous section suggests that there may be many plausible models on which assessment could be based. It may be possible to rule some of these out by goodness-of-fit tests, diagnostic plots, or arguments based on mechanisms or auxiliary variables. These methods are informal (e.g., there is no clear criterion for choosing the level of a goodness-of-fit test), but honest use would usually retain several models for assessment. Thus model uncertainty is a part of the uncertainty in estimates of change.

It is likely that none of the models remaining is "correct". A strategy is to begin with a broad enough range of realistic models to have a high likelihood that at least one of them is close enough to the truth for effective decision making. Impact assessment could benefit from a "kit" of generic stochastic spatio-temporal models which can allow for major systematic and random effects and reflect the physiology and behavior of groups of organisms, but allow comparison of neighboring sites over time without an excess of unknown parameters. A possible starting set might consist of the three types of models—additive (perhaps after transformation), ratio, and predictive—discussed in the previous section.

If all the uneliminated models give similar answers, model uncertainty could be displayed by giving the results from the simplest or most plausible model, with bounds showing the range of variation due to model differences. But "similar answers" may not be easy to define: e.g., models assuming multiplicative effects must give different answers from models assuming additive effects. This is a case where estimates and confidence intervals for effect sizes are messier than *P*-values for a test of "no effect" (see Stewart-Oaten, Chapter 2)—although the tidiness of the latter is misleading, since tests using different models are



testing different things (Sampson and Guttorp 1991). Many of the problems in this area are discussed by Cochran and Rubin (1974); there are few tidy answers.

Some Suggestions. Models giving $E\{N_I(t) \mid N_C(t)\}$, like Equations 12 or 13, are the easiest to deal with. For any given N_C, we can calculate confidence intervals for $E_B\{N_I(t) \mid N_C\} - E_A\{N_I(t) \mid N_C\}$, the difference between the Before and After mean Impact values when the Control has the value N_C. Thus we could compute a "typical" loss, e.g., with N_C = the average of all $N_C(t)$ values, both Before and After. Bence et al (Chapter 8) suggest constructing a confidence interval for the "average" percent loss by a jackknife method, using the values L_i = estimated average percent loss when the i^{th} sampling time is omitted from the data set. The result compares well with the estimate based on Equations 6 to 8, using $N(t) = \log(\text{abundance})$.

It is harder to deal with models which give Before and After estimates of $E\{F(N_I(t), N_C(t))\}$, for some function F: e.g., $F(N_I(t), N_C(t)) = 1/N_I(t) - 1/N_C(t)$ for the "difference" model, Equation 7, with the reciprocal transformation, or $F(N_I(t), N_C(t)) = N_I(t)/[N_C(t) + N_I(t)]$ for the ratio model. For these, one approach might be to choose a "typical" Control value, N_C^* (e.g., the average of $N_C(t)$ for the entire study), equate $F(N, N_C^*)$ to its Before and After means, and solve to find Before and After values of N, interpreted as "typical" Before and After Impact values when Control is at N_C^*. Thus if $Dp_\bullet$ is the average of $F(N_I(t_i), N_C(t_i))$ for period P, and the equation $F(N, N_C^*) = D_{P\bullet}$ has the solution $N = G(N_C^*, D_{P\bullet})$, we could estimate the change in the typical Impact value as $G(N_C^*, D_{B\bullet}) - G(N_C^*, D_{A\bullet})$. If G is expanded in Taylor series, with $G_1 = \partial G/\partial D$, we obtain the approximation $V\{G(N_C^*, D_{P\bullet})\} \approx V\{D_{P\bullet}\}G_1{}^2(N_C^*, D_{P\bullet})/2$. This could be used for an approximate confidence interval for the change. E.g., if $F(N_I, N_C) = 1/N_I - 1/N_C$, so $D_{P\bullet}$ = the average of $1/N_I(t_i) - 1/N_C(t_i)$ in period P, then $G(N_C^*, D_{P\cdot\bullet}) = [1/N_C^* + D_{P\bullet}]^{-1} = \hat{N}_{IP}$, say; so an approximate confidence interval is $\hat{N}_{IB} - \hat{N}_{IA} \pm t\sqrt{\{s_B^2 \hat{N}_{IB}^4/n(B) + s_A^2 \hat{N}_{IA}^4/n(A)\}}$ where t is from the t distribution and $n(P)$ and s_P^2 are the number of observations and the sample variance of $1/N_I(t_i) - 1/N_C(t_i)$, in period P.

A similar approach is to use the estimate of δ = the change in the mean of $F(N_I(t_i), N_C(t_i))$ (i.e., the Before mean minus the After mean, as in Equation 8) directly. If $D = D_{B\bullet} - D_{A\bullet}$ is the estimate of δ, we construct a "no alteration" sample consisting of the Before values of $N_I(t_{Bi})$ and the estimated After values $\hat{N}_I(t_{Ai})$ which solve $F(\hat{N}_I(t_{Ai}), N_C(t_{Ai})) = F(N_I(t_{Ai}), N_C(t_{Ai})) + D$. Thus, $\hat{N}_I(t_{Ai})$ estimates the value we would have got at time t_{Ai} had the alteration not occurred. We also construct an "alteration" sample consisting of the After values $N_I(t_{Ai})$ and the estimated Before values $\hat{N}_I(t_{Bi})$ which solve $F(\hat{N}_I(t_{Bi}), N_C(t_{Bi})) = F(N_I(t_{Bi}), N_C(t_{Bi})) - D$. Thus, $\hat{N}_I(t_{Bi})$ estimates the value we would have got at time t_{Bi} had the alteration existed then. We then estimate the "typical" effect by the difference between the averages of these samples. A more elaborate scheme would be to use the upper and lower boundaries of the confidence interval for δ, instead of using D. Thus if the confidence interval for the change in $E\{1/N_{\mathrm{I}}(t) - 1/N_{\mathrm{C}}(t)\}$ is $(D_{\mathrm{L}}, D_{\mathrm{U}})$, then the upper boundary for an approximate confidence interval for the change in N_{I} is found by (i) construct the "no alteration" sample $N_{\mathrm{I}}(t_{\mathrm{B1}}), \ldots, N_{\mathrm{I}}(t_{\mathrm{B}n(\mathrm{B})}), \hat{N}_{\mathrm{I}}(t_{\mathrm{A1}}), \ldots, \hat{N}_{\mathrm{I}}(t_{\mathrm{A}n(\mathrm{A})})$, where $\hat{N}_{\mathrm{I}}(t_{\mathrm{A}k}) = [1/N_{\mathrm{I}}(t_{\mathrm{A}k}) + D_L]^{-1}$, and calculate its mean, $M_{\mathrm{U,NA}}$; (ii) construct the "alteration" sample, $\hat{N}_{\mathrm{I}}(t_{\mathrm{B1}}), \ldots, \hat{N}_{\mathrm{I}}(t_{\mathrm{B}n(\mathrm{B})}), N_{\mathrm{I}}(t_{\mathrm{A1}}), \ldots, N_{\mathrm{I}}(t_{\mathrm{A}n(\mathrm{A})})$ where $\hat{N}_{\mathrm{I}}(t_{\mathrm{B}k}) = [1/N_{\mathrm{I}}(t_{\mathrm{B}k}) - D_{\mathrm{L}}]^{-1}$, and calculate its mean, $M_{\mathrm{U,A}}$. The approximate upper confidence limit is $M_{\mathrm{U,NA}} - M_{\mathrm{U,A}}$. ($D_{\mathrm{L}}$ is used for the upper limit because $F(N_{\mathrm{I}}, N_{\mathrm{C}}) = 1/N_{\mathrm{I}} - 1/N_{\mathrm{C}}$ is a decreasing function of N_{I}. If the alteration reduces abundance, D_{L} and D_{U} should be negative.)

Approximate Answers to the Right Questions. Such comparisons are very rough. Presenting confidence intervals from several different models is an attempt to combine the ranges of their effect estimates and their error (standard deviation) estimates. The approach ignores both the "errors in variables" problem and that means are not preserved by nonlinear transformations, e.g., $E\{1/N_{\mathrm{I}}(t)\} \neq 1/E\{N_{\mathrm{I}}(t)\}$. The last problem may not be severe for differences, since the errors may approximately cancel. A Taylor series approach to it is described by Sampson and Guttorp (1991), but seems hard to apply here: the pairs $\{(N_{\mathrm{I}}(t_i), N_{\mathrm{C}}(t_i))\}$ are assumed to be independent (for different times). Medians are preserved, so it might be possible to improve model comparisons by using confidence intervals for medians (e.g., based on the sign test) rather than for means.

But these, or similar, comparisons could be useful. Although the models may have very different forms, e.g., additive versus multiplicative, they may give similar results, especially if the Before and After series of Control values are similar. When the series are dissimilar, we need to distinguish a change at one site that does not occur at the other from a change in a comparison measure (the difference or the ratio) that is due solely to a natural change over the entire region. Even when a single model seems clearly "best", we may want to present the results in a different measure or "scale": e.g., the ratio model may be the most plausible, and fit the data best, but a decision maker might want to know "about how many individuals" of a given species will be "lost", i.e., the arithmetic difference between N_{I} and what it would have been. When several dissimilar models remain in the running, P-values for a test of "no effect" might be useful, not as a measure of the strength of the effect but to help indicate the compatibility of the models. In some cases it may be necessary to report more than one set of results, with arguments for preferring some model's to others.

There are ways to avoid (or evade) model uncertainty. One is to choose a standard model (e.g., Equation 7 with independent errors, $\varepsilon_{P(t)}$), or a standard analysis (e.g., a t-test or ANOVA), and report the results of this alone. By implying that the model used (often an implausibly simple one) is known with certainty to be true, this approach seems misleading. It is sometimes supported by subjecting the model to a goodness-of-fit test, but other plausible models might also pass this test while giving different assessment results, especially if data are sparse.

Estimating Other Parameters

Underwood (1991, Chapter 9) has suggested that effects on other parameters, notably the variance, should also be assessed. This is attractive, but estimating the variance functions, $V_B(t,H,t_0)$ and $V_A(t,H,t_0)$, or the covariance functions, $C_B(t_1,t_2,H,t_0)$ and $C_A(t_1,t_2,H,t_0)$, would be harder than estimating the means. If we observe $N(t)$ at times t_1, t_2, ..., t_n, the mean (over all possible realizations) of the usual estimate of variance,

$$s_N^2 = \Sigma[N(t_i) - N_\bullet]^2/(n - 1), \qquad (14)$$

is

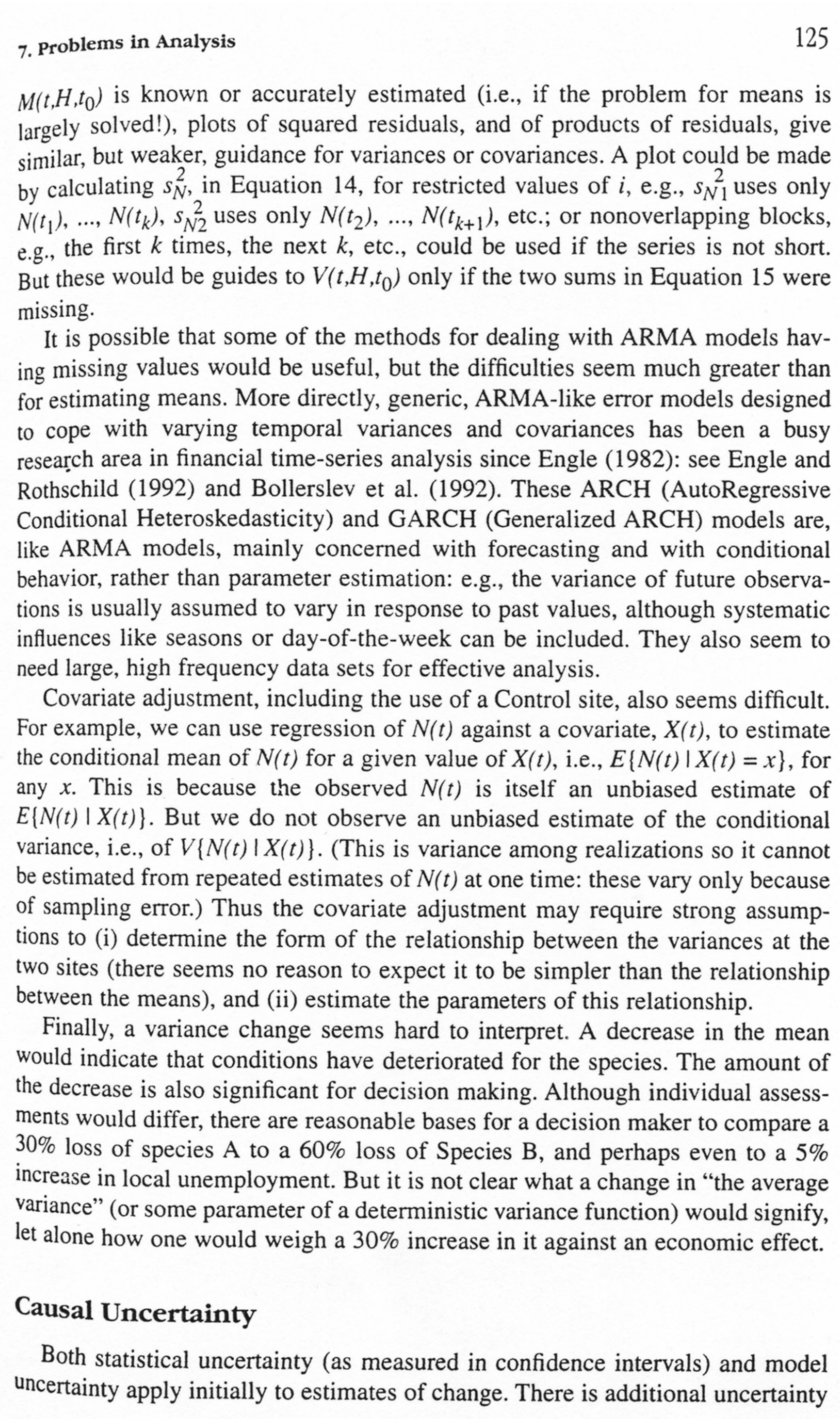
$$\phi = \Sigma[M(t_i,H,t_0) - M(\bullet,H,t_0)]^2/(n - 1) + V(\cdot,H,t_0) - 2\Sigma_j\Sigma_{j<k}C(t_j,t_k,H,t_0)/n(n - 1), \qquad (15)$$

where $M(\cdot,H,t_0) = \Sigma M(t_j,H,t_0)/n$ and $V(\cdot,H,t_0) = \Sigma V(t_j,H,t_0)/n$. If we knew how to adjust to eliminate the two sums in Equation 15, most of the difficulty in inference concerning the mean function would be removed. But inference concerning the variance function would still face all the problems discussed so far.

Inference (estimation and uncertainty measurement) concerning mean functions is difficult because their functional form is unknown, and temporal variation and serial correlation cause (i) effect estimates to have large variances and (ii) these variances to be hard to estimate. Inference concerning variance functions has all three difficulties in more severe form. The estimate of the variance function can also be biased by variation in the mean function, and the variance of this estimate is affected by higher temporal moments, e.g., by the kurtosis at time t, or $E\{N(t_1)N(t_2)N(t_3)N(t_4)\}$ for four distinct times.

The standard "fixes" of deterministic modeling, time-series modeling and the use of covariates are all harder to achieve. Deterministic modeling requires intuition or knowledge about the behavior of these functions. Variances seem as likely to vary over time as means; e.g., $V(t,H,t_0)$ seems likely to be higher if t is in a period with high levels of disturbance (more storms, upwellings, or migrations), and the part due to sampling error may depend on population size. Covariances, $C(t_1,t_2,H,t_0)$, would be expected to be higher if t_1 and t_2 are close, but lower if the period between t_1 and t_2 is one of high disturbance. It seems harder to base plausible deterministic models for variances and covariances on these mechanisms than it is for means. Deterministic models for higher moments seem even more remote. Cox (1981) briefly discusses approaches to monotone and cyclical variation in variances.

If $V(t,H,t_0)$ varies in time, we never observe an estimate of it. In contrast, the observed $N(t)$ is itself an unbiased estimate of $M(t,H,t_0)$, so plotting the path traced out by the observations can give us some indication of functional form. If



$M(t,H,t_0)$ is known or accurately estimated (i.e., if the problem for means is largely solved!), plots of squared residuals, and of products of residuals, give similar, but weaker, guidance for variances or covariances. A plot could be made by calculating s_N^2, in Equation 14, for restricted values of i, e.g., s_{N1}^2 uses only $N(t_1)$, ..., $N(t_k)$, s_{N2}^2 uses only $N(t_2)$, ..., $N(t_{k+1})$, etc.; or nonoverlapping blocks, e.g., the first k times, the next k, etc., could be used if the series is not short. But these would be guides to $V(t,H,t_0)$ only if the two sums in Equation 15 were missing.

It is possible that some of the methods for dealing with ARMA models having missing values would be useful, but the difficulties seem much greater than for estimating means. More directly, generic, ARMA-like error models designed to cope with varying temporal variances and covariances has been a busy research area in financial time-series analysis since Engle (1982): see Engle and Rothschild (1992) and Bollerslev et al. (1992). These ARCH (AutoRegressive Conditional Heteroskedasticity) and GARCH (Generalized ARCH) models are, like ARMA models, mainly concerned with forecasting and with conditional behavior, rather than parameter estimation: e.g., the variance of future observations is usually assumed to vary in response to past values, although systematic influences like seasons or day-of-the-week can be included. They also seem to need large, high frequency data sets for effective analysis.

Covariate adjustment, including the use of a Control site, also seems difficult. For example, we can use regression of $N(t)$ against a covariate, $X(t)$, to estimate the conditional mean of $N(t)$ for a given value of $X(t)$, i.e., $E\{N(t) \mid X(t) = x\}$, for any x. This is because the observed $N(t)$ is itself an unbiased estimate of $E\{N(t) \mid X(t)\}$. But we do not observe an unbiased estimate of the conditional variance, i.e., of $V\{N(t) \mid X(t)\}$. (This is variance among realizations so it cannot be estimated from repeated estimates of $N(t)$ at one time: these vary only because of sampling error.) Thus the covariate adjustment may require strong assumptions to (i) determine the form of the relationship between the variances at the two sites (there seems no reason to expect it to be simpler than the relationship between the means), and (ii) estimate the parameters of this relationship.

Finally, a variance change seems hard to interpret. A decrease in the mean would indicate that conditions have deteriorated for the species. The amount of the decrease is also significant for decision making. Although individual assessments would differ, there are reasonable bases for a decision maker to compare a 30% loss of species A to a 60% loss of Species B, and perhaps even to a 5% increase in local unemployment. But it is not clear what a change in "the average variance" (or some parameter of a deterministic variance function) would signify, let alone how one would weigh a 30% increase in it against an economic effect.

Causal Uncertainty

Both statistical uncertainty (as measured in confidence intervals) and model uncertainty apply initially to estimates of change. There is additional uncertainty as to whether the alteration caused the change (i.e., whether the change is an "effect"), since assignment of sampling times and sites to "unaffected" or "potentially affected" is not under the investigator's control, and in particular is not random.

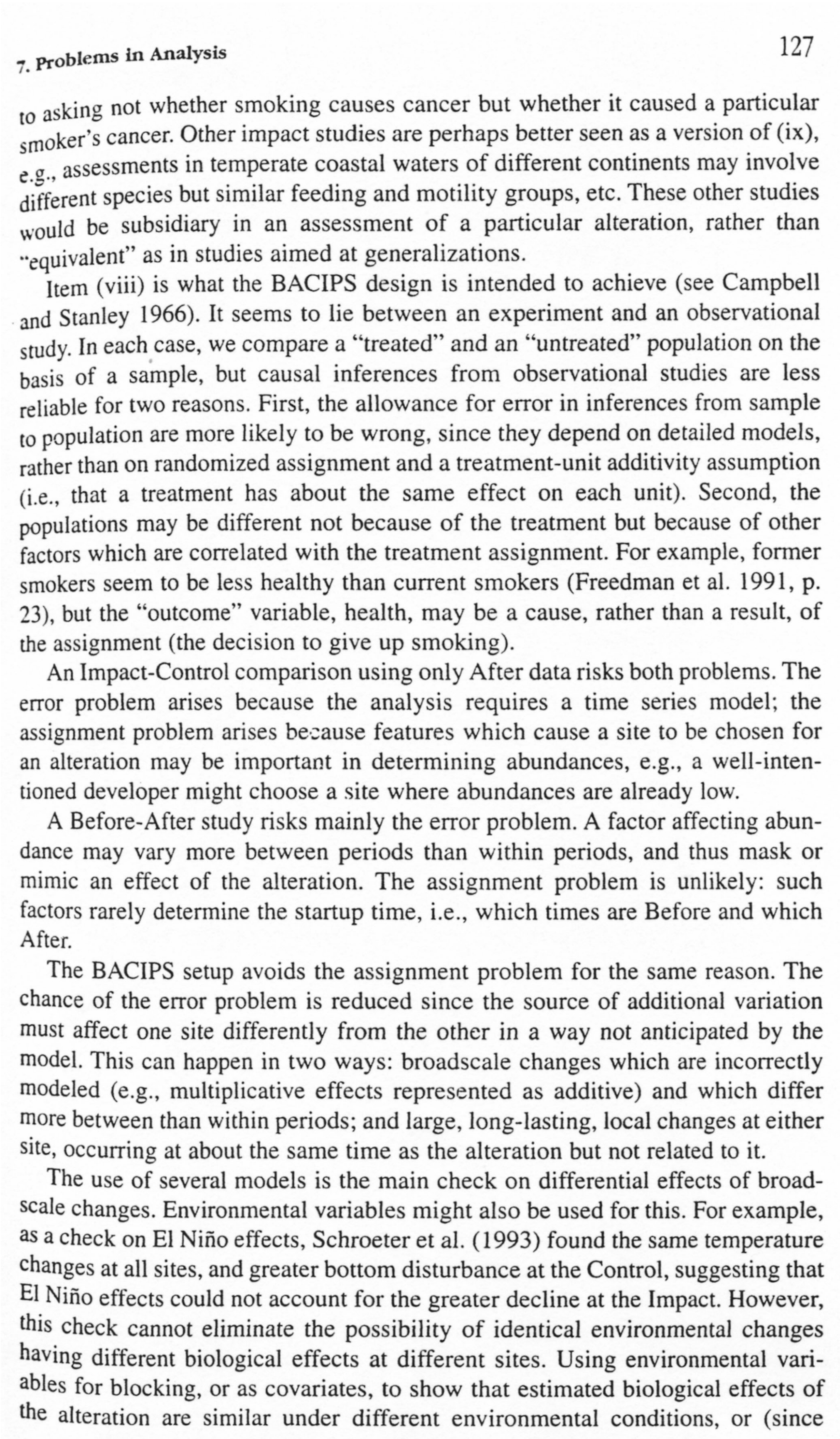
Experiments with randomized assignments seem clearly the best way to establish causes as opposed to associations (Barnard 1982), but these were invented relatively recently. (Fisher 1925). Much accepted scientific "truth" is still probably based on nonrandomized studies. Problems of causality in observational and quasi-experimental studies have attracted increased attention from statisticians recently (e.g., Cochran 1972, Rubin 1974, Pratt and Schlaifer 1984, Rosenbaum 1984, Cox 1992).

Hill (1965, see also 1971, Chapter 24) lists characteristics favoring causal interpretation of results from observational studies: (i) strength of effect, (ii) consistency (among studies), (iii) specificity, (iv) temporality (does the cause precede the effect), (v) biological gradient (monotone dose-response curve), (vi) plausibility, (vii) coherence, (viii) experimental or "semi-experimental" evidence, and (ix) analogy (with similar causes which led to similar effects). In impact assessment, (iv) seems covered by the Before data.

Hill (1965) stresses (i), which seems to favor confidence intervals over hypothesis tests. The larger the effect, the less likely it is to be the result of some overlooked factor. He points out that the measure of "strength" does not need to be the same as the measure of "importance": e.g., in medical studies, the relative difference between groups in death rates from a specific illness may be convincing, even though the absolute rates are both small. Thus an impact analysis in terms of the ratio model can help suggest cause, even though importance may be judged by an estimate of absolute change.

Schroeter et al. (1993) stress (iii), (v), (vi) and (vii): "In the absence of a demonstrated causal chain, a convincing case requires that the results for a number of different species tie together and be consistent, that plausible mechanisms for an ecological impact be identified, and that reasonable alternative mechanisms be explored and ruled out." In their study of kelp bed invertebrates affected by the plume of a power plant cooling system, they (vii) demonstrate similar declines for similar species of snails, at two Impact sites, with (v) the site nearer the plant showing greater effects; (vi) suggest two plausible mechanisms (reduced supplies of drift kelp, and increased abrasion due to the flux of fine particles); and (iii) reject several alternative explanations (e.g., by carrying out separate analyses omitting samples possibly affected by an urchin feeding front). Another version of (iii) might be that species which should *not* be affected should not change. Also, a time trend in the estimated change might indicate the temporary effects of installation, as opposed to the long-term effects of existence or operation.

A version of (ii) may be comparison of results with other assessment studies of similar types. Note, however, that impact studies are usually concerned with effects at a particular place and time, not with generalizations: they are analogous



to asking not whether smoking causes cancer but whether it caused a particular smoker's cancer. Other impact studies are perhaps better seen as a version of (ix), e.g., assessments in temperate coastal waters of different continents may involve different species but similar feeding and motility groups, etc. These other studies would be subsidiary in an assessment of a particular alteration, rather than "equivalent" as in studies aimed at generalizations.

Item (viii) is what the BACIPS design is intended to achieve (see Campbell and Stanley 1966). It seems to lie between an experiment and an observational study. In each case, we compare a "treated" and an "untreated" population on the basis of a sample, but causal inferences from observational studies are less reliable for two reasons. First, the allowance for error in inferences from sample to population are more likely to be wrong, since they depend on detailed models, rather than on randomized assignment and a treatment-unit additivity assumption (i.e., that a treatment has about the same effect on each unit). Second, the populations may be different not because of the treatment but because of other factors which are correlated with the treatment assignment. For example, former smokers seem to be less healthy than current smokers (Freedman et al. 1991, p. 23), but the "outcome" variable, health, may be a cause, rather than a result, of the assignment (the decision to give up smoking).

An Impact-Control comparison using only After data risks both problems. The error problem arises because the analysis requires a time series model; the assignment problem arises because features which cause a site to be chosen for an alteration may be important in determining abundances, e.g., a well-intentioned developer might choose a site where abundances are already low.

A Before-After study risks mainly the error problem. A factor affecting abundance may vary more between periods than within periods, and thus mask or mimic an effect of the alteration. The assignment problem is unlikely: such factors rarely determine the startup time, i.e., which times are Before and which After.

The BACIPS setup avoids the assignment problem for the same reason. The chance of the error problem is reduced since the source of additional variation must affect one site differently from the other in a way not anticipated by the model. This can happen in two ways: broadscale changes which are incorrectly modeled (e.g., multiplicative effects represented as additive) and which differ more between than within periods; and large, long-lasting, local changes at either site, occurring at about the same time as the alteration but not related to it.

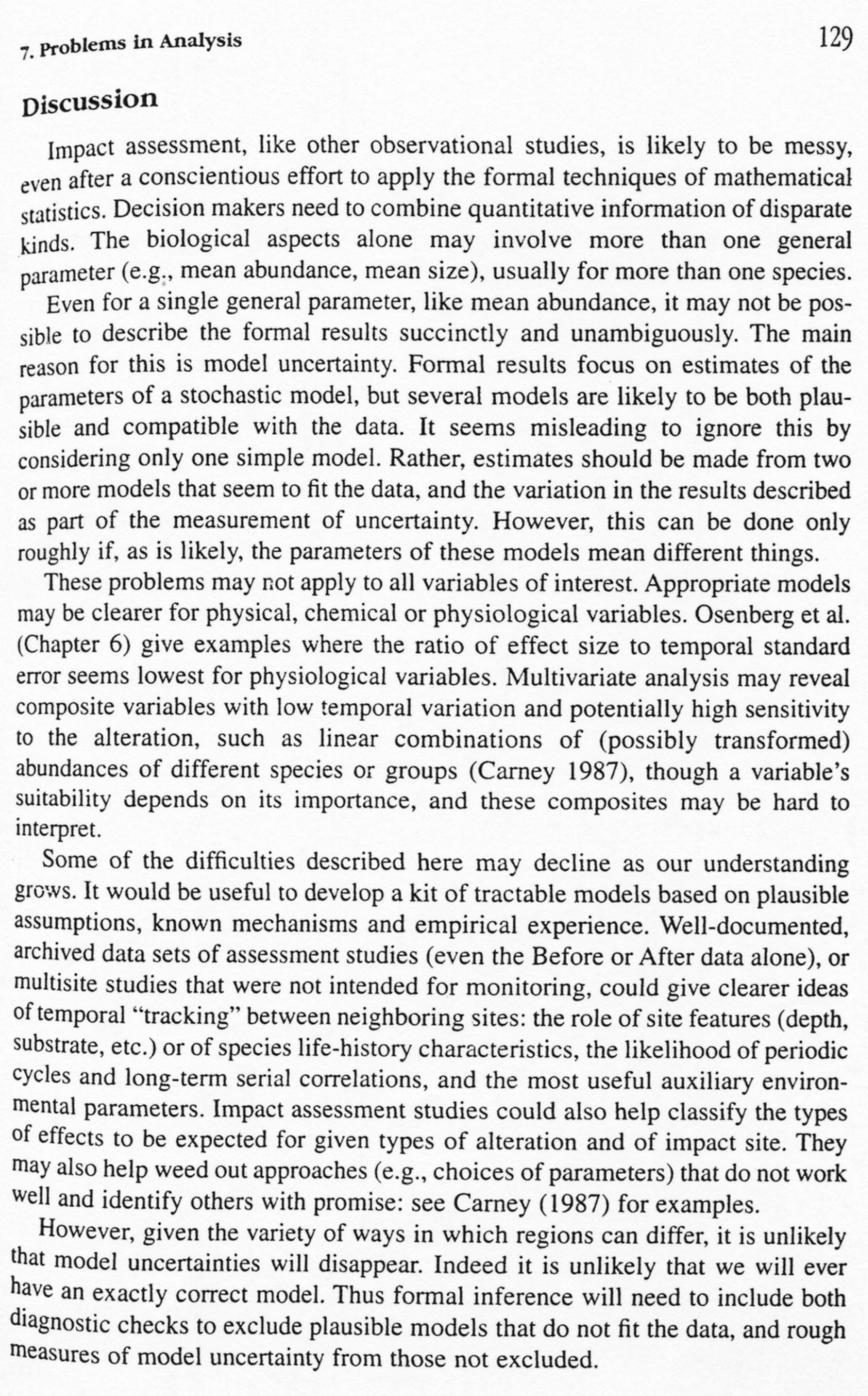
The use of several models is the main check on differential effects of broadscale changes. Environmental variables might also be used for this. For example, as a check on El Niño effects, Schroeter et al. (1993) found the same temperature changes at all sites, and greater bottom disturbance at the Control, suggesting that El Niño effects could not account for the greater decline at the Impact. However, this check cannot eliminate the possibility of identical environmental changes having different biological effects at different sites. Using environmental variables for blocking, or as covariates, to show that estimated biological effects of the alteration are similar under different environmental conditions, or (since some effects may be expected to vary with conditions) that they agree with expectations, may help, but this may be difficult if data are sparse.

Additional checks are possible with multiple Control sites. Results using different Controls should be similar, or the differences explainable. A single model incorporating all Control sites, and allowing for systematic location effects and for both temporal and spatial correlation, may allow more realism and flexibility, and reduce the "errors in variables" problem (since the "unaffected" abundance will be more accurately estimated), and still retain more degrees of freedom (number of observations − number of estimated parameters) for estimating variance. In particular, multiple Controls may give a much better idea of the likely variability due to naturally-arising, large, long-lasting local perturbations which might mimic or mask an alteration effect.

However, multiple Controls offer no guarantees. The sites are not random, since Impact, at least, will have been deliberately chosen. It may well be suggestive (in a study with After data only) that the mean of the Impact site over the After period is (say) smaller than all the Control means, or (in a Before-After study) that the difference between the Before and After means was greater at the Impact site than at any of the Controls, but it is not possible to attach a standard error, confidence or P-value to this without several dubious assumptions: selection of the Impact site was "effectively" random (e.g., the reasons for its selection are unrelated, directly or indirectly, to the abundance of the species under study), and the means or changes in means at the different sites are independent (e.g., neighboring sites are not expected to have more similar changes than distant sites). A "single control" study could get misleading results if an extraneous factor affects the Control. Multiple controls protect against this, but not against an extraneous factor affecting the Impact site, or affecting some Controls but not others.

Multiple controls might even be less reliable than a single control if an extraneous factor affects the Impact site and nearby, but not distant, Controls, or if the more distant Controls track the Impact site poorly. Thus, as noted by Underwood (1992, Chapter 9) and others, multiple Controls (and multiple Impacts, e.g., sites which let us check that the putative effect decreases with distance from the alteration) offer the possibility of more convincing causal arguments and of reductions in the effects of both natural temporal variation and model uncertainty, but these gains require careful modeling and analysis.

Stronger semi-experimental evidence arises if the potential cause can be applied, removed and then reinstated (Cox 1992). Effects on individuals during operation can be compared to effects during brief shutdowns (Raimondi and Schmitt 1992). Effects on populations may require long shutdowns, but an example is given by Granelli et al. (1990): the operation of a sewage plant was suspended for several months to allow comparison with "plant on" conditions. This seems to provide good evidence of the effect of an installation's operation; its existence may have other effects (e.g., provision of substrate or alteration of water movement) that cannot be checked this way.



Discussion

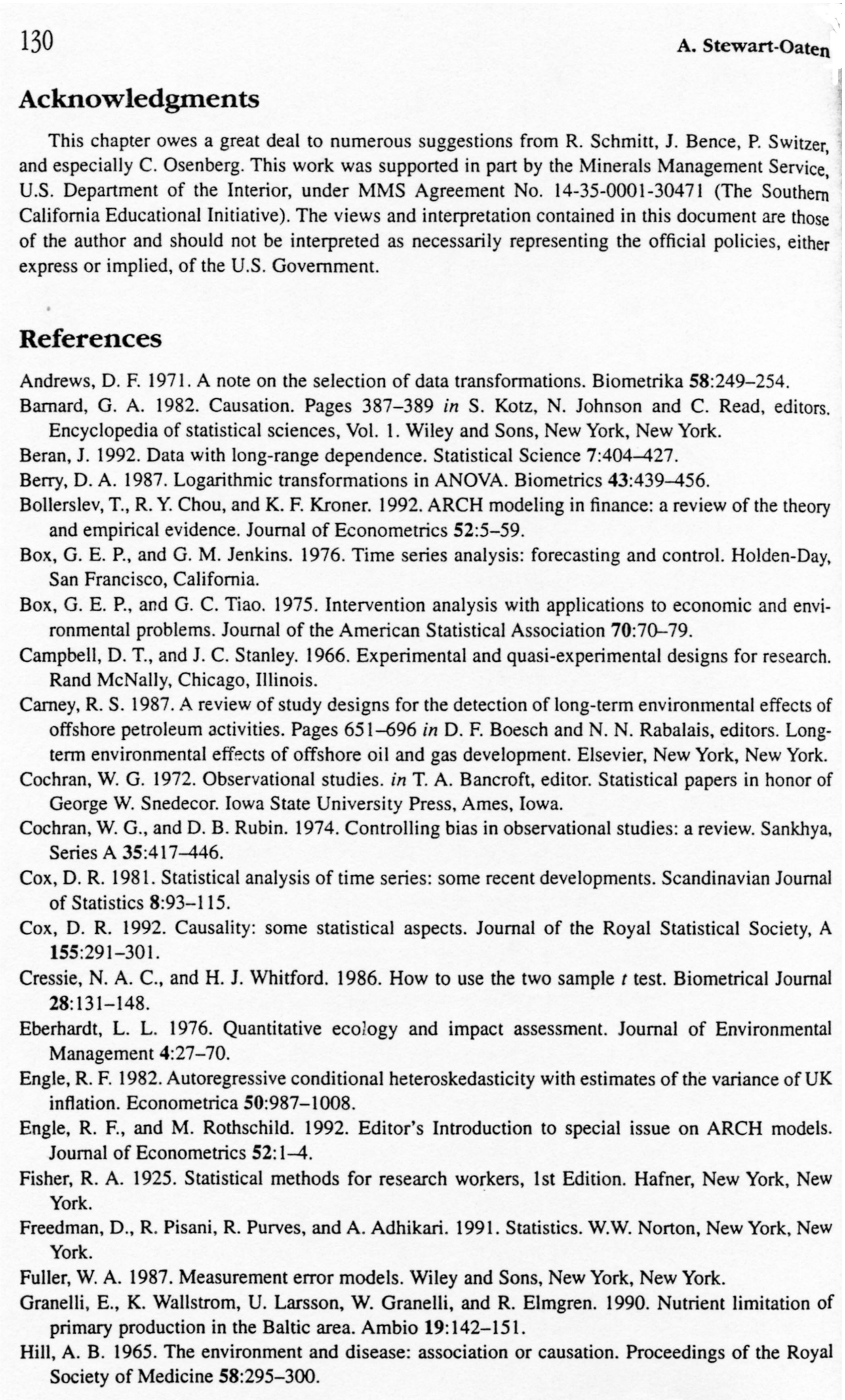
Impact assessment, like other observational studies, is likely to be messy, even after a conscientious effort to apply the formal techniques of mathematical statistics. Decision makers need to combine quantitative information of disparate kinds. The biological aspects alone may involve more than one general parameter (e.g., mean abundance, mean size), usually for more than one species.

Even for a single general parameter, like mean abundance, it may not be possible to describe the formal results succinctly and unambiguously. The main reason for this is model uncertainty. Formal results focus on estimates of the parameters of a stochastic model, but several models are likely to be both plausible and compatible with the data. It seems misleading to ignore this by considering only one simple model. Rather, estimates should be made from two or more models that seem to fit the data, and the variation in the results described as part of the measurement of uncertainty. However, this can be done only roughly if, as is likely, the parameters of these models mean different things.

These problems may not apply to all variables of interest. Appropriate models may be clearer for physical, chemical or physiological variables. Osenberg et al. (Chapter 6) give examples where the ratio of effect size to temporal standard error seems lowest for physiological variables. Multivariate analysis may reveal composite variables with low temporal variation and potentially high sensitivity to the alteration, such as linear combinations of (possibly transformed) abundances of different species or groups (Carney 1987), though a variable's suitability depends on its importance, and these composites may be hard to interpret.

Some of the difficulties described here may decline as our understanding grows. It would be useful to develop a kit of tractable models based on plausible assumptions, known mechanisms and empirical experience. Well-documented, archived data sets of assessment studies (even the Before or After data alone), or multisite studies that were not intended for monitoring, could give clearer ideas of temporal "tracking" between neighboring sites: the role of site features (depth, substrate, etc.) or of species life-history characteristics, the likelihood of periodic cycles and long-term serial correlations, and the most useful auxiliary environmental parameters. Impact assessment studies could also help classify the types of effects to be expected for given types of alteration and of impact site. They may also help weed out approaches (e.g., choices of parameters) that do not work well and identify others with promise: see Carney (1987) for examples.

However, given the variety of ways in which regions can differ, it is unlikely that model uncertainties will disappear. Indeed it is unlikely that we will ever have an exactly correct model. Thus formal inference will need to include both diagnostic checks to exclude plausible models that do not fit the data, and rough measures of model uncertainty from those not excluded.



Acknowledgments

This chapter owes a great deal to numerous suggestions from R. Schmitt, J. Bence, P. Switzer, and especially C. Osenberg. This work was supported in part by the Minerals Management Service, U.S. Department of the Interior, under MMS Agreement No. 14-35-0001-30471 (The Southern California Educational Initiative). The views and interpretation contained in this document are those of the author and should not be interpreted as necessarily representing the official policies, either express or implied, of the U.S. Government.

References

Andrews, D. F. 1971. A note on the selection of data transformations. Biometrika **58**:249–254.

Barnard, G. A. 1982. Causation. Pages 387–389 *in* S. Kotz, N. Johnson and C. Read, editors. Encyclopedia of statistical sciences, Vol. 1. Wiley and Sons, New York, New York.

Beran, J. 1992. Data with long-range dependence. Statistical Science **7**:404–427.

Berry, D. A. 1987. Logarithmic transformations in ANOVA. Biometrics **43**:439–456.

Bollerslev, T., R. Y. Chou, and K. F. Kroner. 1992. ARCH modeling in finance: a review of the theory and empirical evidence. Journal of Econometrics **52**:5–59.

Box, G. E. P., and G. M. Jenkins. 1976. Time series analysis: forecasting and control. Holden-Day, San Francisco, California.

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.

Campbell, D. T., and J. C. Stanley. 1966. Experimental and quasi-experimental designs for research. Rand McNally, Chicago, Illinois.

Carney, R. S. 1987. A review of study designs for the detection of long-term environmental effects of offshore petroleum activities. Pages 651–696 *in* D. F. Boesch and N. N. Rabalais, editors. Long-term environmental effects of offshore oil and gas development. Elsevier, New York, New York.

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

Cochran, W. G., and D. B. Rubin. 1974. Controlling bias in observational studies: a review. Sankhya, Series A **35**:417–446.

Cox, D. R. 1981. Statistical analysis of time series: some recent developments. Scandinavian Journal of Statistics **8**:93–115.

Cox, D. R. 1992. Causality: some statistical aspects. Journal of the Royal Statistical Society, A **155**:291–301.

Cressie, N. A. C., and H. J. Whitford. 1986. How to use the two sample *t* test. Biometrical Journal **28**:131–148.

Eberhardt, L. L. 1976. Quantitative ecology and impact assessment. Journal of Environmental Management **4**:27–70.

Engle, R. F. 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. Econometrica **50**:987–1008.

Engle, R. F., and M. Rothschild. 1992. Editor's Introduction to special issue on ARCH models. Journal of Econometrics **52**:1–4.

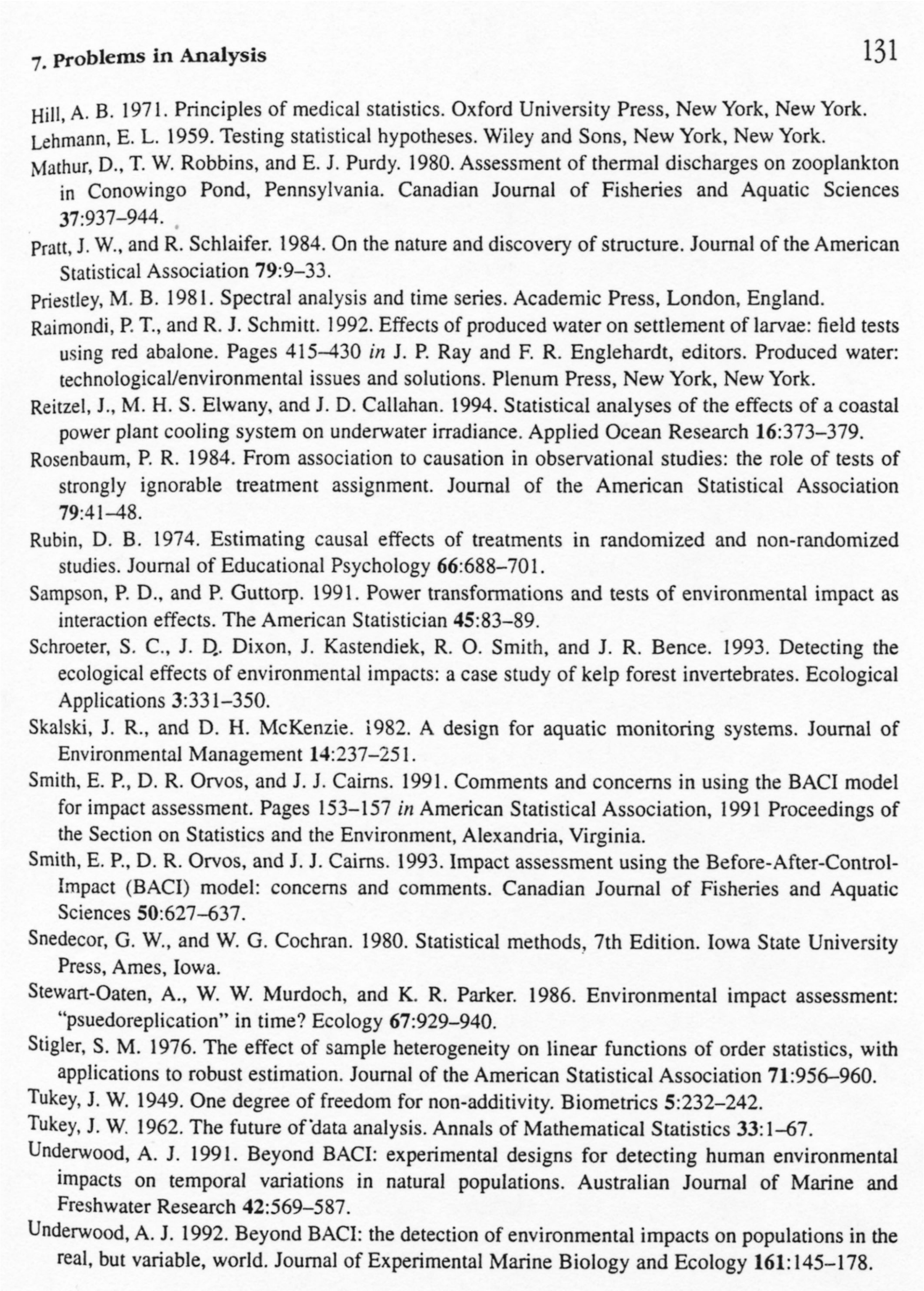
Fisher, R. A. 1925. Statistical methods for research workers, 1st Edition. Hafner, New York, New York.

Freedman, D., R. Pisani, R. Purves, and A. Adhikari. 1991. Statistics. W.W. Norton, New York, New York.

Fuller, W. A. 1987. Measurement error models. Wiley and Sons, New York, New York.

Granelli, E., K. Wallstrom, U. Larsson, W. Granelli, and R. Elmgren. 1990. Nutrient limitation of primary production in the Baltic area. Ambio **19**:142–151.

Hill, A. B. 1965. The environment and disease: association or causation. Proceedings of the Royal Society of Medicine **58**:295–300.



Hill, A. B. 1971. Principles of medical statistics. Oxford University Press, New York, New York.

Lehmann, E. L. 1959. Testing statistical hypotheses. Wiley and Sons, New York, New York.

Mathur, D., T. W. Robbins, and E. J. Purdy. 1980. Assessment of thermal discharges on zooplankton in Conowingo Pond, Pennsylvania. Canadian Journal of Fisheries and Aquatic Sciences **37**:937–944.

Pratt, J. W., and R. Schlaifer. 1984. On the nature and discovery of structure. Journal of the American Statistical Association **79**:9–33.

Priestley, M. B. 1981. Spectral analysis and time series. Academic Press, London, England.

Raimondi, P. T., and R. J. Schmitt. 1992. Effects of produced water on settlement of larvae: field tests using red abalone. Pages 415–430 *in* J. P. Ray and F. R. Englehardt, editors. Produced water: technological/environmental issues and solutions. Plenum Press, New York, New York.

Reitzel, J., M. H. S. Elwany, and J. D. Callahan. 1994. Statistical analyses of the effects of a coastal power plant cooling system on underwater irradiance. Applied Ocean Research **16**:373–379.

Rosenbaum, P. R. 1984. From association to causation in observational studies: the role of tests of strongly ignorable treatment assignment. Journal of the American Statistical Association **79**:41–48.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and non-randomized studies. Journal of Educational Psychology **66**:688–701.

Sampson, P. D., and P. Guttorp. 1991. Power transformations and tests of environmental impact as interaction effects. The American Statistician **45**:83–89.

Schroeter, S. C., J. D. Dixon, J. Kastendiek, R. O. Smith, and J. R. Bence. 1993. Detecting the ecological effects of environmental impacts: a case study of kelp forest invertebrates. Ecological Applications **3**:331–350.

Skalski, J. R., and D. H. McKenzie. 1982. A design for aquatic monitoring systems. Journal of Environmental Management **14**:237–251.

Smith, E. P., D. R. Orvos, and J. J. Cairns. 1991. Comments and concerns in using the BACI model for impact assessment. Pages 153–157 *in* American Statistical Association, 1991 Proceedings of the Section on Statistics and the Environment, Alexandria, Virginia.

Smith, E. P., D. R. Orvos, and J. J. Cairns. 1993. Impact assessment using the Before-After-Control-Impact (BACI) model: concerns and comments. Canadian Journal of Fisheries and Aquatic Sciences **50**:627–637.

Snedecor, G. W., and W. G. Cochran. 1980. Statistical methods, 7th Edition. Iowa State University Press, Ames, Iowa.

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

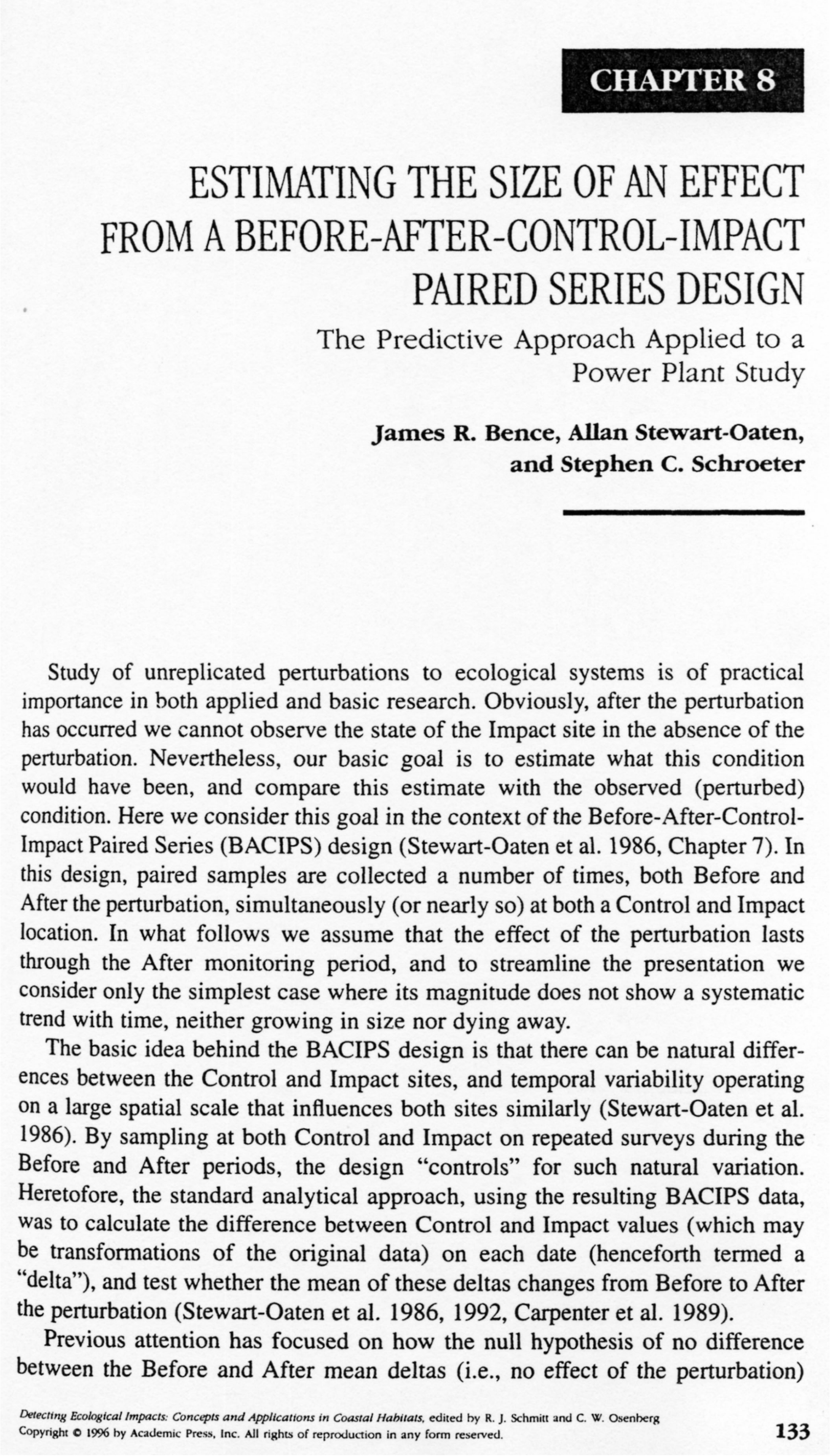
Stigler, S. M. 1976. The effect of sample heterogeneity on linear functions of order statistics, with applications to robust estimation. Journal of the American Statistical Association **71**:956–960.

Tukey, J. W. 1949. One degree of freedom for non-additivity. Biometrics **5**:232–242.

Tukey, J. W. 1962. The future of data analysis. Annals of Mathematical Statistics **33**:1–67.

Underwood, A. J. 1991. Beyond BACI: experimental designs for detecting human environmental impacts on temporal variations in natural populations. Australian Journal of Marine and Freshwater Research **42**:569–587.

Underwood, A. J. 1992. Beyond BACI: the detection of environmental impacts on populations in the real, but variable, world. Journal of Experimental Marine Biology and Ecology **161**:145–178.



CHAPTER 8

ESTIMATING THE SIZE OF AN EFFECT FROM A BEFORE-AFTER-CONTROL-IMPACT PAIRED SERIES DESIGN

The Predictive Approach Applied to a Power Plant Study

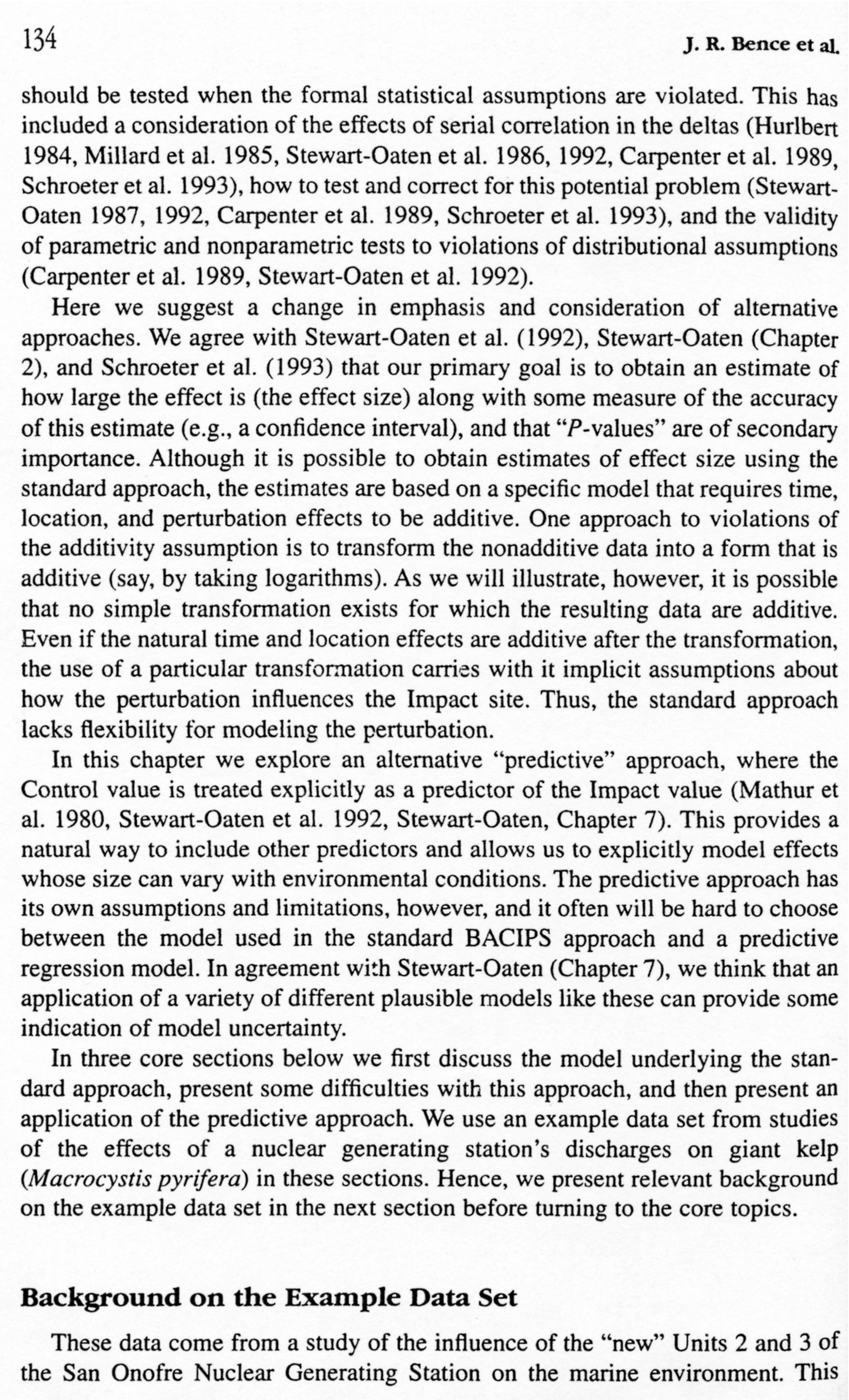
James R. Bence, Allan Stewart-Oaten, and Stephen C. Schroeter

Study of unreplicated perturbations to ecological systems is of practical importance in both applied and basic research. Obviously, after the perturbation has occurred we cannot observe the state of the Impact site in the absence of the perturbation. Nevertheless, our basic goal is to estimate what this condition would have been, and compare this estimate with the observed (perturbed) condition. Here we consider this goal in the context of the Before-After-Control-Impact Paired Series (BACIPS) design (Stewart-Oaten et al. 1986, Chapter 7). In this design, paired samples are collected a number of times, both Before and After the perturbation, simultaneously (or nearly so) at both a Control and Impact location. In what follows we assume that the effect of the perturbation lasts through the After monitoring period, and to streamline the presentation we consider only the simplest case where its magnitude does not show a systematic trend with time, neither growing in size nor dying away.

The basic idea behind the BACIPS design is that there can be natural differences between the Control and Impact sites, and temporal variability operating on a large spatial scale that influences both sites similarly (Stewart-Oaten et al. 1986). By sampling at both Control and Impact on repeated surveys during the Before and After periods, the design "controls" for such natural variation. Heretofore, the standard analytical approach, using the resulting BACIPS data, was to calculate the difference between Control and Impact values (which may be transformations of the original data) on each date (henceforth termed a "delta"), and test whether the mean of these deltas changes from Before to After the perturbation (Stewart-Oaten et al. 1986, 1992, Carpenter et al. 1989).

Previous attention has focused on how the null hypothesis of no difference between the Before and After mean deltas (i.e., no effect of the perturbation)

Detecting Ecological Impacts: Concepts and Applications in Coastal Habitats, edited by R. J. Schmitt and C. W. Osenberg



should be tested when the formal statistical assumptions are violated. This has included a consideration of the effects of serial correlation in the deltas (Hurlbert 1984, Millard et al. 1985, Stewart-Oaten et al. 1986, 1992, Carpenter et al. 1989, Schroeter et al. 1993), how to test and correct for this potential problem (Stewart-Oaten 1987, 1992, Carpenter et al. 1989, Schroeter et al. 1993), and the validity of parametric and nonparametric tests to violations of distributional assumptions (Carpenter et al. 1989, Stewart-Oaten et al. 1992).

Here we suggest a change in emphasis and consideration of alternative approaches. We agree with Stewart-Oaten et al. (1992), Stewart-Oaten (Chapter 2), and Schroeter et al. (1993) that our primary goal is to obtain an estimate of how large the effect is (the effect size) along with some measure of the accuracy of this estimate (e.g., a confidence interval), and that "*P*-values" are of secondary importance. Although it is possible to obtain estimates of effect size using the standard approach, the estimates are based on a specific model that requires time, location, and perturbation effects to be additive. One approach to violations of the additivity assumption is to transform the nonadditive data into a form that is additive (say, by taking logarithms). As we will illustrate, however, it is possible that no simple transformation exists for which the resulting data are additive. Even if the natural time and location effects are additive after the transformation, the use of a particular transformation carries with it implicit assumptions about how the perturbation influences the Impact site. Thus, the standard approach lacks flexibility for modeling the perturbation.

In this chapter we explore an alternative "predictive" approach, where the Control value is treated explicitly as a predictor of the Impact value (Mathur et al. 1980, Stewart-Oaten et al. 1992, Stewart-Oaten, Chapter 7). This provides a natural way to include other predictors and allows us to explicitly model effects whose size can vary with environmental conditions. The predictive approach has its own assumptions and limitations, however, and it often will be hard to choose between the model used in the standard BACIPS approach and a predictive regression model. In agreement with Stewart-Oaten (Chapter 7), we think that an application of a variety of different plausible models like these can provide some indication of model uncertainty.

In three core sections below we first discuss the model underlying the standard approach, present some difficulties with this approach, and then present an application of the predictive approach. We use an example data set from studies of the effects of a nuclear generating station's discharges on giant kelp (*Macrocystis pyrifera*) in these sections. Hence, we present relevant background on the example data set in the next section before turning to the core topics.

Background on the Example Data Set

These data come from a study of the influence of the "new" Units 2 and 3 of the San Onofre Nuclear Generating Station on the marine environment. This facility is located on the southern California coast between Los Angeles and San Diego. The new units became fully operational in May of 1983. *A priori*, the cooling system of these units was predicted to adversely influence the San Onofre kelp forest by increasing particulate flux and reducing the light levels on the ocean's floor within the kelp forest (Murdoch et al. 1980, Ambrose et al., Chapter 18). Giant kelp first settle to the bottom as a microscopic stage and eventually develop and grow to lengths of tens of meters, reaching the sea's surface and forming a canopy. Successful development of the microscopic stages requires that critical levels of light reach the bottom substrate, and ambient levels of light are often near or below the critical levels (Dean and Jacobsen 1984, Deysher and Dean 1984, 1986). Flux of particulates near the bottom can have adverse effects on the early stages of giant kelp through abrasion or burial (Devinny and Volse 1978). Furthermore, successful development of giant kelp in the San Onofre area requires hard substrate, and increased settlement of particulates has the potential to bury hard substrate.

The once-through cooling system of each unit consists of a single intake in shallow water and a diffuser system for returning the seawater extending over the range of depths of the kelp forest, located immediately to the northwest of forest (see Ambrose et al., Figure 18.1). When fully operational, the combined cooling systems of the new units circulate and discharge 100 m^3 per second, and can create a turbid (dirty) plume, both by moving turbid inshore water offshore and by entraining turbid bottom water in the discharge area. The plume of the discharge tends to be moved over the San Onofre kelp forest by the predominant southeast currents. Reductions in light levels (Reitzel et al. 1995), changes in the local current pattern (Elwany et al. 1990), and increases in the particulate settlement rate (Bence et al. 1989), all apparently due to the discharge system, have been reported. Schroeter et al. (1993) report that the discharge of the generating station has led to substantial reductions in the densities of invertebrates associated with the hard bottom of the San Onofre kelp forest.

The example data collected under a BACIPS design are presented in Figure 8.1. These data were collected by side-scan SONAR in the Impact kelp forest (San Onofre) and a Control kelp forest located approximately 5 km northwest of the discharge (San Mateo). Surveys were done at (usually) 6-month intervals over a study period extending from 1978 through 1989. For each survey the side-scan records were examined and maps of different categories of "adult" giant kelp density were constructed (Murdoch et al. 1989). Here we report on areas occupied by moderate and higher density categories, corresponding to densities exceeding approximately 0.04 m^{-2}.

In the following sections these example data are analyzed and manipulated using a variety of methods. The goal of this process is to illustrate an approach for estimating the magnitude of effects due to the power plant. In our opinion, this data set is the best available for estimating such effects because of its relatively long duration of measurements related to kelp density. We stress, however, that the case for adverse effects of the generating station on the kelp forest is

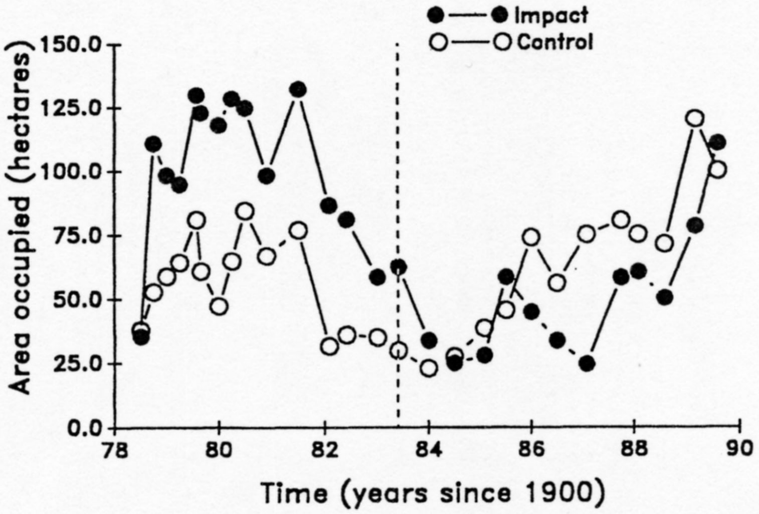


Figure 8.1. Data on areas occupied by densities of giant kelp plants exceeding approximately 0.04 m^{-2}, as determined by side-scan SONAR for the Impact (San Onofre kelp forest) and Control (San Mateo kelp forest) sites over time. Data were collected by Ecosystem Management Associates for 300 survey areas at each site (see Murdoch et al. 1989). Dashed vertical line separates Before and After periods.

based on many more data than those in the example set, including environmental measurements and mechanistic studies cited above, experimental outplants of various stages of giant kelp, other measures of kelp abundance, including counts on fixed transects and estimates from down-looking SONAR, and concomitant studies of potentially confounding factors such as sea urchin abundance and localized changes in the oceanic environment (Bence et al. 1989).

The Standard Approach –The Underlying Model and Implications

The idea that an actual population trajectory over time is a single realization of a stochastic process is central to the standard procedure, and to the alternatives suggested here. This concept is discussed extensively by Stewart-Oaten et al. (1986), and Stewart-Oaten (Chapter 7) and we will not repeat that detailed discussion here. We note that a key consequence is that the actual population value at a given place and time will usually differ from its expected value (the process mean), which itself is time (and space) dependent. Estimates (say of population abundance) can further deviate from this mean because of measurement error. Because replicate observations collected at the same time cannot provide information on the variability of the actual population about its expected value, we treat sampling dates as our level of replication. Of course, within-survey replicates might be collected to reduce sampling error, and our estimate of abundance or other parameters for that survey would then be the average or some other summary of these.

The standard approach assumes "additivity," meaning that in the absence of a perturbation effect the expected value for an observation could be expressed as the sum of time and location effects. We can write this assumption as:

$$\mu_{ij(k)} = L_i + T_{j(k)},$$

where $\mu_{ij(k)}$ is the expected value at location i (I or C for Impact or Control) and time j within the k^{th} period (B or A for Before or After the perturbation), L is the natural location effect and T is the time effect. This additivity assumption implies that the deltas (Impact - Control values) all have the same expected value in the Before period, namely $\Delta_{j(\text{B})} = \Delta_{\text{B}} = L_{\text{I}} - L_{\text{C}}$ for all j. When there is a perturbation effect (i.e., in the After period at the Impact site), we also model this as additive:

$$\mu_{\text{I}j(\text{A})} = L_{\text{I}} + T_{j(\text{A})} + E,$$

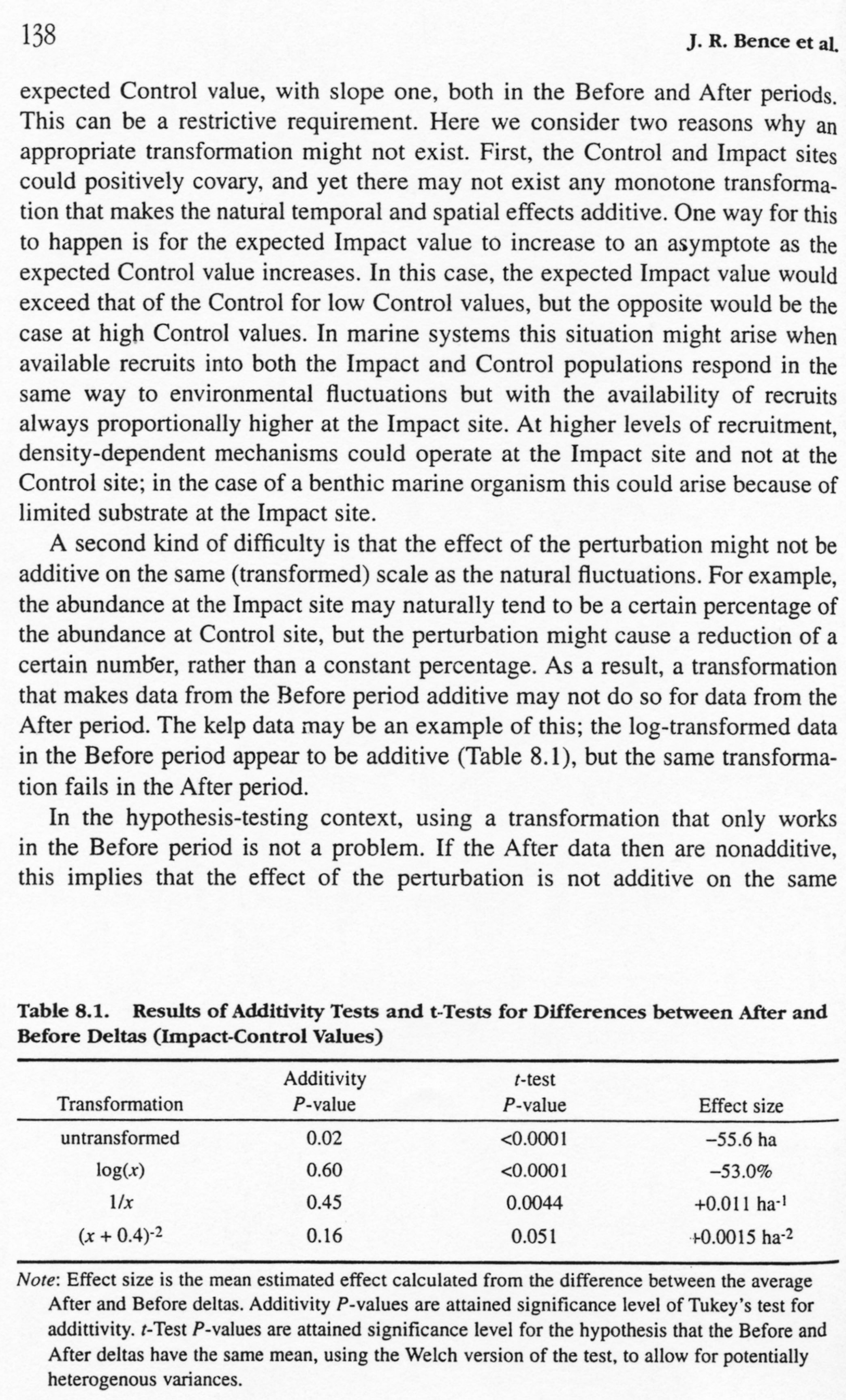
where E is the effect of the perturbation. Therefore $\Delta_{j(\text{A})} = \Delta_{\text{A}} = L_{\text{I}} - L_{\text{C}} + E$. Note that the deltas have the same expected value within periods, but these expected values differ by an amount E between periods. An alternative way of viewing the additivity assumption is to note that it implies a particular linear relationship between expected Impact and Control values, namely $\mu_{\text{I}j(k)} = \Delta_k + \mu_{\text{C}j(k)}$. Thus the slope stays equal to one and the perturbation changes only the intercept.

While the idea of additivity of effects is appealing, it is easy to envision cases where the untransformed data would not be additive. For example, the untransformed data might follow a multiplicative rather than additive model, so that the expected Impact value tends to be a constant multiple of the expected Control value rather than differing by a constant number. In general, failure of the additivity assumption could lead to inefficient tests or to artifactual effects (Stewart-Oaten et al. 1986, 1992). For our multiplicative example, if the Impact value tends to be half the Control value, and overall abundance increases from the Before to After period, the mean Impact - Control delta would decline, even with no effect of the perturbation. We could solve this problem and achieve additivity in this case by taking a logarithmic transformation. More generally, we could consider a class of transformations, say of the Box-Cox form $y = (x + c)^{\lambda}$ (Box and Cox 1964). One could then test for additivity using the Before data, say by the Tukey test for additivity (Tukey 1949), for a range of λ's and c's, and choose a transformation that passes the test.

Difficulties with the Standard Approach

Limited Flexibility of the Standard Approach

As noted above, for estimation we may need to transform the data so that (on the transformed scale) the expected Impact value increases linearly with the



expected Control value, with slope one, both in the Before and After periods. This can be a restrictive requirement. Here we consider two reasons why an appropriate transformation might not exist. First, the Control and Impact sites could positively covary, and yet there may not exist any monotone transformation that makes the natural temporal and spatial effects additive. One way for this to happen is for the expected Impact value to increase to an asymptote as the expected Control value increases. In this case, the expected Impact value would exceed that of the Control for low Control values, but the opposite would be the case at high Control values. In marine systems this situation might arise when available recruits into both the Impact and Control populations respond in the same way to environmental fluctuations but with the availability of recruits always proportionally higher at the Impact site. At higher levels of recruitment, density-dependent mechanisms could operate at the Impact site and not at the Control site; in the case of a benthic marine organism this could arise because of limited substrate at the Impact site.

A second kind of difficulty is that the effect of the perturbation might not be additive on the same (transformed) scale as the natural fluctuations. For example, the abundance at the Impact site may naturally tend to be a certain percentage of the abundance at Control site, but the perturbation might cause a reduction of a certain number, rather than a constant percentage. As a result, a transformation that makes data from the Before period additive may not do so for data from the After period. The kelp data may be an example of this; the log-transformed data in the Before period appear to be additive (Table 8.1), but the same transformation fails in the After period.

In the hypothesis-testing context, using a transformation that only works in the Before period is not a problem. If the After data then are nonadditive, this implies that the effect of the perturbation is not additive on the same

Table 8.1. Results of Additivity Tests and t-Tests for Differences between After and Before Deltas (Impact-Control Values)

Transformation	Additivity *P*-value	*t*-test *P*-value	Effect size
untransformed	0.02	<0.0001	−55.6 ha
log(x)	0.60	<0.0001	−53.0%
$1/x$	0.45	0.0044	+0.011 ha^{-1}
$(x + 0.4)^{-2}$	0.16	0.051	+0.0015 ha^{-2}

Note: Effect size is the mean estimated effect calculated from the difference between the average After and Before deltas. Additivity *P*-values are attained significance level of Tukey's test for addittivity. *t*-Test *P*-values are attained significance level for the hypothesis that the Before and After deltas have the same mean, using the Welch version of the test, to allow for potentially heterogenous variances.

transformation scale that worked for the natural time and location effects. Although such a result indicates that the basic model used by the standard approach is not a realistic portrayal of the system under study, the hypothesis test for a perturbation effect remains valid. This is true because the validity of the test depends on the assumption being true only under the null hypothesis of no effect. However, estimates of effect size, and descriptions of how they vary with conditions, could be markedly off, and we think these should be of primary interest.

Choosing Transformations and Interpreting Effects

Even when a suitable transformation exists, there can still be problems in applying the standard approach. In practice it can be difficult to choose a transformation, or to interpret the effect, which is now measured in the units of the transformed variable. Table 8.1 gives results of *t*-tests used to test for a difference between the mean Before and After deltas for the giant kelp data for four possible treatments of the data. First, the data are untransformed. Second, the data are $\log_e$ transformed. Third, reciprocals are taken (i.e., $1/x$), and fourth we take the Box-Cox transformation $(x + 0.4)^{-2}$. In three cases the *t*-test for an effect is statistically significant at the 0.05 level, and in the fourth it is nearly so. The additivity assumption cannot be rejected for the three transformations, but can be rejected for the untransformed data. Estimated effect sizes are also given in the table.

For the first three treatments of the data these effect sizes have an intuitive and qualitatively consistent interpretation; a specified *decrease* in area (untransformed), a specified percentage *decrease* ($\log_e$ transform), or an *increase* in the resources required to maintain one unit area of kelp forest ($1/x$ transform). The fourth data treatment, an arbitrarily chosen Box-Cox transformation, does not have such an obvious interpretation, although the qualitative result, a reduction in kelp area, is consistent.

Each of the treatments of the data presented above imply that the effect of the perturbation acts to change the relationship between Impact and Control in a particular way. We illustrate this for hypothetical examples (Figure 8.2). In Figure 8.2a we plot relationships between the expected Impact and expected Control values under the assumption that the data follow an additive model and the perturbation causes a decline at the Impact site. We next consider the case where a log-transformation would be appropriate. In this case the expected Impact value is a constant percentage of the expected Control value and this percentage declines by a fixed amount After the perturbation. Here the perturbation causes a change in the slope of the relationship between expected Impact and Control values and the intercept is fixed at zero (Figure 8.2b). The portrayed relationships in Figure 8.2c were chosen as an example where our Box-Cox transformation $(x + 0.4)^{-2}$ would achieve additivity. In the standard approach we suspect that careful evaluation of the implicit assumption of these Impact versus Control relationships is rare.

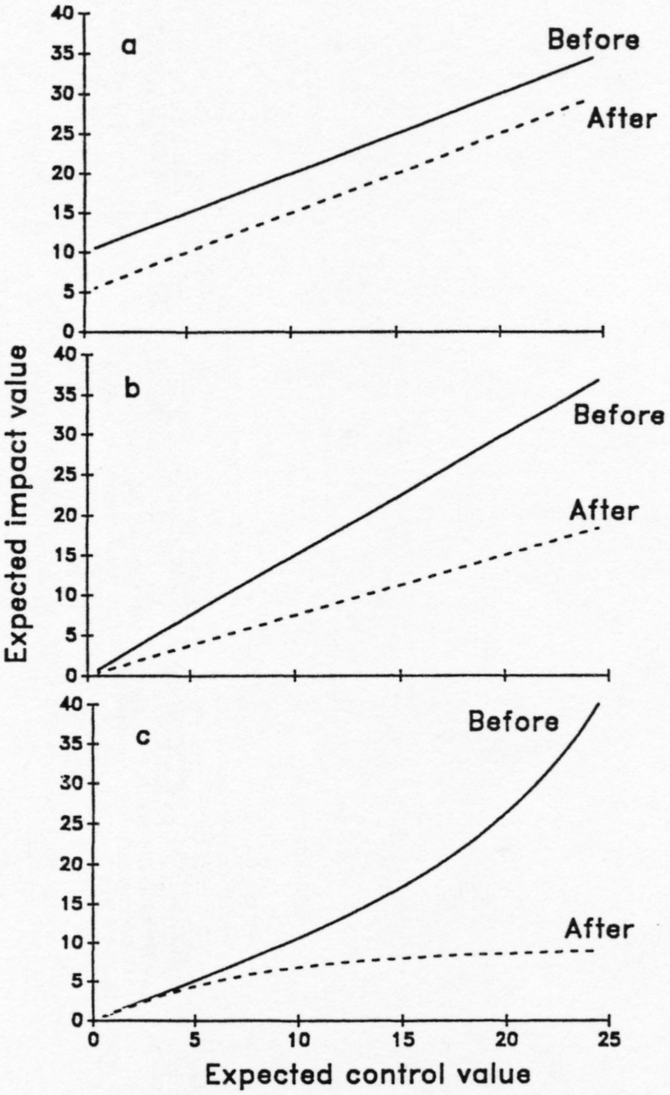


Figure 8.2. Hypothetical relationships between expected Impact and expected Control values. In each case the perturbation has caused a reduction in the expected Impact values. (a) The additive relationship assumed by the standard approach, (b) a multiplicative relationship leading to a constant percentage reduction in the Impact value for all Control values, (c) a relationship chosen so that if the data were $(x + 0.4)^{-2}$ transformed, the transformed data would be additive.

An Alternative: The Predictive Approach

The primary value of the Control is that it acts as a predictor of the Impact value, and here we consider alternative analyses where this concept is an explicit part of the method. The assessment problem can then be thought of as one of comparing two regressions, or fitted functions; the function which best predicts the Impact value from the Control value in the Before period, and the corresponding prediction function for the After period (Stewart-Oaten et al. 1992, Stewart-Oaten, Chapter 7). We could predict the Impact value as a joint function of other variables (e.g., season, current direction, sedimentation) along with the Control value, and could force certain parameters to be the same in both periods (under the assumption that the perturbation did not influence them).

The basic idea behind this approach is that we use the predictive functions to estimate the expected value at the Impact site given the Control for both the Before and After periods. The difference in the expected values, conditional on a particular Control value, is taken as an estimate of the effect of the perturbation (under one set of conditions). This approach attacks the restrictions associated with additivity in two ways. First, any function can be used to model the relationship between Impact and Control values within a period. Second, we can use different functions in the After period than in the Before period, and this allows the perturbation to influence dynamics differently than we might expect from a natural change. A critical implicit assumption is that changes in these conditional expectations (Impact given Control) occur only due to the perturbation. Often this will not be strictly true for reasons related to error in variable problems (e.g., Seber and Wild 1989) (recall that the Control value differs from the expected value of the process that generated it). This is likely to be more of a problem when the distribution of Before and After Control values differ markedly (see Stewart-Oaten, Chapter 7). Because of this difficulty we recommend that the Before and After distributions of Control values be examined to ensure they have similar ranges and variability, and that regression-based estimates be compared with estimates from other models (e.g., the standard BACIPS model).

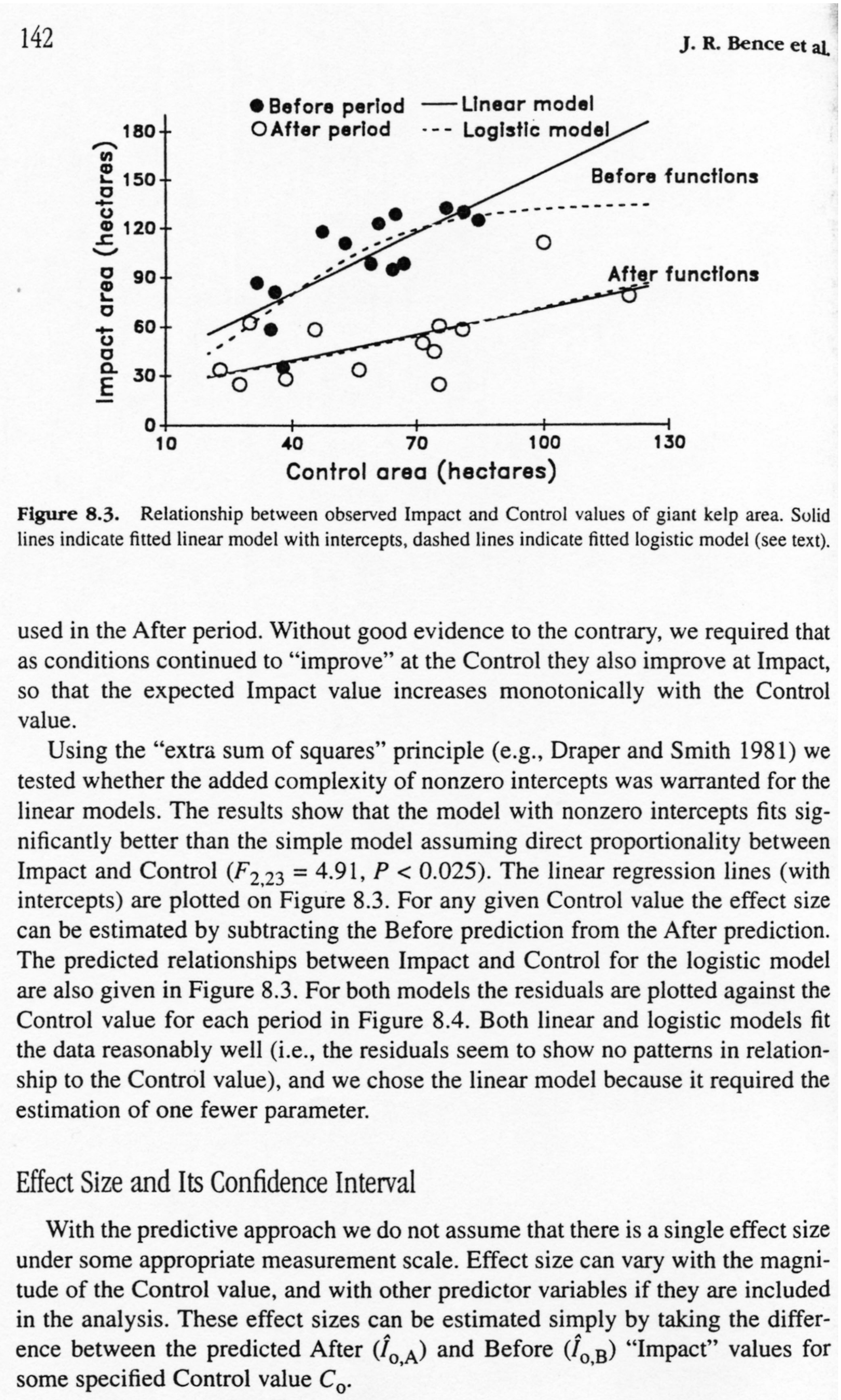
We now illustrate this approach, again using the side-scan SONAR data on area occupied by giant kelp.

Finding Appropriate Functions

Figure 8.3 shows a plot of Impact versus Control areas (same data as Figure 8.1), with Before and After data distinguished. There is a suggestion in these data that the relationship between Impact and Control may be nonlinear in the Before period, with the Impact value reaching an asymptote. Our approach was to fit four models to the data (each separately by period): linear with assumed zero intercepts, linear with intercepts, a quadratic model, and a nonlinear logistic model. For the logistic model we assumed that the asymptote detected during the Before period remained unchanged in the After period. Thus our approach was to fit the logistic model,

$$I = \alpha/(1 + \beta e^{-kC}),$$

to the Before data, then fix the parameter determining the asymptote (α) and estimate the remaining two parameters based on the After data. We discarded the quadratic model early on because it predicts a dome-shaped relationship between Impact and Control for the Before period over the range the function needs to be



● Before period
○ After period
— Linear model
--- Logistic model
Before functions
After functions
Impact area (hectares)
Control area (hectares)
0
30
60
90
120
150
180
10
40
70
100
130

Figure 8.3. Relationship between observed Impact and Control values of giant kelp area. Solid lines indicate fitted linear model with intercepts, dashed lines indicate fitted logistic model (see text).

used in the After period. Without good evidence to the contrary, we required that as conditions continued to "improve" at the Control they also improve at Impact, so that the expected Impact value increases monotonically with the Control value.

Using the "extra sum of squares" principle (e.g., Draper and Smith 1981) we tested whether the added complexity of nonzero intercepts was warranted for the linear models. The results show that the model with nonzero intercepts fits significantly better than the simple model assuming direct proportionality between Impact and Control ($F_{2,23} = 4.91$, $P < 0.025$). The linear regression lines (with intercepts) are plotted on Figure 8.3. For any given Control value the effect size can be estimated by subtracting the Before prediction from the After prediction. The predicted relationships between Impact and Control for the logistic model are also given in Figure 8.3. For both models the residuals are plotted against the Control value for each period in Figure 8.4. Both linear and logistic models fit the data reasonably well (i.e., the residuals seem to show no patterns in relationship to the Control value), and we chose the linear model because it required the estimation of one fewer parameter.

Effect Size and Its Confidence Interval

With the predictive approach we do not assume that there is a single effect size under some appropriate measurement scale. Effect size can vary with the magnitude of the Control value, and with other predictor variables if they are included in the analysis. These effect sizes can be estimated simply by taking the difference between the predicted After ($\hat{I}_{o,A}$) and Before ($\hat{I}_{o,B}$) "Impact" values for some specified Control value C_o.

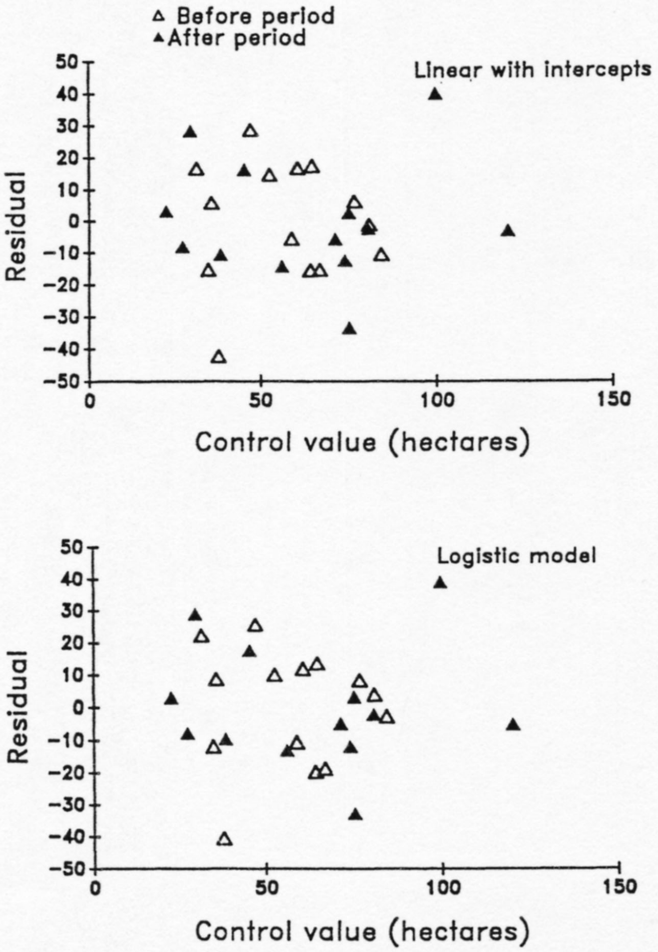


Figure 8.4. Residuals from linear model with intercepts and logistic model (see text) versus Control value. Before residuals are indicated by open triangles, After residuals by solid triangles.

We constructed approximate (1 − α) confidence intervals for the estimated effects at a specified Control value, $\hat{E}_o = \hat{I}_{o,B} - \hat{I}_{o,A}$, for the linear model with intercepts as

$$\hat{E}_o \pm t^{\frac{\alpha}{2}}_{\nu_o}\hat{\sigma}_o$$

where t refers to critical value of the t distribution, ν_o indicates the degrees of freedom, and σ_o is the estimated standard deviation of the estimated effect at C_o. σ_o was calculated as $\sqrt{(\hat{\sigma}^2_{o,B} + \hat{\sigma}^2_{o,A})}$ where $\hat{\sigma}_{o,B}$ and $\hat{\sigma}_{o,A}$ are the standard deviations of $\hat{I}_{o,B}$ and $\hat{I}_{o,A}$, which were calculated using standard regression approaches (e.g., Draper and Smith 1981). Taking into account that the standard deviations of the estimates of the expected Impact value can differ between the periods, ν_o can be estimated with a Satterthwaite approximation (e.g., Ames and Webster 1991):

$$\nu_0 = \frac{(\hat{\sigma}^2_{o,B} + \hat{\sigma}^2_{o,A})^2}{\dfrac{\hat{\sigma}^4_{o,B}}{\nu_B} + \dfrac{\hat{\sigma}^4_{o,A}}{\nu_A}}$$

where ν_k is the degrees of freedom (n - number of parameters estimated) associated with period k. Note that our approach for constructing confidence intervals does not require that the prediction equations be linear; approximate confidence intervals can be calculated whenever estimates of the appropriate variances and degrees of freedom are available.

Estimated effect sizes (still for the linear model with intercepts) as a function of Control values, along with their approximate 95% confidence intervals, are plotted in Figure 8.5. Effect size increases with the Control value and the effect size as a percentage of the unaffected Impact value (i.e., the Before prediction) increases modestly with the Control value.

Although effect size can depend upon the value of the Control or other variables, the concept of an average or net effect size (and its associated confidence interval) is of practical importance. For example, some estimate of the average percent loss might be required in order to implement a mitigation plan. Then an artificial reef that is expected to provide an equal amount of kelp could be designed. In this situation some consideration needs to be given to what we mean by an "average" effect. Here we are interested in the average long-term effect, so we generate an estimate of the long-term distribution of Control values, and calculate the expected loss (both in area and as a percentage) over this distribution. Our distribution of Control values is taken as given, and is assumed to be the

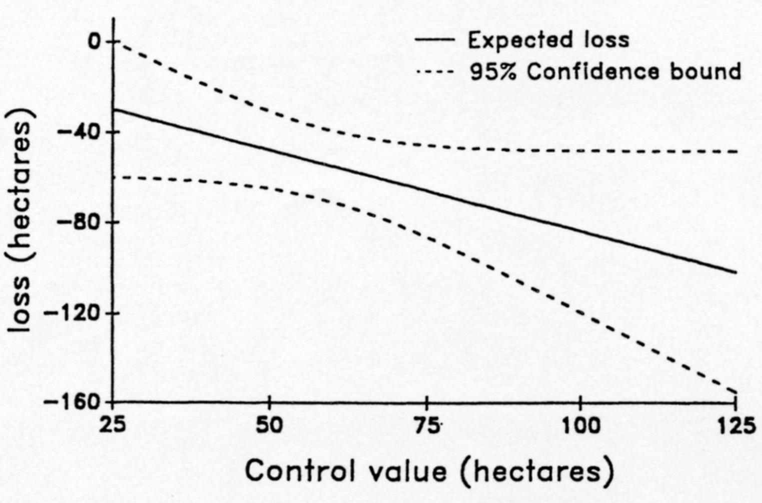


Figure 8.5. Estimated expected loss (effect size) in kelp area as a function of Control value, with 95% confidence intervals, for the linear model with intercepts.

observed set of Control values over the 11-year study including both Before and After data. Based on this set of Control values, the estimated average loss was 55.0 ha, or 52.0%. A confidence interval for the average losses was generated using a nonparametric jackknife method (e.g., Miller 1974). This method requires leaving out each of the 27 surveys one at a time, refitting the regression models with the data point deleted and calculating the average reduction leaving out the selected observation in this calculation also. The variation in these averages was used, following usual jackknife procedures, to calculate a confidence interval. The confidence intervals were (-71.1, -38.0) ha, and (-65.3, -39.2)%. This estimate of "net effect" and its confidence interval was similar to what we obtained for the standard approach using log-transformed data (effect = -53.0%, 95% confidence interval = (-65.7, -35.7)%), although we stress that this is not a guaranteed result (see Discussion).

The Independence Assumption

Although the problem of independence is not peculiar to the predictive approach, the possibility of violating this assumption needs to be considered in virtually all applications, including our example. To this point all our estimation and hypothesis testing have been done under the assumption that residual errors are uncorrelated. If there is substantial autocorrelation then (i) hypothesis tests given above are biased, (ii) the ordinary least squares estimates are not the most efficient, and (iii) we should take the true error structure into account when calculating confidence intervals. Within the context of general linear models, the usual approach is to test for first-order autocorrelation using the Durbin-Watson statistic Q. For the linear model with intercepts, the estimated first order autocorrelation was negative, small in magnitude (-0.091), and not statistically significant by the Durbin-Watson test ($Q = 1.81$, $P > 0.05$: the null distribution of the Durbin-Watson statistic was approximated by matching the first three moments of a beta distribution to $a + bQ$ for suitably chosen a and b, as suggested by Henshaw (1966) and evaluated by Durbin and Watson (1971)). Thus the available evidence indicates that first-order autocorrelation of the residuals is relatively weak for the kelp example.

Other types of violations of the independence assumption are possible. For example, residuals within a year (or some large block of time) could be correlated. Rather than performing many specific tests (none of which we have a good *a priori* basis for), we have plotted the residuals (still for the logistic model) against sampling date. There is some suggestion that the effect might be increasing during the first few surveys of the After period (i.e., the residuals are decreasing), but otherwise there are no obvious patterns (Figure 8.6).

We conclude this section by noting that there could still be weak autocorrelation that we failed to detect, or violations of the independence assumption in ways we have failed to consider. This is one more reason to treat confidence intervals only as approximate guides in interpretation.

Discussion

In the predictive approach we consider a Control to be a predictor, perhaps combined with others, of an Impact value. By comparing predictions based on data collected Before a perturbation with predictions based on data from the After period we can estimate the magnitude of an effect. We have contrasted this approach with the standard (BACIPS *t*-test) approach. The predictive approach is more flexible, in the sense that it can deal with effects of a perturbation, or natural temporal and spatial effects, that are not additive. In another sense, however, our implementation of the predictive approach is more restrictive; it assumes that changes in the expected value of Impact given the observed Control will change only due to effects of the perturbation [see Stewart-Oaten (Chapter 7) for discussion of this assumption]. Because the different models make different assumptions, and it may be hard to choose between them, we recommend that evaluations include estimates of effects based on several different models or approaches. When estimates from different methods are similar this will increase confidence in the conclusions, while large differences would indicate that model uncertainty may add greatly to the stated uncertainty obtained from any single method (Stewart-Oaten, Chapter 7).

For the kelp example, the estimated net effect is not appreciably different when using the predictive approach than for the standard approach on log-transformed data. This suggests that the formal estimates of effects and their confidence intervals are not highly dependent on arbitrary properties of a particular model. In large part, this favorable result occurred because the predictive approach yielded Impact-Control relationships that were similar to those implied

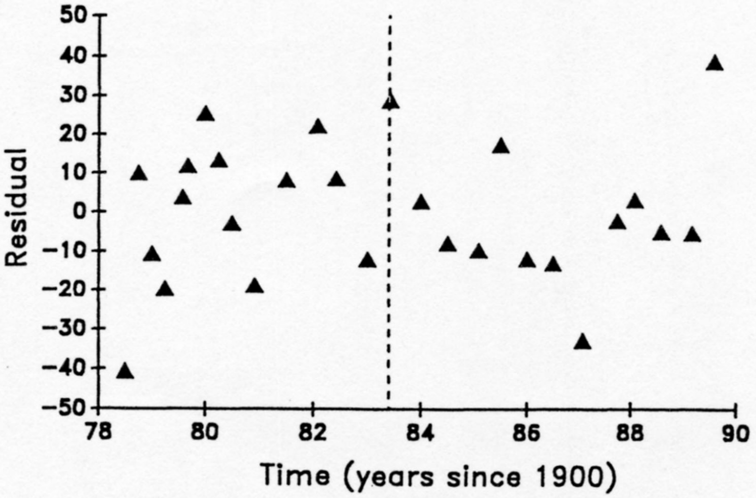


Figure 8.6. Residuals from linear regression (with intercepts) between Impact and Control giant kelp areas plotted against time. Dashed vertical line separates Before and After periods.

by the standard BACI *t*-test model applied to log-transformed data (compare Figures 8.2b and 8.3). This similarity, of course, is not guaranteed.

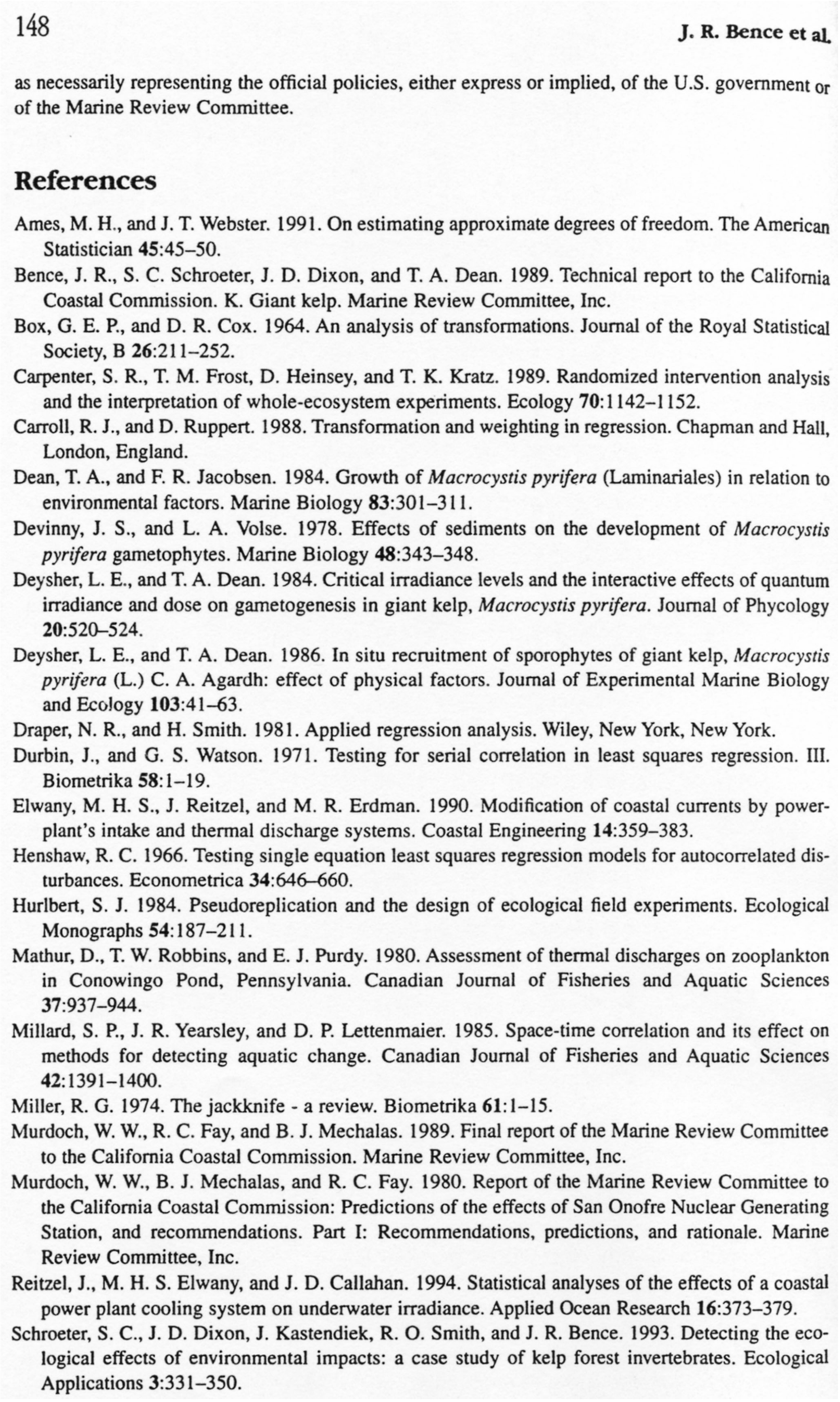
Use of the predictive approach may help us rule out some models. When using the predictive approach we are explicit about the fact that the Control is a predictor. This naturally leads us to think about whether our data fit the regression (or other model) we use, and whether our estimates of effects in the After period require us to predict outside the bounds of the Before data. This is possible to some degree when using a *t*-test, but there is nothing inherent in the test to promote plotting Impact versus Control, or to look at whether the relationship changes at extreme values. In addition, the course of action when lack of fit is observed may not be obvious for the *t*-test. This is still tricky in regression analyses but is a standard part of the approach (e.g., Draper and Smith 1981, Carroll and Ruppert 1988, Seber and Wild 1989).

We know of two examples in the literature where a form of the predictive approach has been used. Mathur et al. (1980) tested for differences in zooplankton abundance between Before and After operations of a power plant's cooling system started by analysis of covariance, using temperature, stream water flow and abundance at a Control site as covariates. Reitzel et al. (1995) analyzed irradiance data from the San Onofre study; they stratified the data on the basis of current direction and followed the standard approach (*t* test on deltas) within the strata. This stratification by current direction allowed them to detect effects whose sign depended upon current direction. Both of these studies incorporated elements of the predictive approach, which we think provided insight beyond what could have been obtained based only on a test comparing the means of the Before and After deltas. We encourage an even more flexible and exploratory approach.

We conclude this largely statistical chapter by stressing that statistics can be only part of the equation. We recognize that any analysis will rest on assumptions, not all of which can be adequately tested (Schroeter et al. 1993, Stewart-Oaten, Chapter 7). In the end, decisions about the reality and importance of an apparent effect should depend upon the weight of all the available evidence–including the results of mechanistic studies, consideration of potential alternative explanations, and consistency among different sets of data–not just the *P*-value from a single test, or even estimates of the effect and associated confidence intervals.

Acknowledgments

R. Ambrose and W. Murdoch provided comments as our work progressed and R. Schmitt and C. Osenberg provided input on earlier drafts of this chapter. This work was supported in part by the Minerals Management Service, U.S. Department of the Interior, under MMS Agreement No. 14-35-0001-30471 (The Southern California Educational Initiative) and by the Marine Review Committee. The side-scan SONAR data were collected by Ecosystems Management Associates under a cooperative agreement between the Marine Review Committee and Southern California Edison. The views and interpretation contained in this document are those of the authors and should not be interpreted



as necessarily representing the official policies, either express or implied, of the U.S. government or of the Marine Review Committee.

References

Ames, M. H., and J. T. Webster. 1991. On estimating approximate degrees of freedom. The American Statistician **45**:45–50.

Bence, J. R., S. C. Schroeter, J. D. Dixon, and T. A. Dean. 1989. Technical report to the California Coastal Commission. K. Giant kelp. Marine Review Committee, Inc.

Box, G. E. P., and D. R. Cox. 1964. An analysis of transformations. Journal of the Royal Statistical Society, B **26**:211–252.

Carpenter, S. R., T. M. Frost, D. Heinsey, and T. K. Kratz. 1989. Randomized intervention analysis and the interpretation of whole-ecosystem experiments. Ecology **70**:1142–1152.

Carroll, R. J., and D. Ruppert. 1988. Transformation and weighting in regression. Chapman and Hall, London, England.

Dean, T. A., and F. R. Jacobsen. 1984. Growth of *Macrocystis pyrifera* (Laminariales) in relation to environmental factors. Marine Biology **83**:301–311.

Devinny, J. S., and L. A. Volse. 1978. Effects of sediments on the development of *Macrocystis pyrifera* gametophytes. Marine Biology **48**:343–348.

Deysher, L. E., and T. A. Dean. 1984. Critical irradiance levels and the interactive effects of quantum irradiance and dose on gametogenesis in giant kelp, *Macrocystis pyrifera*. Journal of Phycology **20**:520–524.

Deysher, L. E., and T. A. Dean. 1986. In situ recruitment of sporophytes of giant kelp, *Macrocystis pyrifera* (L.) C. A. Agardh: effect of physical factors. Journal of Experimental Marine Biology and Ecology **103**:41–63.

Draper, N. R., and H. Smith. 1981. Applied regression analysis. Wiley, New York, New York.

Durbin, J., and G. S. Watson. 1971. Testing for serial correlation in least squares regression. III. Biometrika **58**:1–19.

Elwany, M. H. S., J. Reitzel, and M. R. Erdman. 1990. Modification of coastal currents by power-plant's intake and thermal discharge systems. Coastal Engineering **14**:359–383.

Henshaw, R. C. 1966. Testing single equation least squares regression models for autocorrelated disturbances. Econometrica **34**:646–660.

Hurlbert, S. J. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs **54**:187–211.

Mathur, D., T. W. Robbins, and E. J. Purdy. 1980. Assessment of thermal discharges on zooplankton in Conowingo Pond, Pennsylvania. Canadian Journal of Fisheries and Aquatic Sciences **37**:937–944.

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

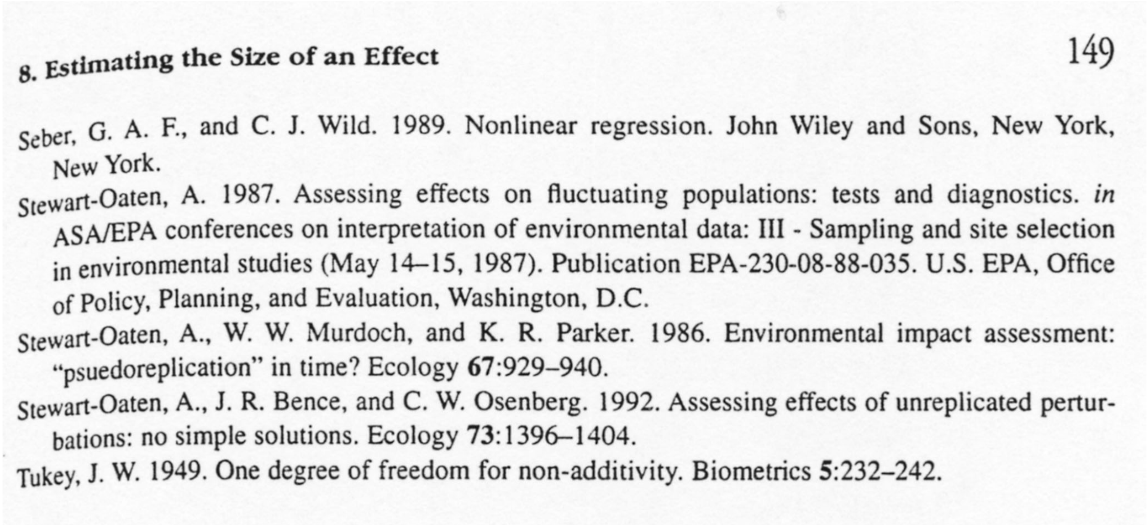
Miller, R. G. 1974. The jackknife - a review. Biometrika **61**:1–15.

Murdoch, W. W., R. C. Fay, and B. J. Mechalas. 1989. Final report of the Marine Review Committee to the California Coastal Commission. Marine Review Committee, Inc.

Murdoch, W. W., B. J. Mechalas, and R. C. Fay. 1980. Report of the Marine Review Committee to the California Coastal Commission: Predictions of the effects of San Onofre Nuclear Generating Station, and recommendations. Part I: Recommendations, predictions, and rationale. Marine Review Committee, Inc.

Reitzel, J., M. H. S. Elwany, and J. D. Callahan. 1994. Statistical analyses of the effects of a coastal power plant cooling system on underwater irradiance. Applied Ocean Research **16**:373–379.

Schroeter, S. C., J. D. Dixon, J. Kastendiek, R. O. Smith, and J. R. Bence. 1993. Detecting the ecological effects of environmental impacts: a case study of kelp forest invertebrates. Ecological Applications **3**:331–350.



Seber, G. A. F., and C. J. Wild. 1989. Nonlinear regression. John Wiley and Sons, New York, New York.

Stewart-Oaten, A. 1987. Assessing effects on fluctuating populations: tests and diagnostics. *in* ASA/EPA conferences on interpretation of environmental data: III - Sampling and site selection in environmental studies (May 14–15, 1987). Publication EPA-230-08-88-035. U.S. EPA, Office of Policy, Planning, and Evaluation, Washington, D.C.

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

Stewart-Oaten, A., J. R. Bence, and C. W. Osenberg. 1992. Assessing effects of unreplicated perturbations: no simple solutions. Ecology **73**:1396–1404.

Tukey, J. W. 1949. One degree of freedom for non-additivity. Biometrics **5**:232–242.

BIOMETRICS 52, 38–49
March 1996

Sequential Estimation of Log(Abundance)

Allan Stewart-Oaten

Biology Department, University of California,
Santa Barbara, California 93106, U.S.A.

SUMMARY

I discuss the estimation of the abundance of a biological population, its logarithm, and the variances of these estimates, from a sequential sampling scheme with minimum and maximum sample sizes. Observations are counts of organisms in randomly chosen "packets" such as cores, branches, bushes, and so forth. For preassigned values m, n_1 and n_2, samples are taken until (a) at least n_1 packets and (b) either m positive packets or a total of n_2 packets have been observed.

Abundance estimates are based on an estimate of the fraction of positive packets given by Kremers (1987, *Technometrics* **29,** 109–112), with a modification to avoid estimates of zero. Estimates of log abundance are given by log(estimated abundance) with an adjustment for bias due to the concavity of the log function. Two adjustments are considered, one based on Taylor series expansion (the delta method) and the other on the bootstrap. These techniques are also used to estimate the variance of the estimate of log(abundance). Simulations suggest that both methods are better than not adjusting, though the gain is small compared to the standard deviation of the estimates. The bootstrap estimates are less biased than the Taylor series estimates but have larger variances, so that the Taylor series estimates have smaller mean squared errors. The variances of the sequential estimates of log(abundance) tend to be only weakly dependent on the true abundance.

1. Introduction

Population abundances of species are perhaps the fundamental currency of ecology. Most measures of environments or communities are functions of them. Their estimation under various conditions is one of the major tasks of biometry and statistical ecology.

Assessments of change, or comparisons of the temporal variability of populations of different species, or of populations in different places, can be distorted because of differences in overall abundances. A natural way to reduce or eliminate such distortions is to base the comparisons on log(abundance). For example, Williamson (1984) found that the standard deviation of $\log(\hat{\alpha})$, where $\hat{\alpha}$ is the estimated population abundance, has been the most popular measure of temporal variability. Because the simplest models of population growth are exponential, plots of log(abundance) against time are likely to be informative about rates. Also, one way to assess the local effect of a planned intervention on a species is to compare the mean (over time) of the difference between the affected population and the population in a nearby unaffected area, before the intervention, to the mean difference after the intervention; if temporal variation tends to be multiplicative, use of log(abundance) may be more efficient and more valid than use of raw abundance (e.g., Stewart-Oaten, Murdoch, and Parker 1986).

In virtually all practical cases, the true population abundance is not known but must be estimated from samples. Three problems arise from this.

(i) The estimated abundance may be zero, and log(0) is undefined. This is a common problem, with no clear resolution. The usual response is to use either $\log(\hat{\alpha}+c)$ or $\max\{\log(\hat{\alpha}), \log(c)\}$ for some constant, c. The choice of c can have a strong influence on results of analyses, but it is largely arbitrary: The best known formal method (Box and Cox, 1964) does not perform well in this case (Atkinson 1985; Berry, 1987), and other arguments have been *ad hoc* ("pretend we saw half an extra animal"), "rather recondite" (and not given: Mosteller and Tukey, 1977),

Key words: Abundance; Bias; Bootstrap; Delta method; Sequential sampling; Zeros.

or, perhaps best, based on minimizing the effects of nuisance parameters in particular models (Anscombe 1948; Berry, 1987).

(ii) A fixed sampling effort often leads to greater errors in the estimate of log(abundance) when abundance is small than when it is large. If the variance of $\hat{\alpha}$ is σ^2 and zeros can be avoided, the variance of $\log(\hat{\alpha})$ is approximately σ^2/α^2, where α is the true abundance. If $\sigma^2 = b\alpha^2$ for some constant b (a case of Taylor's [1961] "Power Law"), the variance of $\log(\hat{\alpha})$ is approximately independent of α, but this assumption may not hold at small abundances or for sampling on small scales. If a fraction α of possible quadrats has one animal each, while the rest are empty, then σ^2 is approximately proportional to α, not to α^2, and the variance of $\log(\hat{\alpha})$ will be larger for small α.

(iii) The log of a positive unbiased estimate of α, the true abundance, will not be an unbiased estimate of $\log(\alpha)$, but of something smaller, due to Jensen's Inequality (Feller, 1966, p. 152). By Taylor series expansion, the bias is roughly $-\sigma^2/2\alpha^2$, so $\log(\alpha)$ may be underestimated by more when α is small than when it is large, in view of problem (ii). This could affect estimates and comparisons of temporal variability of $\log(\alpha)$.

Problems (i) and (ii) suggest sequential sampling, to increase the sampling effort at low abundances and decrease the likelihood of sample zeros, without prior knowledge of the (often highly variable) abundance. In Section 2, I propose a positive, "almost" unbiased estimator of abundance, based on a sequential sampling scheme described by Kim and Nachlas (1984) and by Kremers (1987). In Section 3, I describe two ways of adjusting the log of this estimator for an "almost" unbiased estimate of log(abundance), one based on Taylor series expansion and the other on the bootstrap. These methods also yield estimates of the variance of the estimator. Section 4 gives an example, and Section 5 reports some numerical results.

2. Sequential Estimation of the Abundance

Our samples consist of counts of animals in randomly chosen "packets." The packets might be quadrats; aliquots; net hauls, which "sieve" equal volumes of water; core samples which remove equal volumes of sand or soil; or twigs or plants of approximately equal size. Write $C_1, C_2, \ldots,$ for the counts on the 1st, 2nd, ..., sample packets. Define

$$\alpha = EC_i = \text{the mean number of animals per packet in the habitat} \tag{2.1}$$

where "the habitat" is the region from which the sampled packets are randomly chosen. A standard sequential sampling plan is to keep sampling until a specified number of positive packets have been observed. This is not completely realistic: one must stop sampling eventually. We may also want to specify a minimum number of packets; if we do not do this, then the minimum sample size is the required number of positive packets, and this may be too small for estimating variances or other distributional parameters.

I consider the following sequential scheme.

(i) Choose

$$\begin{aligned} n_1 &= \text{minimum sample size,} \\ n_2 &= \text{maximum sample size,} \\ m &= \text{minimum number of positive packets required.} \end{aligned} \tag{2.2}$$

(ii) Randomly sample n_1 packets. If these contain at least m positive packets, stop sampling. Otherwise, continue sampling until either there are m positive packets or there are n_2 packets altogether. Let

$$\begin{aligned} N &= \text{the number of packets actually obtained,} \\ N_+ &= \text{the number of positive packets actually obtained.} \end{aligned} \tag{2.3}$$

(iii) Let

$$\hat{P}_K = \begin{cases} N_+/n_1 & \text{if } N = n_1, \\ N_+/n_2 & \text{if } N = n_2 \text{ and } 0 \le N_+ < m, \\ (m-1)/(N-1) & \text{otherwise.} \end{cases} \tag{2.4}$$

Thus, $\hat{P}_K$ is the usual sequential sample estimate of

$$p = \text{the probability of a positive packet} \tag{1.5}$$

if the sequential rule causes the sampling to stop at $n_1 < N \leq n_2$ and is, otherwise, the usual fixed sample size estimate, for a sample of n_1 or n_2. To avoid estimates of zero, we make an arbitrary choice between 0 and the smallest obtainable positive value when $N_+ = 0$. Our biased estimate is

$$\hat{P}_b = \begin{cases} \hat{P}_K & \text{if } N_+ > 0, \\ 1/2n_2 & \text{if } N = n_2 \text{ and } N_+ = 0. \end{cases} \tag{2.6}$$

(iv) Let

$$C_i = \text{the number of organisms counted in the } i\text{th sampled packet,}$$

$$C_+ = \begin{cases} \sum C_i/N_+ & \text{if } N_+ > 0, \text{ and} \\ 1 & \text{if } N_+ = 0. \end{cases} \tag{2.7}$$

Thus, C_+ is the average number of animals found in the positive packets, with the minimum possible value used if none are observed.

(v) Our (biased) estimator of the abundance, α, is

$$\hat{\alpha}_b = C_+\hat{p}_b. \tag{2.8}$$

The sampling scheme in steps (i) and (ii) is a special case of that proposed by Kim and Nachlas (1984). Their scheme is more general in that it allows a minimum required number of zero packets, as well as of positive packets. The estimator $\hat{p}_K$, in step (iii), was given by Kremers (1987), who showed it to be unbiased, and an improvement over the estimator of Kim and Nachlas (1984). Steps (iv) and (v) are obvious extensions to the case where we wish to estimate the average abundance per packet = (proportion of positive packets) × (average abundance per positive packet).

It is easy to show that $A_K = C_+\hat{p}_K$ is an unbiased estimator of α. Thus, the bias in $\hat{\alpha}_b$, as given by (2.8), is $P\{C_+ = 0\}/2n_2$.

(vi) The variance of $\hat{\alpha}_b$ is close to that of A_K unless p is very close to 0. It can be shown that

$$V(A_K) = \sigma^2_{C+}E\{{p_K}^2/x\} + \alpha^2_+V\{\hat{p}_K\}, \tag{2.9}$$

where α_+ and σ^2_{C+} are the mean and variance of the numbers in positive packets (i.e., of C_i given $C_i > 0$). We estimate $V\{\hat{\alpha}_b\}$ by substituting estimates for the four terms in (2.9).

We estimate $\sigma^2_{C_+}$ by

$$s^2_{C+} = \begin{cases} \text{sample variance of positive packets} & \text{if } N_+ > 1, \\ C_+(1 - \exp(1 - C_+)) & \text{otherwise,} \end{cases} \tag{2.10}$$

where C_+ is given by (1.7). The estimate for $N_+ \leq 1$ assumes a truncated Poisson distribution for the positive packets.

We estimate α^2_+ by $C^2_+ - s^2_{C+}/N_+$, using the relations $E\{C^2_+\} = \alpha^2_+ + V\{C_+\}$ and $V\{C_+\} = \sigma^2_{C_+}/N_+$ (because C_+ is an average of N_+ observations).

Formulae for $V\{\hat{p}_K\}$ and $E\{\hat{p}^2_K/N_+\}$ are messier. Kremers (1987) showed that $V\{\hat{p}_K\}$ is the same as the variance of the estimator of Kim and Nachlas (1984), but with $n_1 + 1$ and $n_2 + 1$ replacing n_1 and n_2. However, the formula given by Kim and Nachlas is incorrect: It is actually appropriate for Kremers' estimator. For the general case with m_1 nonzero and m_2 zero packets required, straightforward bookkeeping gives

$$\begin{aligned} V\{\hat{p}_K\} = {} & \sum_{i=0}^{m_1-1} [i/n_2]^2 b(i|n_2,p) + \sum_{i=n_2-m_2+1}^{n_2} [i/n_2]^2 b(i|n_2,p) + \sum_{i=m_1}^{n_1-m_2} [i/n_1]^2 b(i|n_1,p) \\ & + \sum_{i=n_1+1}^{n_2} [(m_1-1)/(i-1)]^2 W(i|m_1,p) + \sum_{i=n_1+1}^{n_2} [(i-m_2)/(i-1)]^2 W(i|m_2,q) - p^2 \end{aligned} \tag{2.11}$$

and

$$E\{\hat{p}_K^2/N_+\} = \sum_{i=0}^{m_1-1} ib(i|n_2,p)/n_2^2 + \sum_{i=n_2-m_2+1}^{n_2} ib(i|n_2,p)/n_2^2 + \sum_{i=m_1}^{n_1-m_2} ib(i|n_1,p)/n_1^2$$
$$+ \sum_{i=n_1+1}^{n_2} (m_1-1)^2 W(i|m_1,p)/[m_1(i-1)^2] + \sum_{i=n_1+1}^{n_2} (i-m_2)W(i|m_2,q)/(i-1)^2, \tag{2.12}$$

where $q = 1 - p$,

$$b(i|n,p) = n!p^i q^{n-i}/[i!(n-i)!], \tag{2.13}$$

the probability of i successes in n binomial trials, and

$$W(i|n,p) = (i-1)!p^n q^{i-n}/[(n-1)!(i-n)!], \tag{2.14}$$

the probability of waiting until the ith trial to obtain the nth success. Thus, our estimate of the variance of $\hat{\alpha}_b$ is

$$s^2 = s_{C+}^2 \hat{E}\{\hat{p}_K^2/N_+\} + (C_+^2 - s_{C+}^2/N_+)\hat{V}\{\hat{p}_K\}, \tag{2.15}$$

where $\hat{V}\{\hat{p}_K\}$ and $\hat{E}\{\hat{p}_K^2/N_+\}$ are obtained by replacing "p" and "m_2" by $\hat{p}_K$ and 0 in (2.11) and (2.12).

If the maximum allowable sample size, n_2, is large, these formulae involve formidable hand-calculating, although negligible computer time. They can be simplified by ignoring or approximating sums of small terms. The simplest estimates, $\hat{p}_K(1-\hat{p}_K)/n_1$ for $\hat{V}\{\hat{p}_K\}$ and $\hat{p}_K/n_1$ for $E\{\hat{p}_K^2/N_+\}$ are negligibly different from the messy formulae if $N = n_1$, that is, sampling stops at the minimum sample. For larger N, an approximation based on equation (4.13) of DeGroot (1959) seems very accurate.

3. Estimation of Log(Abundance)

Even if the bias in $\hat{\alpha}_b$, as an estimator of α, can be regarded as negligible, $\log(\hat{\alpha}_b)$ would still have a negative bias as an estimator of $\log(\alpha)$. We also usually want an estimate of the variance of our estimate of log(a). This section describes two methods of obtaining these estimates. Throughout, "log" refers to the natural logarithm, to base e. Our interest is in $\log(\alpha) = \log(E\{C_i\})$, not in $E\{\log(C_i)\}$: in most practical cases, the latter is undefined, because $P\{C_i = 0\} > 0$; in other cases, it is likely to depend nonlinearly on the size of the sampling unit—the quadrat, core, aliquot, and so forth,—which is usually arbitrary.

Taylor series expansion. If $\varepsilon = \hat{\alpha}_b - \alpha$ is small compared to α, then $\log(\hat{\alpha}_b) \approx \log(\alpha) + \varepsilon/\alpha - \varepsilon^2/2\alpha^2 + \cdots$. This suggests that, if $E\{\varepsilon\} = 0$ and $E\{s^2\} \approx \sigma^2 = V(\hat{\alpha}_b)$, where s^2 is given in (2.15) and E and V indicate expectation and variance, then an approximately unbiased estimator of $\log(\alpha)$ is

$$\hat{L}_{TS} = \log(\hat{\alpha}_b) + s^2/2\hat{\alpha}_b^2, \tag{3.1}$$

and that the variance of the first term can be estimated by an estimate of $V\{\varepsilon/\alpha\}$:

$$\hat{V}_{TS} = s^2/\hat{\alpha}_b^2. \tag{3.2}$$

Equations (3.1) and (3.2) are obtained by substituting $s^2/\hat{\alpha}_b^2$ for σ^2/α^2. There are *ad hoc* arguments leading to other estimates of σ^2/α^2—for example, "$E\{1/\hat{\alpha}_b^2\} \approx 1/\alpha^2 + 3\sigma^2/\alpha^4$, so $\sigma^2/\alpha^2 \approx (\sqrt{\{1+12s^2/\hat{\alpha}_b^2\}} - 1)/6$." These did not fare better than $s^2/\hat{\alpha}_b^2$ in numerical simulations.

Bootstrap estimates. A functional, $\lambda(F)$, of the (unknown) distribution function, F, can be estimated by drawing a sample, $\boldsymbol{y}$, of observations from F according to a design, D, using it to compute an estimate, $F^*(\boldsymbol{y}) = \hat{F}$, of F and computing the estimate $\hat{\lambda} = \lambda(\hat{F})$. The distribution of $\hat{\lambda}$ can be estimated by that of the variable $\hat{\lambda}_1$, which is obtained in exactly the same way as $\hat{\lambda}$ except that the observations, $\boldsymbol{y}$, are distributed according to the known $\hat{F}$, rather than the unknown F. Thus, the distribution of $\hat{\lambda}$ is $G(x) = P[\lambda(F^*(\boldsymbol{y})) \leq x|F, D]$, whereas that of $\hat{\lambda}_1$ is the bootstrap distribution $G_B(x) = P[\lambda(F^*(\boldsymbol{y})) \leq x|\hat{F}, D]$ (e.g., Efron and Tibshirani 1993). In particular, the bias of $\hat{\lambda}$ can be estimated by $E\{\hat{\lambda}_1\} - \hat{\lambda}$ and the variance of $\hat{\lambda}$ by $V\{\hat{\lambda}_1\}$.

The bootstrap distribution is usually estimated by taking a random set of samples $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_k$, from $\hat{F}$, using design D (in our case, the sequential scheme): $G_{Bk}(x)$ is the fraction of samples for which $\lambda(F^*(\boldsymbol{y}_i)) \leq x$. But when $\hat{F}$ gives positive probability to only a small number of values, and $\boldsymbol{y}$ has only a few components, one can enumerate all possible samples from $\hat{F}$, and their probabilities, and compute G_B exactly. This would often be the case with the sequential sampling scheme because the maximum number of nonzero values in any sample is n_1; frequently it will be less. Enumeration removes one source of error from bootstrap estimates of variance and bias, the difference between G_B and G_{Bk}, though the difference between $G_B(x)$ and $G(x)$ remains. The calculations described below are for enumeration, although they could be used for random sampling.

In our case, the preliminary estimate, "$\lambda(\hat{F})$," is $\log(\hat{\alpha}_b) = \log(\hat{p}_b) + \log(C_+)$. Let S_+ be the observed set of positive values, that is, the N_+ positive C_i's, so C_+ = average of the values in S_+. For instance, if the original sample had three nonempty packets, with values 1, 2, and 2, then $S_+ = \{1, 2, 2\}$. Then, $\hat{F}$ is given by

$$P\{C = 0\} = 1 - \hat{p}_b \quad \text{and} \quad P\{C = j \mid C > 0\} = (\text{number of } j\text{'s in } S_+)/N_+. \tag{3.3}$$

When $N_+ = 0$, $P\{C = 1 \mid C > 0\} = 1$.

An estimate of the bias of $\log(\hat{\alpha}_b)$ is

$$B_1 = B_{p1} + B_{+1}, \tag{3.4}$$

where

$$B_{p1} = \log(\hat{p}_b) - E\{\log(\hat{p}_{b1}) \mid \hat{p}_b\}, \tag{3.5}$$

the bias estimate for $\log(\hat{p}_b)$, and

$$\begin{aligned} B_{+1} &= \log(C_+) - E\{\log(C_{+1}) \mid \hat{p}_b, S_+\} \\ &= \log(C_+) - \sum_k E\{\log(C_{+1}) \mid S_+,\ N_{+1} = k\} P\{N_{+1} = k \mid \hat{p}_b\}, \end{aligned} \tag{3.6}$$

the bias estimate for $\log(C_+)$. Here, $\hat{p}_{b1}$, C_{+1} and N_{+1} are given by equations (2.3)–(2.7), applied to a random sample drawn as in equations (2.2) and (2.3) from a population distributed as in (3.3). Under the sequential sampling scheme, both $\hat{p}_b$ and $\hat{p}_{b1}$ have the same set of possible values $(1/2n_2, 1/n_2, \ldots, (m-1)/n_2, (m-1)/(n_2-1), \ldots, (m-1)/n_1, m/n_1, \ldots, 1)$, so $P\{N_{+1} = k \mid \hat{p}_b\}$, all $P\{\hat{p}_{b1} \mid \hat{p}_b\}$, and $E\{\log(\hat{p}_{b1}) \mid \hat{p}_b\}$ are easy to compute. $E\{\log(C_{+1}) \mid S_+,\ N_{+1} = k\}$ is the mean of log(sample average) over all samples of size k, with replacement, from the set S_+. Computation is manageable unless k and the number of distinct values in S_+ are both large.

Thus, our bias-corrected bootstrap estimate of $\log(\alpha)$ was

$$\hat{L}_B = \log(\hat{\alpha}_b) + B_{p1} + B_{1+}. \tag{3.7}$$

In estimating the variance, I considered only the variance of $\log(\hat{\alpha}_b)$, assuming that the bias corrections would vary relatively little. (The task of simulating the bootstrap of a bootstrap for Section 5 was also daunting.) We have

$$V\{\log(\hat{\alpha}_b)\} = V\{\log(\hat{p}_b)\} + V\{\log(C_+)\} + 2\operatorname{cov}\{\log(\hat{p}_b),\ \log(C_+)\}. \tag{3.8}$$

With $\hat{\alpha}_{b1}$, $\hat{p}_{b1}$, C_{+1} and N_{+1} given by equations (2.2)–(2.8), but with the distribution of the C_i's given by that of C in (3.3), the bootstrap variance estimate is

$$\hat{V}_B = V\{\log(\hat{\alpha}_{b1})\} = V\{\log(\hat{p}_{b1})\} + V\{\log(C_{+1})\} + 2\operatorname{cov}\{\log(\hat{p}_{b1}),\ \log(C_{+1})\}. \tag{3.9}$$

$V\{\log(\hat{p}_{b1})\} = E\{\log(\hat{p}_{b1})^2\} - E^2\{\log(\hat{p}_{b1})\}$ again uses the common set of possible values of $\hat{p}_b$ and $\hat{p}_{b1}$ and the set of probabilities $P\{\hat{p}_{b1} \mid \hat{p}_b\}$.

For the second term of (3.9), we have

$$V\{\log(C_{+1})\} = \sum E\{\log(C_{+1})^2 \mid N_{+1} = k\} P\{N_{+1} = k\} - E^2\{\log(C_{+1})\}, \tag{3.10}$$

Both terms are computed like the second term in (3.6), discussed earlier.

For the third term in (3.9):

$$\operatorname{cov}\{\log(\hat{p}_{b1}),\ \log(C_{+1})\} = \sum E\{\log(\hat{p}_{b1}) \log(C_{+1}) \mid N_{+1} = k\} P\{k\} - E\{\log(\hat{p}_{b1})\} E\{\log(C_{+1})\}.$$

But, given N_{+1}, C_{+1} and $\hat{p}_{b1}$ are independent, because the distribution of C_{+1} depends only on the sample of N_{+1} drawn from S_+ while that of $\hat{p}_{b1}$ depends only on N_1, the number of packets sampled (cf. equations (2.3) and (2.4)). Thus,

$$E\{\log(\hat{p}_{b1})\log(C_{+1}) \mid N_{+1}=k\} = E\{\log(\hat{p}_{b1}) \mid N_{+1}=k\}E\{\log(C_{+1}) \mid N_{+1}=k\}. \qquad (3.11)$$

The first term is obtained from the set of possible values for $\hat{p}_b$: In fact, "$N_{+1} = k$" uniquely determines $\log(\hat{p}_{b1})$ unless $k = m$. The second term was discussed following (3.6).

4. Example

Table 1 shows counts of the California red scale, *Aonidiella aurantii*, on a grapefruit tree (W. W. Murdoch, unpublished data). Samples A and B were taken in October 1986, C and D in January 1987. Each was a fixed sample of 10 "packets." Each packet used the section consisting of the most recent four flushes (growth spurts) of a randomly chosen twig. A "packet" was the section's stem for A and C and its leaves for B and D.

Table 1

Counts of red scale on leaves and twigs

	1	2	3	4	5	6	7	8	9	10
A	7	1	25	41	6	17	3	0	0	7
B	0	0	0	6	0	1	0	0	0	3
C	0	0	1	4	0	0	0	0	1	0
D	1	0	0	0	0	0	0	0	0	0

	$\hat{\alpha}_b$	s^2	$\log(\hat{\alpha}_b)$	$\hat{L}_{TS}$	$\hat{L}_B$	$\hat{V}_{TS}$	$\hat{V}_B$
	Estimates for samples A, B and C, if $n_1 = 10$, $n_2 = 100$, $m = 3$						
A	10.701	7.4356	2.3702	2.4464	2.4478	0.1523	0.1709
B	1.00	0.3524	−0.0000	0.1762	0.1872	0.3524	0.4189
C	0.60	0.1417	−0.5108	−0.3140	−0.3090	0.3937	0.4446
	Estimates for $n_1 = 6$, $n_2 = 100$, $m = 2$						
A	15.503	9.9833	2.7408	2.8241	2.8219	0.1664	0.1844
B	1.17	0.8533	0.1542	0.4676	0.5037	0.6269	0.8595
C	0.83	0.3727	−0.1823	0.0860	0.1103	0.5366	0.6910
	Estimates for $n_1 = 6$, $n_2 = 100$, $m = 3$						
A	15.503	9.9833	2.7408	2.8241	2.8219	0.1664	0.1844
B	0.74	0.3195	−0.3001	−0.0089	−0.0556	0.5823	0.4941
C	0.50	0.1498	−0.6931	−0.3936	−0.4411	0.5992	0.5083

The sequential estimates are illustrated under three setups. Each has a maximum sample of $n_2 = 100$. They differ in the initial sample, n_1, and the required number of positives, m. Sample D is incomplete for all setups and would lead to further sampling. It was problems of this sort that led to the present work on sequential sampling, which has not been implemented yet. For the first and second setups, none of the samples would have gone into the sequential phase; the estimates are similar to, though not identical to, the fixed sample size estimates. In the third case, samples B and C both need the sequential phase to obtain the required number of positive packets. The observations for sample B are the same as for the first setup, but the estimates are different: In particular, $P\{\text{positive}\}$ is estimated by $(3-1)/(10-1)$, not by $3/10$.

5. Numerical Results

For all simulations summarized here, the C_i's were independent observations from a negative binomial distribution, classified by its mean, α, and by its value of $c = (\text{variance} - \alpha)/\alpha^2$. If C_i is seen as the number of failures before the kth success, when $P\{\text{success}\} = p$, then $\alpha = k(1-p)/p$. With $Q = 1/p$ and $P = Q - 1$, the probabilities are the terms of the expansion of $(Q-P)^{-k}$, and $\alpha = kP$. In both cases, $c = 1/k$; this seems preferable to k as a measure of "clumping" because the Poisson distribution has $c = 0$, and increases in c imply increases in clumping. The probability of an empty packet is $p^k = (c\alpha + 1)^{-1/c}$ (or $e^{-\alpha}$ for the Poisson).

The simulations used 10,000 replicates on each combination of $n_1 = 10$, $n_2 = 100$, $m = 2$ or 4, $\alpha = 0.02$, 0.1, 1, 10, or 100, and $c = 0$, 3 or 9. In practice, we would expect values of c between 0 and 3 and $\alpha \geq 0.1$. Values outside this range were intended as severe tests—for example, when $\alpha = 0.02$, many samples of 100 will have ≤ 1 nonempty packets.

Sequential sample results were compared to those for the "corresponding" fixed samples, that is, those whose size is the average size of the sequential sample (rounded up if the fractional part is > 0.1). This average size depends not only on the sequential sampler's choice of m but also on the population parameters α and c. These would not be known in practice, so a comparison of a sequential summary and the corresponding fixed sample loads the dice in favor of the fixed sample.

Abundance. Estimation of α was essentially unbiased: The maximum bias in $\hat{\alpha}_b$ was 4.5% for $\alpha = 0.02$, and less than 1% otherwise. Fixed sample biases were slightly larger, but also negligible.

Precision of abundance estimates. Variances of abundance estimates were smaller for fixed than for sequential samples, but most differences were small. The ratio (sequential variance)/(fixed variance) was < 1.1 for $\alpha \geq 1$, except for 1.25 when $\alpha = 1$ and $c = 9$; it was about 2 for $\alpha = 0.1$ and for $\alpha = 0.02$ when $m = 4$; and reached 4 when $\alpha = .02$ and $m = 2$. The standard deviation (SD) of the sequential estimate is about α for $\alpha = 0.02$ or 0.1 or for $c = 9$, about 0.6α for $c = 3$ and $\alpha > 0.1$, and about $0.3\sqrt{\alpha}$ for $c = 0$. These values are relevant to the Taylor series expansions, which assume small values of SD/abundance. The variance estimate, using (2.15), was negligibly biased for all combinations.

Log(abundance). Table 2 gives the biases of the estimates of $\log(\alpha)$, that is, the average of [estimate$-$ $\log(\alpha)$]. The negative bias of $\log(\hat{\alpha}_b)$ is roughly an increasing function of $V(\hat{\alpha}_b)/\alpha^2$ and is non-negligible except for $c = 0$ and $\alpha \geq 1$. Both adjustments have smaller biases than $\log(\hat{\alpha}_b)$, except for the Taylor series when abundance = 0.02, when it overcorrects. The bootstrap usually has a smaller bias than the Taylor series estimate, but this improvement is rather small except when the abundance is 0.02 or (with $m = 2$) 0.1. The biases for the fixed sample estimates are essentially the same, except when $m = 2$ and $\alpha = 0.02$ or 0.1, when they are about half the biases of the sequential estimates.

Table 2

Biases of log(abundance) estimates for $n_1 = 10$, $n_2 = 100$

					$\log(\alpha)$		
m	c		-3.91	-2.30	0.00	2.30	4.61
2	0	$\log(\hat{\alpha}_b)$	-0.33	-0.36	-0.05	-0.00	-0.00
		$\hat{L}_{TS}$	0.92	0.20	0.00	0.00	0.00
		$\hat{L}_B$	-0.19	-0.03	0.01	-0.00	0.00
	3	$\log(\hat{\alpha}_b)$	-0.33	-0.38	-0.22	-0.16	-0.16
		$\hat{L}_{TS}$	0.93	0.23	-0.05	-0.04	-0.04
		$\hat{L}_B$	-0.18	-0.04	-0.03	-0.02	-0.03
	9	$\log(\hat{\alpha}_b)$	-0.36	-0.44	-0.46	-0.48	-0.50
		$\hat{L}_{TS}$	0.94	0.27	-0.13	-0.24	-0.27
		$\hat{L}_B$	-0.22	-0.07	-0.14	-0.17	-0.15
4	0	$\log(\hat{\alpha}_b)$	-0.19	-0.15	-0.06	-0.00	-0.00
		$\hat{L}_{TS}$	0.24	0.05	-0.01	0.00	0.00
		$\hat{L}_B$	-0.11	-0.00	-0.00	0.00	0.00
	3	$\log(\hat{\alpha}_b)$	-0.20	-0.16	-0.18	-0.16	-0.16
		$\hat{L}_{TS}$	0.25	0.05	-0.04	-0.04	-0.04
		$\hat{L}_B$	-0.12	0.00	-0.05	-0.03	-0.03
	9	$\log(\hat{\alpha}_b)$	-0.22	-0.19	-0.28	-0.40	-0.46
		$\hat{L}_{TS}$	0.25	0.05	-0.05	-0.17	-0.23
		$\hat{L}_B$	-0.14	-0.01	-0.08	-0.14	-0.15

Precision of log(abundance) estimates. Table 3 gives the SD of the estimates of $\log(\alpha)$. One of the aims of the sequential samples—that the precision of the estimate of $\log(\alpha)$ should be only weakly dependent on α itself—is seen to be achieved, although there is a trend (smaller variances for larger α) for the Poisson ($c = 0$) distribution. Except for the Taylor series estimates when $m = 2$ and $\alpha = 0.02$, the SDs are always greater than the biases, often by factors of 5 for $\log(\hat{\alpha}_b)$ and of 10 for the other estimates. The SDs are rather similar for the three estimates in most cases, but where they do differ ($\alpha = 0.02$, and $\alpha = 0.1$ and 1 for $m = 2$), the Taylor series estimate's SD tends to be smallest. (The Taylor series SD's are smaller for $m = 2$ than for $m = 4$, when $\alpha = 0.02$; this is due to sampling error: the minimum of m positive packets was frequently not reached in both cases.) Results for fixed sample estimates are very similar: The extreme values of (fixed SD)—(sequential SD) for the Taylor series estimates were 0.1 and −0.07; these occurred for $\alpha = 1$, $c = 9$, $m = 4$, and $\alpha = 10$, $c = 9$, $m = 4$, respectively.

Table 3

SDs of log(abundance) estimates for $n_1 = 10$, $n_2 = 100$

			log(α)				
m	c		−3.91	−2.30	0.00	2.30	4.61
2	0	$\log(\hat{\alpha}_b)$	0.73	0.82	0.34	0.10	0.03
		$\hat{L}_{TS}$	0.36	0.56	0.32	0.10	0.03
		$\hat{L}_B$	0.91	0.76	0.32	0.10	0.03
	3	$\log(\hat{\alpha}_b)$	0.74	0.85	0.72	0.61	0.59
		$\hat{L}_{TS}$	0.37	0.60	0.67	0.61	0.59
		$\hat{L}_B$	0.92	0.82	0.68	0.61	0.60
	9	$\log(\hat{\alpha}_b)$	0.74	0.90	1.03	1.12	1.18
		$\hat{L}_{TS}$	0.39	0.66	0.95	1.11	1.18
		$\hat{L}_B$	0.95	0.89	1.01	1.15	1.22
4	0	$\log(\hat{\alpha}_b)$	0.69	0.52	0.34	0.10	0.03
		$\hat{L}_{TS}$	0.47	0.50	0.33	0.10	0.03
		$\hat{L}_B$	0.84	0.51	0.33	0.10	0.03
	3	$\log(\hat{\alpha}_b)$	0.71	0.56	0.60	0.61	0.59
		$\hat{L}_{TS}$	0.49	0.54	0.59	0.61	0.60
		$\hat{L}_B$	0.88	0.55	0.60	0.62	0.60
	9	$\log(\hat{\alpha}_b)$	0.72	0.60	0.76	0.96	1.09
		$\hat{L}_{TS}$	0.50	0.58	0.76	0.97	1.10
		$\hat{L}_B$	0.90	0.59	0.77	0.99	1.13

When Tables 2 and 3 are combined to give the root mean square error (RMSE), the bias differences are mostly swamped by variances. The rmse is not very different for the three estimators, but the difference tends to favor the Taylor series estimate. As for the SDs, the RMSE's are only weakly dependent on α except when $c = 0$. RMSE's are again very similar for the fixed sample estimates.

Estimated variances of log(abundance) estimates. Nominally, the variance estimates $\hat{V}_{TS}$ and $\hat{V}_B$ described in equations (3.2) and (3.9) are aimed at $V\{\log(\hat{\alpha}_b)\}$, but the performance of the Taylor Series estimate in Table 3 makes $V\{\hat{L}_{TS}\}$ of interest. These target values are the squares of the SD's in Table 3. $\hat{V}_{TS}$ overestimates both targets when α is small and underestimates when α is large; its bias is below 10% when $c = 0$ and $\alpha \geq 1$ (except for 17% for $m = 4$, $\alpha = 1$) and below 35% when $c = 3$, $\alpha \geq 1$ or $c = 9$, $\alpha = 1$, but it is otherwise over 50%, often far over. $\hat{V}_B$ underestimates $V\{\log(\hat{\alpha}_b)\}$ in all but three cases, but its bias is always below 30% and usually below 20%. It overestimates $V\{\hat{L}_{TS}\}$ for $\alpha < 1$ and underestimates for $\alpha \geq 1$ (except for $m = 2$, $c = 0$, $\alpha = 1$); the bias is less than 32% except when $\alpha = 0.02$ or $m = 2$, $\alpha = 0.1$.

The SDs of these variance estimates also tend to be moderately larger than the biases. However the Taylor series estimate shows no clear tendency to be smallest: $\hat{V}_{TS}$ has the largest SD's when

α is small, but the smallest when α is large. As a result, $\hat{V}_B$ seems to do best overall, when judged by RMSE, although the improvement is not great. Table 4 shows (RMSE)/$V\{\hat{L}_{TS}\}$. $\hat{V}_B$ is clearly better for $\alpha < 1$, whereas the difference is negligible for $\alpha \geq 1$.

Comments. These results suggest that the Taylor series adjustment to $\log(\hat{\alpha}_b)$ gives a useful improvement, whereas the extra effort demanded by the bootstrap may result in an inferior estimator. There may be exceptions to this rule. For example, biased estimation could either inflate or deflate an estimate of the variability of log(abundance) over time, depending on whether clumping increases or decreases as abundance increases, so bias could play a larger role in the comparison of populations whose clumping varies with abundance.

The Taylor series estimate can be in error either because $\sigma^2/2\alpha^2$ is a poor approximation to the bias, or because $s^2/2\hat{\alpha}_b^2$ is a poor estimate of it. In fact, $\log(\hat{\alpha}_b) + \sigma^2/2\alpha^2$ had low bias except when $\alpha = 0.02$, notably low when α and c were large (even though the Taylor series justification seems weak, with $\sigma/\alpha = 0.5$ or more). Because σ^2/α^2 is the Taylor Series approximation to $V\{\log(\hat{\alpha}_b)\}$, I looked at $\log(\hat{\alpha}_b) + \hat{V}_B/2$ and $\log(\hat{\alpha}_b) + V\{\log(\hat{\alpha}_b)\}/2$. These performed well (average squared biases <0.005), but this may be fortuitous, and the former's decrease in bias may be outweighed by the increase in variance, although I did not check this.

Table 4
RMSE/V$\{L_{TS}\}$ *for estimates of* $V\{L_{TS}\}$, *where* L_{TS} = *Taylor series estimate of log(abundance) for* n_1 = 10, n_2 = 100

			log(α)				
m	c		−3.91	−2.30	0.00	2.30	4.61
2	0	$\hat{V}_{TS}$	19.35	3.14	0.62	0.48	0.47
		$\hat{V}_B$	2.57	1.01	0.93	0.45	0.44
	3	$\hat{V}_{TS}$	18.20	2.90	0.56	0.50	0.50
		$\hat{V}_B$	2.46	0.82	0.52	0.67	0.70
	9	$\hat{V}_{TS}$	16.93	2.72	0.53	0.64	0.68
		$\hat{V}_B$	2.24	0.69	0.44	0.61	0.76
4	0	$\hat{V}_{TS}$	3.62	0.68	0.53	0.47	0.47
		$\hat{V}_B$	0.86	0.28	0.56	0.45	0.44
	3	$\hat{V}_{TS}$	3.42	0.58	0.41	0.50	0.50
		$\hat{V}_B$	0.79	0.30	0.39	0.58	0.68
	9	$\hat{V}_{TS}$	3.33	0.53	0.36	0.56	0.65
		$\hat{V}_B$	0.84	0.41	0.44	0.58	0.68

The bootstrap variance estimate, $\hat{V}_B = V\{\log(\hat{\alpha}_{b1})\}$ performed better overall than the Taylor series estimate, but only because it did much better for small α. It also performed better as an estimate of $V\{\log(\hat{\alpha}_b)\}$, its nominal target, than as an estimate of $V\{\hat{L}_{TS}\}$. This suggests that "$V\{\hat{L}_{TS1}\}$," the bootstrap estimate aimed at the Taylor series estimate of log(abundance), might perform better still. A more elaborate possibility arises if $\log(\hat{\alpha}_b) + \hat{V}_B/2$ should improve on $\hat{L}_{TS}$: A bootstrap estimate of its variance would require bootstrapping a bootstrap. The computation required is likely to be manageable, because the sequential samples are likely to have only a few distinct values, but simulating the performance of such an estimate may be difficult.

Bootstrapping has an additional advantage not studied here: It can be used to compute confidence intervals without any distributional assumptions, such as Normality. This may be less important than the point estimate for log(abundance) at a particular time, if our ultimate aim is to study the mean or variability over time for impact or theory assessment, but could be valuable in other cases.

Apparent dead ends. Several variations on the bootstrap methods were also tried.

B_{p1} (equation (3.5)) is itself likely to be biased. Its bias is $\log(p) - E\{\log(\hat{p}_b) \mid p\} - E\{\log(\hat{p}_b) - E\{\log(\hat{p}_{b1}) \mid \hat{p}_b\} \mid p\}$, which can also be estimated by the bootstrap, by replacing p, $\hat{p}_b$ and $\hat{p}_{b1}$ with $\hat{p}_b$, $\hat{p}_{b1}$, and $\hat{p}_{b2}$ respectively, where $\hat{p}_{b2}$ is obtained by first sampling as in Section 2 from the distribution given by $\hat{p}_b$ and using (2.6) to obtain $\hat{p}_{b1}$, then sampling from the distribution given by $\hat{p}_{b1}$ to obtain $\hat{p}_{b2}$. This estimate, B_{p2}, has a bias that can be estimated by B_{p3}, and so on, giving a series of estimates of $\log(p)$. The sequence $B_{p1}, B_{p2}, \ldots$, is easy to program. Given m, n_1, and n_2, the only possible values of $\hat{p}_b$ (or $\hat{p}_{b1}, \ldots$) are $1/2n_2, 1/n_2, \ldots, (m-1)/n_2, (m-1)/(n_2-1), \ldots, (m-1)/n_1, m/n_1, \ldots, 1$. $E\{\log(\hat{p}_{b1}) \mid \hat{p}_b\}$, and $P\{\hat{p}_{b1} \mid \hat{p}_b\}$, $P\{\hat{p}_{b2} \mid \hat{p}_b\} = \sum P\{\hat{p}_{b2} \mid \hat{p}_{b1}\} P\{\hat{p}_{b1} \mid \hat{p}_b\}$,and so on, are easy to record. I looked at the first two iterations of this sequence. A similar iteration is possible for $V\{\log(\hat{p}_{b1})\}$: I looked at the first step of this sequence, by adding $V\{\log(\hat{p}_{b1})\} - E\{V\{\log(\hat{p}_{b2}) \mid \hat{p}_{b1}\}\}$. None of these more elaborate estimates performed better than the simpler ones. A similar iteration is possible for B_{1+} but requires far more computing, so I did not attempt it.

Two other estimates of variance were also considered. One was given by (3.9) but with "$V\{\log(C_{+1})\}$" multiplied by $N_+/(N_+ - 1)$ when $N_+ \geq 2$, by analogy with the standard unbiased estimate of variance where the same multiplier is used on the variance of the empirical distribution function. The other applies the same analogy to the covariance term, using (3.8) but with "$V\{\log(C_{+1})\} + 2\,\mathrm{cov}\{\log(\hat{p}_{b1}),\ \log(C_{+1})\}$" multiplied by $N_+/(N_+ - 1)$ for $N_+ \geq 2$. Both estimates had slightly smaller biases overall than $\hat{V}_B$, but their SD's and RMSEs were larger.

6. Discussion

The sequential sampling scheme described in Section 2 is easy to carry out in many cases. Its scope is wider than may at first appear. In many practical cases, such as small animals hidden in core samples from the sea bottom, the sampler cannot distinguish between "zero" and "positive" packets until the samples are analyzed in the laboratory. In such cases, however, the main cost of sampling is often in the laboratory analysis, not in the collection. Thus, one can sample the full set of n_2 packets in the field, but analyze only the number needed to give the required number of positive packets.

The estimates are easy to calculate, except for s^2 in (2.15), which requires negligible computing, or can be approximated, and the bootstrap estimates, which require intensive computation although the programming is simple. The bootstrap estimates could be avoided; the bootstrap bias adjustment seems inferior to the Taylor series adjustment, based on RMSE, and the bootstrap variance estimate may not be a sufficient improvement to justify the effort. On the other hand, more elaborate bootstrap estimates using smoothing (e.g., Young 1994) might do better.

The errors are unacceptably large in some cases, but most of these involve extreme, unrepresentative means or clumping. Given an abundance of 0.02 per packet, most biologists will use a larger packet, a larger n_2, or a different sampling method—or will study a different organism! Clumping of $c = 9$, implying a coefficient of variation of $> 300\%$, seems very rare. The ranges $\alpha \geq 0.1$ and $c = 0$ to $c = 3$, seem much more common, and most results are adequate, although less than perfect, for this.

The assumption that the sampled "packets" are roughly equal in size is in fact not needed. For example, the packets could be plants or branches or twigs of varying sizes. Estimation of the mean abundance (the average number of animals per plant, per branch, or per twig) and of its log would proceed as earlier. If the sizes of the plants, branches or twigs are recorded, the abundance estimate can be converted into an estimated abundance per unit of area in the usual way. The only requirement is that the sampled plants (say) be randomly sampled: In particular, there should be no tendency to choose the larger ones. The drawback is that, although (say) number of animals and surface area can be measured on each plant, only the number can be used in the abundance estimate: For example, one could not adjust the estimate if the area measurements indicated that, despite the randomization used, the sampled plants were in fact larger than average. Such an adjustment would improve the method, but it requires a model specifying the mean and variance of the numbers of animals on a plant as functions of its size.

Much of the concern here has been with the development of unbiased estimates. There are objections to the use of unbiasedness as a guiding principle. It can lead to unappealing (and inadmissible) estimates—for example, negative estimates of variance components. It involves an average over the sample space, so it depends on what was not observed (but could have been), as well as on what was observed. In sequential sampling, the unobserved possibilities depend on the sampler's intentions—for example, our formulae give different results for the same sample, depending on the values of m, n_1, and n_2. Thus, our estimate depends not only on what nature tells us, but also on what was in the sampler's mind, which seems irrelevant.

However, unbiasedness makes better sense when a sampling plan is likely to be repeated many times, by one or more investigators (Anscombe 1963). This seems likely for impact assessment schemes, measures of temporal variation, and other uses of $\log(\alpha)$.

Second, for small m (which seems the best choice), the initial sample of n_1 will usually yield the required m positive twigs, unless positive twigs are rare (in which case all the estimates are biased). In these cases, $\hat{\alpha}_b$ is just $C.$, the average number per packet, and the variance estimator, s^2, is fairly well approximated by (sample variance)$/n_1$: the difference seems to be $\leq 5\%$, usually. Neither $C.$ nor the sample variance depends on m or n_2, and they depend on n_1 only because this was the actual sample size, not because of the sampler's intentions.

Third, the Taylor series estimate seems to reduce both bias and variance (see Table 3), so other criteria (e.g., Bayesian expected squared error) may also favor this estimator.

A final comment is that methods similar to those used here could be used for estimating other functions of abundance defined only for positive values—for example, habitat quality may sometimes be best indicated by the reciprocal of abundance, that is, the amount of habitat needed to support one individual.

Acknowledgements

I thank Craig Osenberg, Milton Sobel, Girish Aras, and especially Bill Murdoch for commenting on several drafts and Bill Murdoch and Sue Swarbrick for the data. Very detailed comments by an Associate Editor and a referee much improved the focus. This work was supported in part by the Minerals Management Service, U.S. Department of the Interior, under MMS agreement 14-35-0001-30471 (The Southern California Educational Initiative). The views and interpretation contained in this document are those of the author and should not be interpreted as necessarily representing the official policies, either express or implied, of the U.S. Government.

Résumé

Je traite de l'estimation de l'abondance d'une population biologique, les logarithmes et les variances de ces estimations, à partir d'un schéma d'échantillonnage séquentiel avec des tailles minimum et maximum. Les observations sont les dénombrements d'organismes dans des "paquets" tirés au sort tels que des carottes, des branches, des buissons, ... Des échantillons sont constitués pour des valeurs pré-définies de m, n_1 et n_2 jusqu'à (a) au moins n_1 paquets et (b) soit m paquets positifs ou n_2 paquets observés.

Les estimations d'abondance sont basées sur une estimation de la fraction de paquets positifs, donnée par Kremers (1987, *Technometrics* **29**, 102–112), modifiée afin d'éviter les estimations de zéro. Les estimations du log de l'abondance sont données par Log (abondance estimée) avec une correction du biais dû à la concavité de la fonction Log. Deux ajustements sont considérés, l'un basé sur un dévéloppement en série de Taylor (la 'delta method') et l'autre sur le bootstrap. Ces techniques sont aussi utilisées pour estimer la variance de l'estimateur de la Log (Abondance). Des simulations suggèrent que les deux mèthodes sont préférables à ne pas ajuster, bien que le gain soit faible comparé à l'estimation de l'écart-type standard. Les estimations bootstrap sont moins biaisées que les estimations des séries de Taylor, mais elles sont de plus grandes variances, de sorte que les estimations de séries de Taylor ont des carrés moyens plus petits. Les variances des estimations séquentielles de la Log (Abondance) ont tendance a être seulement légèrement dépendantes de l'abondance vraie.

References

Anscombe, F. J. (1948). The transformation of Poisson, binomial and negative binomial data. *Biometrika* **35,** 246–254.

Anscombe, F. J. (1963). Sequential medical trials. *Journal of the American Statistical Association* **58,** 365–383.

Atkinson, A. C. (1985). *Plots, Transformations and Regression.* Oxford, Clarendon.

Berry, D. A. (1987). Logarithmic transformations in ANOVA. *Biometrics* **43,** 439–456.

Box, G. E. P., and Cox. D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, 'B'* **26,** 211-252.

DeGroot, M. H. (1959). Unbiased sequential estimation for binomial populations. *Annals of Mathematical Statistics* **30,** 80–101.

Efron, B. and R. J. Tibshirani. (1993). *An Introduction to the Bootstrap.* New York, Chapman and Hall.

Feller, W. (1966). *An Introduction to Probability Theory and Its Applications.* Volume I. Wiley, New York.

Kim, S. and Nachlas. J. A. (1984). Estimation in Bernoulli trials under a generalized sampling plan. *Technometrics* **26,** 379–387.

Kremers, W. K. (1987). An improved estimator of the mean for a sequential binomial sampling plan. *Technometrics* **29,** 109–112.

Mosteller, F. and Tukey. J. W. (1977). *Data Analysis and Regression.* Reading, Massachusetts: Addison Wesley.

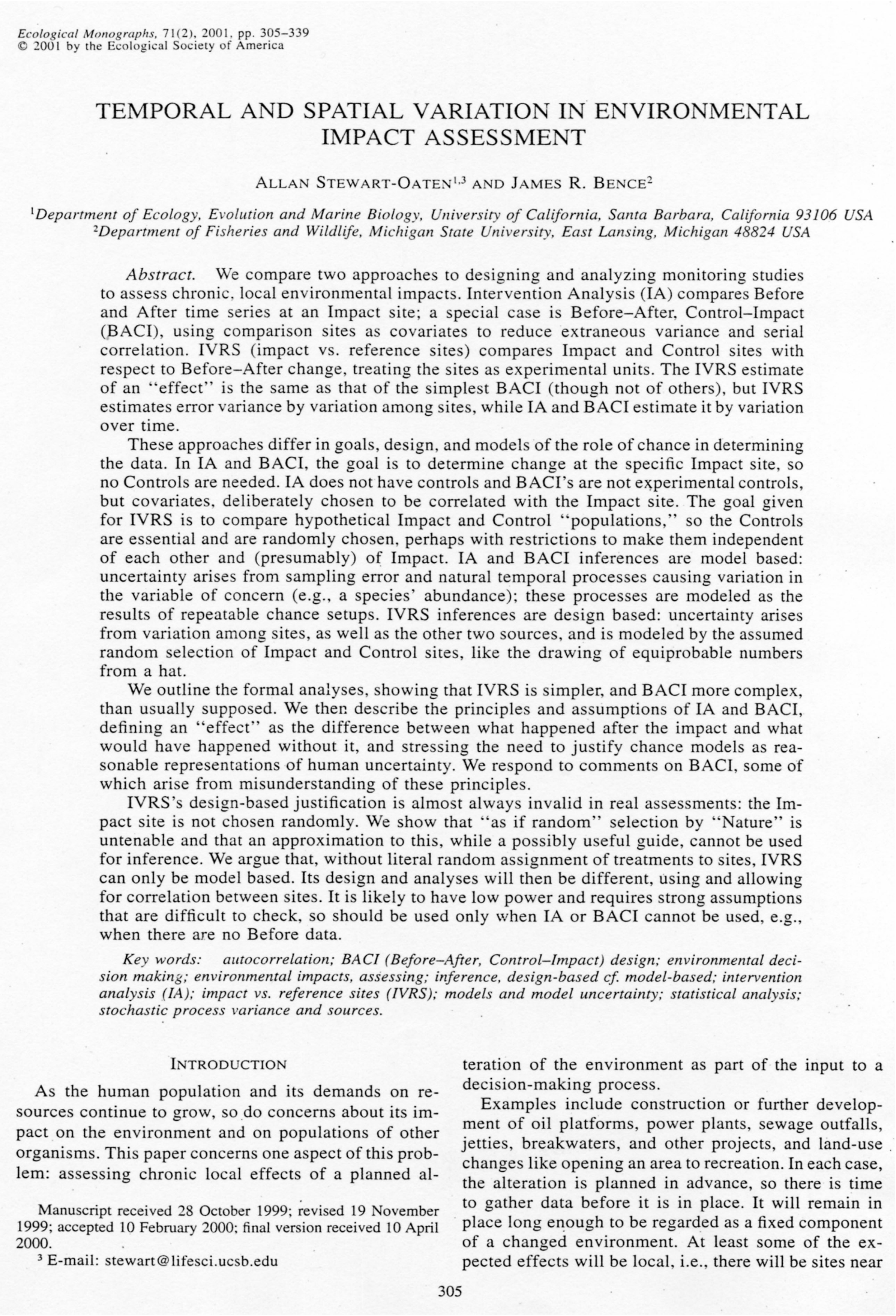
Stewart-Oaten, A., Murdoch, W. W. and Parker. K. R. (1986). Environmental impact assessment: "Pseudoreplication" in time? *Ecology* **67,** 929–940.

Taylor, L. R. (1961). Aggregation, variance and the mean. *Nature* **189,** 732–735.

Williamson, M. H. (1984). The measurement of population variability. *Ecological Entomology* **9,** 239–241.

Young, G. A. (1994). Bootstrap: more than a stab in the dark? *Statistical Science* **9,** 382–415.

Received August 1993; *revised July* 1994; *accepted April* 1995.



Ecological Monographs, 71(2), 2001, pp. 305–339

TEMPORAL AND SPATIAL VARIATION IN ENVIRONMENTAL IMPACT ASSESSMENT

ALLAN STEWART-OATEN[1,3] AND JAMES R. BENCE[2]

[1]*Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara, California 93106 USA*
[2]*Department of Fisheries and Wildlife, Michigan State University, East Lansing, Michigan 48824 USA*

Abstract. We compare two approaches to designing and analyzing monitoring studies to assess chronic, local environmental impacts. Intervention Analysis (IA) compares Before and After time series at an Impact site; a special case is Before–After, Control–Impact (BACI), using comparison sites as covariates to reduce extraneous variance and serial correlation. IVRS (impact vs. reference sites) compares Impact and Control sites with respect to Before–After change, treating the sites as experimental units. The IVRS estimate of an "effect" is the same as that of the simplest BACI (though not of others), but IVRS estimates error variance by variation among sites, while IA and BACI estimate it by variation over time.

These approaches differ in goals, design, and models of the role of chance in determining the data. In IA and BACI, the goal is to determine change at the specific Impact site, so no Controls are needed. IA does not have controls and BACI's are not experimental controls, but covariates, deliberately chosen to be correlated with the Impact site. The goal given for IVRS is to compare hypothetical Impact and Control "populations," so the Controls are essential and are randomly chosen, perhaps with restrictions to make them independent of each other and (presumably) of Impact. IA and BACI inferences are model based: uncertainty arises from sampling error and natural temporal processes causing variation in the variable of concern (e.g., a species' abundance); these processes are modeled as the results of repeatable chance setups. IVRS inferences are design based: uncertainty arises from variation among sites, as well as the other two sources, and is modeled by the assumed random selection of Impact and Control sites, like the drawing of equiprobable numbers from a hat.

We outline the formal analyses, showing that IVRS is simpler, and BACI more complex, than usually supposed. We then describe the principles and assumptions of IA and BACI, defining an "effect" as the difference between what happened after the impact and what would have happened without it, and stressing the need to justify chance models as reasonable representations of human uncertainty. We respond to comments on BACI, some of which arise from misunderstanding of these principles.

IVRS's design-based justification is almost always invalid in real assessments: the Impact site is not chosen randomly. We show that "as if random" selection by "Nature" is untenable and that an approximation to this, while a possibly useful guide, cannot be used for inference. We argue that, without literal random assignment of treatments to sites, IVRS can only be model based. Its design and analyses will then be different, using and allowing for correlation between sites. It is likely to have low power and requires strong assumptions that are difficult to check, so should be used only when IA or BACI cannot be used, e.g., when there are no Before data.

Key words: autocorrelation; BACI (Before–After, Control–Impact) design; environmental decision making; environmental impacts, assessing; inference, design-based cf. model-based; intervention analysis (IA); impact vs. reference sites (IVRS); models and model uncertainty; statistical analysis; stochastic process variance and sources.

INTRODUCTION

As the human population and its demands on resources continue to grow, so do concerns about its impact on the environment and on populations of other organisms. This paper concerns one aspect of this problem: assessing chronic local effects of a planned alteration of the environment as part of the input to a decision-making process.

Examples include construction or further development of oil platforms, power plants, sewage outfalls, jetties, breakwaters, and other projects, and land-use changes like opening an area to recreation. In each case, the alteration is planned in advance, so there is time to gather data before it is in place. It will remain in place long enough to be regarded as a fixed component of a changed environment. At least some of the expected effects will be local, i.e., there will be sites near

Manuscript received 28 October 1999; revised 19 November 1999; accepted 10 February 2000; final version received 10 April 2000.

[3] E-mail: stewart@lifesci.ucsb.edu

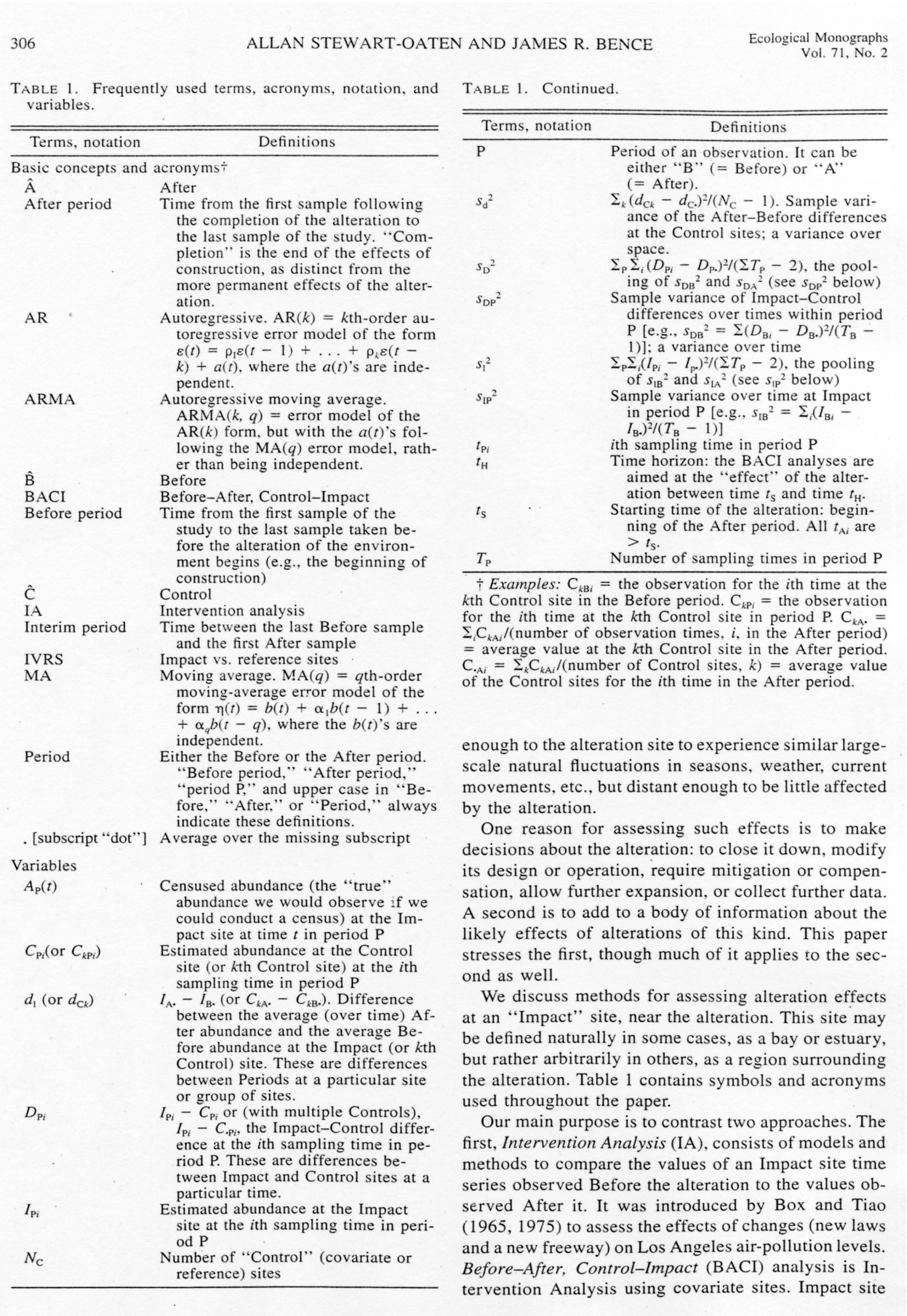


TABLE 1. Frequently used terms, acronyms, notation, and variables.

Terms, notation	Definitions
Basic concepts and acronyms†	
Â	After
After period	Time from the first sample following the completion of the alteration to the last sample of the study. "Completion" is the end of the effects of construction, as distinct from the more permanent effects of the alteration.
AR	Autoregressive. AR(k) = kth-order autoregressive error model of the form $\varepsilon(t) = \rho_1\varepsilon(t-1) + \ldots + \rho_k\varepsilon(t-k) + a(t)$, where the $a(t)$'s are independent.
ARMA	Autoregressive moving average. ARMA(k, q) = error model of the AR(k) form, but with the $a(t)$'s following the MA(q) error model, rather than being independent.
B̂	Before
BACI	Before–After, Control–Impact
Before period	Time from the first sample of the study to the last sample taken before the alteration of the environment begins (e.g., the beginning of construction)
Ĉ	Control
IA	Intervention analysis
Interim period	Time between the last Before sample and the first After sample
IVRS	Impact vs. reference sites
MA	Moving average. MA(q) = qth-order moving-average error model of the form $\eta(t) = b(t) + \alpha_1 b(t-1) + \ldots + \alpha_q b(t-q)$, where the $b(t)$'s are independent.
Period	Either the Before or the After period. "Before period," "After period," "period P," and upper case in "Before," "After," or "Period," always indicate these definitions.
. [subscript "dot"]	Average over the missing subscript
Variables	
$A_P(t)$	Censused abundance (the "true" abundance we would observe if we could conduct a census) at the Impact site at time t in period P
C_{Pi}(or C_{kPi})	Estimated abundance at the Control site (or kth Control site) at the ith sampling time in period P
d_I (or d_{Ck})	$I_{A\cdot} - I_{B\cdot}$ (or $C_{kA\cdot} - C_{kB\cdot}$). Difference between the average (over time) After abundance and the average Before abundance at the Impact (or kth Control) site. These are differences between Periods at a particular site or group of sites.
D_{Pi}	$I_{Pi} - C_{Pi}$ or (with multiple Controls), $I_{Pi} - C_{\cdot Pi}$, the Impact–Control difference at the ith sampling time in period P. These are differences between Impact and Control sites at a particular time.
I_{Pi}	Estimated abundance at the Impact site at the ith sampling time in period P
N_C	Number of "Control" (covariate or reference) sites
P	Period of an observation. It can be either "B" (= Before) or "A" (= After).
s_d^2	$\Sigma_k (d_{Ck} - d_{C\cdot})^2/(N_C - 1)$. Sample variance of the After–Before differences at the Control sites; a variance over space.
s_D^2	$\Sigma_P \Sigma_i (D_{Pi} - D_{P\cdot})^2/(\Sigma T_P - 2)$, the pooling of s_{DB}^2 and s_{DA}^2 (see s_{DP}^2 below)
s_{DP}^2	Sample variance of Impact–Control differences over times within period P [e.g., $s_{DB}^2 = \Sigma(D_{Bi} - D_{B\cdot})^2/(T_B - 1)$]; a variance over time
s_I^2	$\Sigma_P \Sigma_i (I_{Pi} - I_{p\cdot})^2/(\Sigma T_P - 2)$, the pooling of s_{IB}^2 and s_{IA}^2 (see s_{IP}^2 below)
s_{IP}^2	Sample variance over time at Impact in period P [e.g., $s_{IB}^2 = \Sigma_i(I_{Bi} - I_{B\cdot})^2/(T_B - 1)$]
t_{Pi}	ith sampling time in period P
t_H	Time horizon: the BACI analyses are aimed at the "effect" of the alteration between time t_S and time t_H.
t_S	Starting time of the alteration: beginning of the After period. All t_{Ai} are $> t_S$.
T_P	Number of sampling times in period P

† *Examples:* C_{kBi} = the observation for the ith time at the kth Control site in the Before period. C_{kPi} = the observation for the ith time at the kth Control site in period P. $C_{kA\cdot} = \Sigma_i C_{kAi}$/(number of observation times, i, in the After period) = average value at the kth Control site in the After period. $C_{\cdot Ai} = \Sigma_k C_{kAi}$/(number of Control sites, k) = average value of the Control sites for the ith time in the After period.

enough to the alteration site to experience similar large-scale natural fluctuations in seasons, weather, current movements, etc., but distant enough to be little affected by the alteration.

One reason for assessing such effects is to make decisions about the alteration: to close it down, modify its design or operation, require mitigation or compensation, allow further expansion, or collect further data. A second is to add to a body of information about the likely effects of alterations of this kind. This paper stresses the first, though much of it applies to the second as well.

We discuss methods for assessing alteration effects at an "Impact" site, near the alteration. This site may be defined naturally in some cases, as a bay or estuary, but rather arbitrarily in others, as a region surrounding the alteration. Table 1 contains symbols and acronyms used throughout the paper.

Our main purpose is to contrast two approaches. The first, *Intervention Analysis* (IA), consists of models and methods to compare the values of an Impact site time series observed Before the alteration to the values observed After it. It was introduced by Box and Tiao (1965, 1975) to assess the effects of changes (new laws and a new freeway) on Los Angeles air-pollution levels. *Before–After, Control–Impact* (BACI) analysis is Intervention Analysis using covariate sites. Impact site

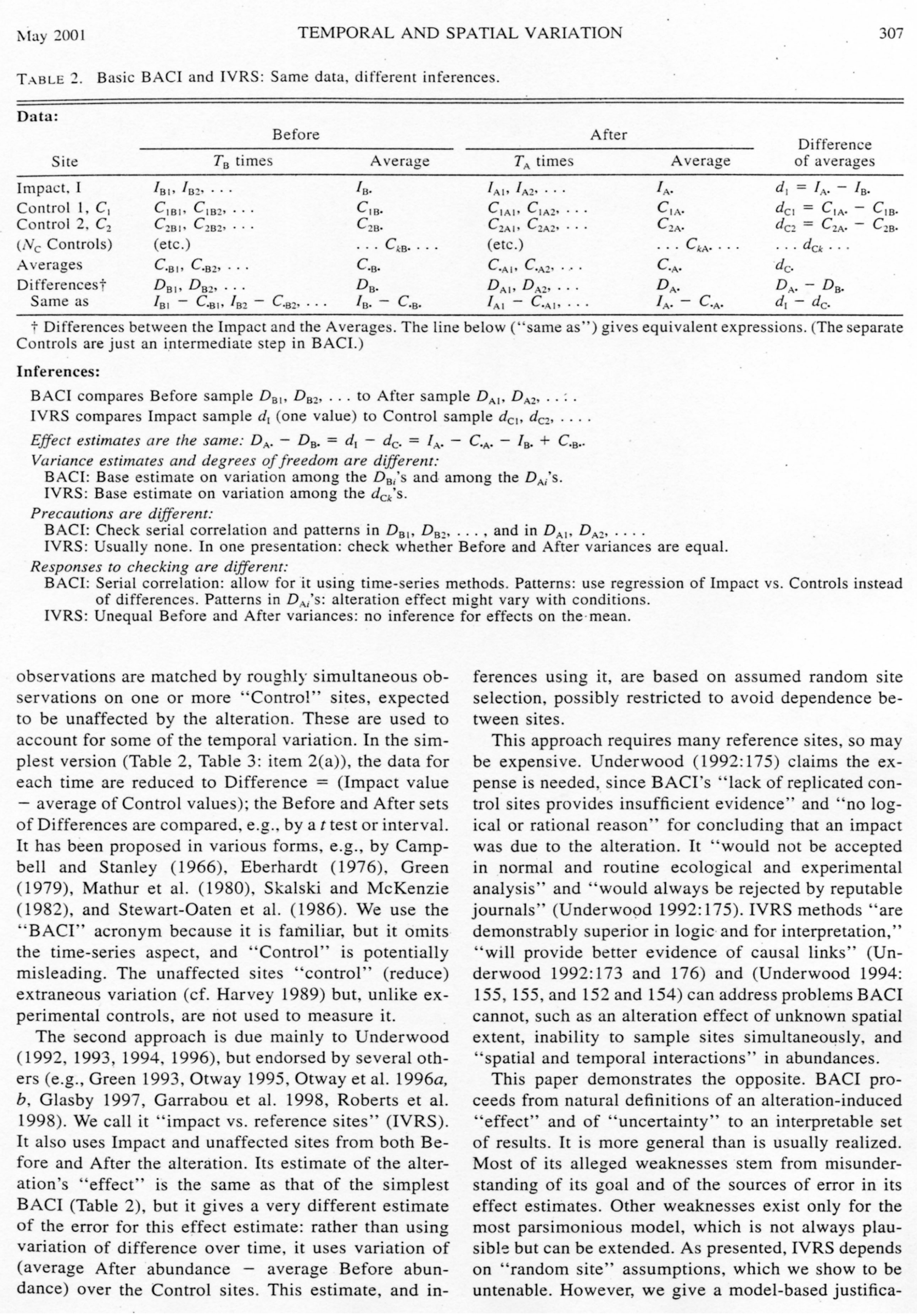


TABLE 2. Basic BACI and IVRS: Same data, different inferences.

Data:

Site	Before		After		Difference of averages
	T_B times	Average	T_A times	Average	
Impact, I	$I_{B1}, I_{B2}, \ldots$	$I_{B\cdot}$	$I_{A1}, I_{A2}, \ldots$	$I_{A\cdot}$	$d_I = I_{A\cdot} - I_{B\cdot}$
Control 1, C_1	$C_{1B1}, C_{1B2}, \ldots$	$C_{1B\cdot}$	$C_{1A1}, C_{1A2}, \ldots$	$C_{1A\cdot}$	$d_{C1} = C_{1A\cdot} - C_{1B\cdot}$
Control 2, C_2	$C_{2B1}, C_{2B2}, \ldots$	$C_{2B\cdot}$	$C_{2A1}, C_{2A2}, \ldots$	$C_{2A\cdot}$	$d_{C2} = C_{2A\cdot} - C_{2B\cdot}$
(N_C Controls)	(etc.)	$\ldots C_{kB\cdot} \ldots$	(etc.)	$\ldots C_{kA\cdot} \ldots$	$\ldots d_{Ck} \ldots$
Averages	$C_{\cdot B1}, C_{\cdot B2}, \ldots$	$C_{\cdot B\cdot}$	$C_{\cdot A1}, C_{\cdot A2}, \ldots$	$C_{\cdot A\cdot}$	$d_{C\cdot}$
Differences†	$D_{B1}, D_{B2}, \ldots$	$D_{B\cdot}$	$D_{A1}, D_{A2}, \ldots$	$D_{A\cdot}$	$D_{A\cdot} - D_{B\cdot}$
Same as	$I_{B1} - C_{\cdot B1}, I_{B2} - C_{\cdot B2}, \ldots$	$I_{B\cdot} - C_{\cdot B\cdot}$	$I_{A1} - C_{\cdot A1}, \ldots$	$I_{A\cdot} - C_{\cdot A\cdot}$	$d_I - d_{C\cdot}$

† Differences between the Impact and the Averages. The line below ("same as") gives equivalent expressions. (The separate Controls are just an intermediate step in BACI.)

Inferences:

BACI compares Before sample $D_{B1}, D_{B2}, \ldots$ to After sample $D_{A1}, D_{A2}, \ldots$.

IVRS compares Impact sample d_I (one value) to Control sample $d_{C1}, d_{C2}, \ldots$.

Effect estimates are the same: $D_{A\cdot} - D_{B\cdot} = d_I - d_{C\cdot} = I_{A\cdot} - C_{\cdot A\cdot} - I_{B\cdot} + C_{\cdot B\cdot}$.

Variance estimates and degrees of freedom are different:
BACI: Base estimate on variation among the D_{Bi}'s and among the D_{Ai}'s.
IVRS: Base estimate on variation among the d_{Ck}'s.

Precautions are different:
BACI: Check serial correlation and patterns in $D_{B1}, D_{B2}, \ldots$, and in $D_{A1}, D_{A2}, \ldots$.
IVRS: Usually none. In one presentation: check whether Before and After variances are equal.

Responses to checking are different:
BACI: Serial correlation: allow for it using time-series methods. Patterns: use regression of Impact vs. Controls instead of differences. Patterns in D_{Ai}'s: alteration effect might vary with conditions.
IVRS: Unequal Before and After variances: no inference for effects on the mean.

observations are matched by roughly simultaneous observations on one or more "Control" sites, expected to be unaffected by the alteration. These are used to account for some of the temporal variation. In the simplest version (Table 2, Table 3: item 2(a)), the data for each time are reduced to Difference = (Impact value − average of Control values); the Before and After sets of Differences are compared, e.g., by a *t* test or interval. It has been proposed in various forms, e.g., by Campbell and Stanley (1966), Eberhardt (1976), Green (1979), Mathur et al. (1980), Skalski and McKenzie (1982), and Stewart-Oaten et al. (1986). We use the "BACI" acronym because it is familiar, but it omits the time-series aspect, and "Control" is potentially misleading. The unaffected sites "control" (reduce) extraneous variation (cf. Harvey 1989) but, unlike experimental controls, are not used to measure it.

The second approach is due mainly to Underwood (1992, 1993, 1994, 1996), but endorsed by several others (e.g., Green 1993, Otway 1995, Otway et al. 1996*a*, *b*, Glasby 1997, Garrabou et al. 1998, Roberts et al. 1998). We call it "impact vs. reference sites" (IVRS). It also uses Impact and unaffected sites from both Before and After the alteration. Its estimate of the alteration's "effect" is the same as that of the simplest BACI (Table 2), but it gives a very different estimate of the error for this effect estimate: rather than using variation of difference over time, it uses variation of (average After abundance − average Before abundance) over the Control sites. This estimate, and inferences using it, are based on assumed random site selection, possibly restricted to avoid dependence between sites.

This approach requires many reference sites, so may be expensive. Underwood (1992:175) claims the expense is needed, since BACI's "lack of replicated control sites provides insufficient evidence" and "no logical or rational reason" for concluding that an impact was due to the alteration. It "would not be accepted in normal and routine ecological and experimental analysis" and "would always be rejected by reputable journals" (Underwood 1992:175). IVRS methods "are demonstrably superior in logic and for interpretation," "will provide better evidence of causal links" (Underwood 1992:173 and 176) and (Underwood 1994: 155, 155, and 152 and 154) can address problems BACI cannot, such as an alteration effect of unknown spatial extent, inability to sample sites simultaneously, and "spatial and temporal interactions" in abundances.

This paper demonstrates the opposite. BACI proceeds from natural definitions of an alteration-induced "effect" and of "uncertainty" to an interpretable set of results. It is more general than is usually realized. Most of its alleged weaknesses stem from misunderstanding of its goal and of the sources of error in its effect estimates. Other weaknesses exist only for the most parsimonious model, which is not always plausible but can be extended. As presented, IVRS depends on "random site" assumptions, which we show to be untenable. However, we give a model-based justifica-

TABLE 3. Outline of analysis methods discussed in this paper.

Intervention analysis and BACI

- Group 1. Intervention Analysis, IA
  - Data: one value at Impact for each of T_B Before times and T_A After times
  - a) Compare Before and After means, assuming independent errors
  - b) Same as (a), but allow for correlated errors, seasonal variation, and covariates
- Group 2. Before–After-Control–Impact, BACI
  - Data: Like IA, but one value at Impact and at each of N_C Controls at each time
  - a) Like 1(a) or 1(b), but use $I-C$ differences instead of I values (See Table 2.)
  - b) Like 1(a) or 1(b), but use C (or average of C's) as covariate
  - c) Like 2(b), but multiple regression, using controls, other covariates; nonlinear models
  - d) Gradient extension: several impact sites, effect = function of distance from alteration

ANOVA-based analyses

- Data: Same as Group 2, except for 3(c), 5(b) and 6(d)
- Group 3. Impact vs. reference sites (IVRS)
  - a) Compare single Impact value to sample of Control values (See Table 2.)
  - b) Use of 3(a) at 2 levels to gauge unknown extent of effect
  - c) All sites potentially affected: compare variance of B − A differences among sites to pooled Time × Site interaction from within Before and After periods
  - d) Repeated-measures ANOVA, comparing Impact to sample of Control sites
- Group 4. Changes in mean using residual variation for error
  - a) t approach comparing B−A difference at Impact vs. a sample of Control sites
  - b) Use of 4(a) at two levels, as in 3(b)
- Group 5. Changes in mean using temporal variation for error
  - a) Equal variance version of 1(a) or 2(a)
  - b) t approach when sampling times not matched between sites
- Group 6. Changes in temporal variance (see Appendix)
  - a) F test comparing Before and After temporal variation (with or without Controls)
  - b) F test comparing temporal variation to sampling error
  - c) Variance tests at two effect levels as in 3(b) and 4(b)
  - d) N_L sites, all potentially affected; F test cn Time × Location interaction
  - e) Multiple use of 6(a), 6(b) when times within period (Before or After) are unevenly spaced and divided into subperiods

tion for an IVRS approach with a different design and analysis than Underwood's. It has stronger assumptions and less verification, flexibility, and power than BACI, but is important because it may be the only option when there are no Before data.

We also describe and briefly review three other approaches. One uses sampling error to estimate the error in the effect estimate. A second uses temporal variation, like IA and BACI, but uses multiple sites without matching the sampling times. The third targets change in variance rather than mean. None of these is valid except under implausible assumptions.

We outline the analyses in the next section. In the third section we derive the IA and BACI models from first principles, discussing appropriate aims and the interpretation of "chance" in models of processes with unpredictable temporal or spatial components. The fourth section reviews alleged weaknesses of this approach, and some of the claimed improvements. The fifth section addresses the "random sites" justification for the IVRS approach, and discusses a model-based justification, comparing it with IA and BACI. The Discussion draws some lessons.

ANALYSES

This section outlines five Groups of analyses. To fix ideas, we assume that the data consist of estimated abundances (e.g., from core samples, diver counts, net hauls, etc.) taken at a set of times Before the alteration and at another set After it, at the Impact site and at a set of sites called "Controls," which are believed unlikely to be affected by the alteration. Frequently used symbols are in Table 1.

The first two Groups, Intervention analysis (IA) and BACI, compare the Before alteration part of a time series of Impact site values to the After alteration part. The other Groups were all proposed as ANOVA-based analyses by Underwood (1991, 1992, 1993, 1994). Group 3, Impact vs. reference sites (IVRS), compares a summary of the Impact site time series with the corresponding summaries at a set of reference sites. Groups 4 and 5 compare the Before and After time series using sampling error (4) or temporal variation (5) as the error term. Tables 3 and 4 outline/summarize the analyses.

The main contrast in this paper is between the time-series approaches (IA and BACI) and the spatial approach (IVRS, especially Group 3(a) analysis below). Table 2 illustrates how this draws different inferences than the simplest BACI analysis, Group 2(a), from the same data.

The analyses for IA and BACI are incomplete because these approaches are open ended: they cope with serial correlation, covariates, and non-additivity by including them in explicit models, of which there are too many to list. Those for the other Groups are more complete, but sources and some details are in the Appendix.

Throughout this section, "t test" stands for "t test or t confidence interval." We prefer the latter since tests are rarely useful in environmental decision making (Stewart-Oaten 1996*b*).

Intervention analysis and BACI

These analyses estimate mean abundance (and related quantities) at the Impact site from the Before data, $\{I_B\}$, and again from the After data, $\{I_A\}$; the "effect" is estimated by comparing the two estimates. Error arises from sampling error and natural temporal variation.

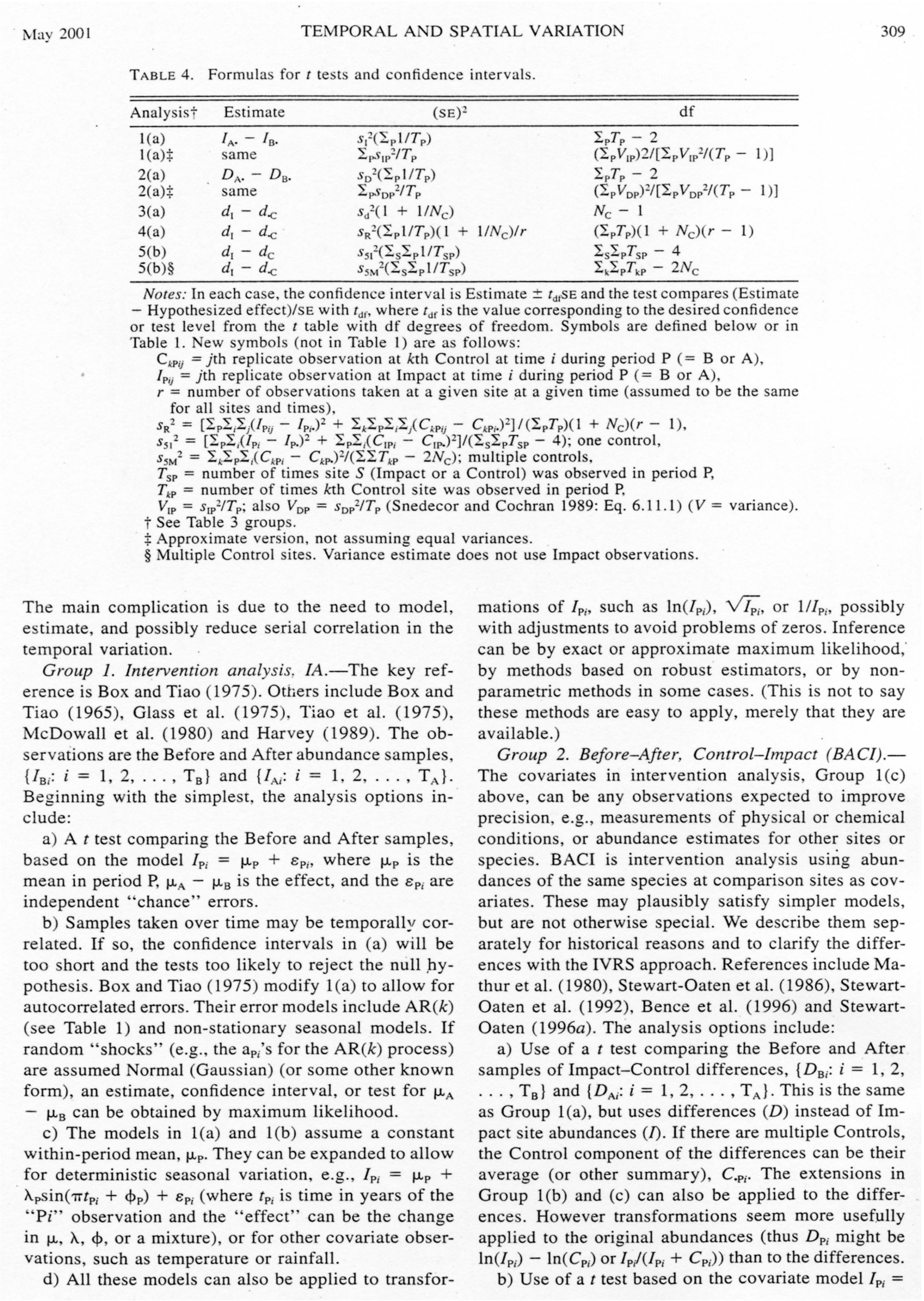


TABLE 4. Formulas for t tests and confidence intervals.

Analysis†	Estimate	(SE)^2	df
1(a)	$I_{A\cdot} - I_{B\cdot}$	$s_I^2(\Sigma_P 1/T_P)$	$\Sigma_P T_P - 2$
1(a)‡	same	$\Sigma_P s_{IP}^2/T_P$	$(\Sigma_P V_{IP})2/[\Sigma_P V_{IP}^2/(T_P - 1)]$
2(a)	$D_{A\cdot} - D_{B\cdot}$	$s_D^2(\Sigma_P 1/T_P)$	$\Sigma_P T_P - 2$
2(a)‡	same	$\Sigma_P s_{DP}^2/T_P$	$(\Sigma_P V_{DP})^2/[\Sigma_P V_{DP}^2/(T_P - 1)]$
3(a)	$d_I - d_{\cdot C}$	$s_d^2(1 + 1/N_C)$	$N_C - 1$
4(a)	$d_I - d_{\cdot C}$	$s_R^2(\Sigma_P 1/T_P)(1 + 1/N_C)/r$	$(\Sigma_P T_P)(1 + N_C)(r - 1)$
5(b)	$d_I - d_C$	$s_{5I}^2(\Sigma_S \Sigma_P 1/T_{SP})$	$\Sigma_S \Sigma_P T_{SP} - 4$
5(b)§	$d_I - d_{\cdot C}$	$s_{5M}^2(\Sigma_S \Sigma_P 1/T_{SP})$	$\Sigma_k \Sigma_P T_{kP} - 2N_C$

Notes: In each case, the confidence interval is Estimate ± t_{df}SE and the test compares (Estimate − Hypothesized effect)/SE with t_{df}, where t_{df} is the value corresponding to the desired confidence or test level from the t table with df degrees of freedom. Symbols are defined below or in Table 1. New symbols (not in Table 1) are as follows:

C_{kPij} = jth replicate observation at kth Control at time i during period P (= B or A),
I_{Pij} = jth replicate observation at Impact at time i during period P (= B or A),
r = number of observations taken at a given site at a given time (assumed to be the same for all sites and times),
$s_R^2 = [\Sigma_P \Sigma_i \Sigma_j (I_{Pij} - I_{Pi\cdot})^2 + \Sigma_k \Sigma_P \Sigma_i \Sigma_j (C_{kPij} - C_{kPi\cdot})^2]/(\Sigma_P T_P)(1 + N_C)(r - 1)$,
$s_{51}^2 = [\Sigma_P \Sigma_i (I_{Pi} - I_{P\cdot})^2 + \Sigma_P \Sigma_i (C_{IPi} - C_{IP\cdot})^2]/(\Sigma_S \Sigma_P T_{SP} - 4)$; one control,
$s_{5M}^2 = \Sigma_k \Sigma_P \Sigma_i (C_{kPi} - C_{kP\cdot})^2/(\Sigma\Sigma T_{kP} - 2N_C)$; multiple controls,
T_{SP} = number of times site S (Impact or a Control) was observed in period P,
T_{kP} = number of times kth Control site was observed in period P,
$V_{IP} = s_{IP}^2/T_P$; also $V_{DP} = s_{DP}^2/T_P$ (Snedecor and Cochran 1989: Eq. 6.11.1) (V = variance).

† See Table 3 groups.
‡ Approximate version, not assuming equal variances.
§ Multiple Control sites. Variance estimate does not use Impact observations.

The main complication is due to the need to model, estimate, and possibly reduce serial correlation in the temporal variation.

Group 1. Intervention analysis, IA.—The key reference is Box and Tiao (1975). Others include Box and Tiao (1965), Glass et al. (1975), Tiao et al. (1975), McDowall et al. (1980) and Harvey (1989). The observations are the Before and After abundance samples, $\{I_{Bi}: i = 1, 2, \ldots, T_B\}$ and $\{I_{Ai}: i = 1, 2, \ldots, T_A\}$. Beginning with the simplest, the analysis options include:

a) A t test comparing the Before and After samples, based on the model $I_{Pi} = \mu_P + \varepsilon_{Pi}$, where μ_P is the mean in period P, $\mu_A - \mu_B$ is the effect, and the ε_{Pi} are independent "chance" errors.

b) Samples taken over time may be temporally correlated. If so, the confidence intervals in (a) will be too short and the tests too likely to reject the null hypothesis. Box and Tiao (1975) modify 1(a) to allow for autocorrelated errors. Their error models include AR(k) (see Table 1) and non-stationary seasonal models. If random "shocks" (e.g., the a_{Pi}'s for the AR(k) process) are assumed Normal (Gaussian) (or some other known form), an estimate, confidence interval, or test for $\mu_A - \mu_B$ can be obtained by maximum likelihood.

c) The models in 1(a) and 1(b) assume a constant within-period mean, μ_P. They can be expanded to allow for deterministic seasonal variation, e.g., $I_{Pi} = \mu_P + \lambda_P \sin(\pi t_{Pi} + \phi_P) + \varepsilon_{Pi}$ (where t_{Pi} is time in years of the "Pi" observation and the "effect" can be the change in μ, λ, ϕ, or a mixture), or for other covariate observations, such as temperature or rainfall.

d) All these models can also be applied to transformations of I_{Pi}, such as $\ln(I_{Pi})$, $\sqrt{I_{Pi}}$, or $1/I_{Pi}$, possibly with adjustments to avoid problems of zeros. Inference can be by exact or approximate maximum likelihood, by methods based on robust estimators, or by nonparametric methods in some cases. (This is not to say these methods are easy to apply, merely that they are available.)

Group 2. Before–After, Control–Impact (BACI).—The covariates in intervention analysis, Group 1(c) above, can be any observations expected to improve precision, e.g., measurements of physical or chemical conditions, or abundance estimates for other sites or species. BACI is intervention analysis using abundances of the same species at comparison sites as covariates. These may plausibly satisfy simpler models, but are not otherwise special. We describe them separately for historical reasons and to clarify the differences with the IVRS approach. References include Mathur et al. (1980), Stewart-Oaten et al. (1986), Stewart-Oaten et al. (1992), Bence et al. (1996) and Stewart-Oaten (1996*a*). The analysis options include:

a) Use of a t test comparing the Before and After samples of Impact–Control differences, $\{D_{Bi}: i = 1, 2, \ldots, T_B\}$ and $\{D_{Ai}: i = 1, 2, \ldots, T_A\}$. This is the same as Group 1(a), but uses differences (D) instead of Impact site abundances (I). If there are multiple Controls, the Control component of the differences can be their average (or other summary), $C_{\cdot Pi}$. The extensions in Group 1(b) and (c) can also be applied to the differences. However transformations seem more usefully applied to the original abundances (thus D_{Pi} might be $\ln(I_{Pi}) - \ln(C_{Pi})$ or $I_{Pi}/(I_{Pi} + C_{Pi})$) than to the differences.

b) Use of a t test based on the covariate model I_{Pi} =

$\mu_P + \beta_P C_{Pi} + \varepsilon_{Pi}$ (P = B or A). "C_{Pi}" could be the average of multiple Control sites. The "effect" can be on the intercept (e.g., $\mu_A - \mu_B$), the slope ($\beta_A - \beta_B$), or taken to be the change at a particular value of the Control site, representing "standard" conditions: $\mu_A + \beta_A C - (\mu_B + \beta_B C)$. Unequal Before and After variances can be dealt with—easily if μ and β can both be affected (each regression can be fitted separately and the results combined), less easily otherwise. These analyses can also be extended to allow the errors to be serially correlated. "I" and "C" can represent transformations of the raw abundance estimates.

c) The model in 2(b) can be extended to include explicit multiple Controls (rather than a single average), other covariates, and nonlinear models. Control sites could be separate covariates or combined into groups (e.g., North (sites north of the Impact site) and South), using the average of each group as a covariate. Other covariates (like temperature) could be included and autocorrelated errors allowed for. The main variables (I_{Pi}), the Control values (C_{Pi} or C_{kPi}), and other covariates could also be transformed. These methods are also not easy; we briefly discuss model choice, checking, and robustness later (see *Intervention analysis and BACI: BACI using comparison sites and covariates. . . : Errors in variables* and *Feasibility and model uncertainty*, below).

d) "Gradient" extensions of IA and BACI could use several Impact sites, and model the effect at a site as a function of its distance from the alteration. This involves difficult modeling problems, to specify relations between Impact and Control sites or between effect and distance, and for spatial and temporal correlation. Ellis and Schneider (1997) describe a version. Wiens and Parker (1995) discuss it for "After only" data.

ANOVA-based analyses

These analyses usually estimate the "effect" by $I_{A\cdot} - I_{B\cdot} - [C_{\cdot A\cdot} - C_{\cdot B\cdot}]$. This equals both $D_{A\cdot} - D_{B\cdot}$, the difference between the average After and Before differences as in the simple BACI model 2(a), and $d_I - d_{C\cdot}$, the difference between the After−Before change at Impact and the average After−Before change at the Controls (see Table 2). They differ in the error they attach to this estimate (see Table 4).

Group 3 uses variation among the Controls, which can arise from spatial variation, temporal variation that differs among sites, and sampling error. Group 4 uses sampling error. Group 5 uses variation over time, which includes sampling error. Group 5(a) analysis is a simple BACI; the rest of Groups 4 and 5 are valid only under implausible models. We present them, comment briefly, then ignore them. We defer a sixth Group, targeting a change in temporal variance rather than mean, to the *Response* section, below, and the Appendix.

The main references are Underwood (1991, 1992, 1993, 1994) (1996 reprints 1994). His "asymmetric ANOVA" tables usually give only the "Source" and degrees of freedom ("df"), omit models and formulae (the "SS" and "MS" columns), and contain ambiguities and seeming errors. However, their F tests for means are simple t tests, and this form clarifies their motivation and assumptions. The Appendix gives sources, more details, and outlines proof that our descriptions are equivalent to Underwood's.

Group 3. Impact vs. reference sites (IVRS), using spatial variation for error.—

a) Use of a t test comparing the Impact "sample," the single value d_I, to the Control sample, $\{d_{C1}, d_{C2}, \ldots\}$.

b) Use of 3(a) at two levels, for when the effect's extent is unknown. For example, there may be an Impact site and several Control sites in a bay, to test an alteration there. If the test in 3(a) is not significant, these sites are combined (averaged), and the test in 3(a) used to compare this "Impact bay" with a set of Control bays.

c) If all the sampled sites are potentially affected because the effect's extent is unknown: use of an F test comparing the variance among sites of the After−Before differences to the pooled Time × Site interaction mean square (MS) from within the Before and After periods. The motivation may be that an alteration effect will change the sites by different amounts, so the variance of the After−Before differences will be greater than expected from natural variation within periods.

d) Use of repeated-measures ANOVA as a method for analyzing "longitudinal data," where each unit (site) has been measured on several occasions, without collapsing the data to a single value per site as in 3(a). The particular analysis depends on the model for the mean, e.g., how it allows for natural differences between Impact and the Control sites, temporal variation in the alteration effect, and autocorrelated errors.

Group 4. ANOVA, using residual variation for error.—a) Use of a t test like Group 3(a) except that the variance of the effect estimate, $d_I - d_{C\cdot}$, is estimated by the pooled sampling variances from each site visit, rather than by the variation among $d_{C1}, d_{C2}, \ldots$. In some cases this test is done only when the "between vs. within" sites comparison is not significant.

(b) Use of the t test in 4(a) twice, once at each of two levels of possible effect, in the same way as 3(a) is used twice in 3(b).

Comments. This group assumes that actual abundances at the sites (the values a census would give) fluctuate in perfect unison over time. This is not plausible, even as an approximation, nor is made so by a nonsignificant preliminary test. Most statistical analyses are approximations: low-degree polynomial regressions, and time-series models with low-order autocorrelations, are justified by tests or plots showing adequacy rather than absolute truth. But these approximations are plausible, and often can predict future values better than more complicated models. They are preferred for interpretability and tractability, not for greater nominal precision. Using the residual variance in assessment replaces tractable and interpretable models that may be nearly right by a model known to be wrong. If it gives a smaller *P* value or shorter confidence interval, then we know that we are claiming more precision than is justified; if it doesn't, it has no point.

Group 5. ANOVA, using temporal variation for error.—

a) Use of the standard (equal variances) version of the *t* test of Group 1(a) or Group 2(a).

(b) Use of a *t* confidence interval or *t* test when sampling times differ between sites (see Table 4).

Comments. Analysis in Group 5(a) is a special case of Group 1(a) or Group 2(a) analyses.

Underwood (1992:160) claims that analysis in Group 5(b) is "a considerable advance on attempting to analyze the data using *t* tests." It *is* a *t* test, whose test statistic can be written as (estimated effect)/(1SE of estimate), and this form clarifies its assumptions: that observations from different sites are independent and that observations from the same site at different times are independent with the same mean and variance. If these were true, the method of Group 1(a) (the naive IA) would be easier and better: the Controls would contribute only random error and extra assumptions. They provide "control" only under an unlikely model: a natural change near the start-up time affects all later observations at all sites by the same non-negligible amount, while deviations at all other times are uncorrelated between sites and between times. Correlation between values taken over time is the main concern in IA, and a main reason for BACI's comparison sites. (Reducing temporal variance is the other.) The analysis in Group 5(b) wishes away the main problems addressed by IA and BACI.

Intervention Analysis and BACI

This section describes Intervention Analysis (IA) and Before–After, Control–Impact (BACI). IA is a way to use samples taken at an Impact site at several times Before and After an "intervention" to determine its effect on some variable at that site. A major problem in using IA to assess effects on biological variables like abundance is high uncertainty (or low power) due to natural variation and serial correlation. Such problems are often reduced by using covariates. BACI is IA with the Impact site data matched by data from one or more comparison sites to control for some of the natural variation.

We focus on concepts rather than mathematics for substantive reasons as well as readability. Inappropriate use of probability concepts cause some of the errors we will later see in the IVRS (impact vs. reference sites) approach. The relation of models to reality, the definitions of quantities of interest, and the meaning of chance and uncertainty require more attention as statistical applications widen into areas where uncertainty arises from sources other than sampling and measurement error (e.g., Chatfield 1995, Draper 1995, Buckland et al. 1997, Mallows 1998). Most time series and spatial data sets are of this type.

Assessments are largely time-series problems. The target quantities and the "chances" appearing in confidences, *P* values, and other measures, may be more than purely social constructs, but are not objective physical quantities that can be taken for granted. The validity of assessment methods rests in part on their aims and definitions being acceptable to "reasonable" non-scientists, not only because final decisions will often be made by non-scientists but also because the methods use concepts like "chance," "mean," etc., which describe human states of mind as well as the physical world.

We assume here that the assessment tasks are (1) to describe the effect of the alteration on the abundance of a given species at the Impact site and (2) to measure the reliability (or uncertainty) of this description. Both tasks involve definition that is not as simple as may appear.

IA Problem statement 1: defining an "effect"

What is an effect?—We define *the effect of concern* to be the difference between the abundance at the Impact site after the alteration and the abundance the site would have had if the alteration had not occurred. This natural definition of a treatment effect is widely accepted (e.g., Rubin 1974, Rosenbaum 1984 and 1987, Holland 1986, Cox 1992, Freedman 1994).

Censused abundance and temporal variation.—We define a site's *censused abundance* to be the actual number of individuals at the site (or the number per unit area or volume). It differs from the *observed abundance,* which is affected by sampling error. Observed abundances are expected to vary over time, but this is due to both sampling error and temporal variation in censused abundances stemming from a variety of sources (Table 5).

Thus the "effect" is the difference between two functions of time:

$$A_A(t),\ A_B(t) = \text{the censused abundances at time } t, \text{ under ``alteration'' (After) and ``no alteration'' (Before) conditions.} \quad (1)$$

The omniscient investigator would provide the decision maker with both functions or with a suitable description of the difference, such as $A_A(t) - A_B(t)$ or $A_A(t)/A_B(t)$, for all values of *t* following the alteration. Of course, only one of these functions can exist at any time and we cannot observe future values of either.

Not all of this information would be used, even if we had it. A manager would base decisions on only a limited range of times. This would usually extend beyond the study period, but not to very large values of *t*—e.g., the lifetime of the alteration might be used, or the first 100 yr after its installation. The shorter horizon

TABLE 5. Temporal variation in censused abundances: sources and examples.

Type of variation	Source of variation	
	Environmental examples	Biological examples
Trends	1) sedimentation, erosion	2) evolution, succession
Cycles	3) days, seasons, ENSO†	4) intra-species, closely linked species‡
Irregular	5) storms, upwellings, runoff, spells (dry, hot, windy, . . .)	6) epidemics, invasions, other migrations, births, deaths, encounters
Interactions	7) Storms, upwellings, or invasions during an annual recruitment period	

Note: The examples of sources of variation are numbered for ease of reference.

† El Niño/Southern Oscillation.

‡ Parasitoids, pathogens, specialist predators, essential resources, competitors.

makes prediction easier, since geological changes can be ignored.

Even over this range, the full sets of function values would not be used—e.g., small or short-lived fluctuations would rarely be of interest. Instead, a few summaries would be calculated, for comparison with similar summaries of alteration effects on other species and on non-biological variables (e.g., economic). Among these might be the mean (calculated overall, or for particular seasons, or for conditions like El Niño or northerly currents), the times to local extinction and subsequent recovery, and measures of change in the amplitude or frequency of fluctuations.

Prediction tasks.—Our targets are the functions $A_A(t)$ and $A_B(t)$ from startup or installation of the alteration to the end of the time horizon or period of interest ($t_S < t < t_H$) (see Table 1). Instead of observing these functions exactly and continuousiy, we observe them with error as I_{Bi} (Before) and I_{Ai} (After) for a set of times before (t_{Bi}: $t_{B1} < t_{B2} < \ldots < t_{BT_B} < t_S$) and after ($t_{Ai}$: $t_S < t_{A1} < t_{A2} < \ldots < t_{AT_A}$) the alteration. Therefore, we need to predict the values of these functions (or summaries of these values) from these data.

In this paper we treat the prediction period as being long relative to the sampling period for the After data, so that virtually the entire assessment problem concerns a period after the last sampling time, t_{AT_A}. We thus avoid the task of estimating $A_A(t)$ between the times t_S and t_{AT_A}, which is one of interpolation (Krishnaiah and Rao 1988) rather than prediction. In most practical cases the time period we ignore is small and the interpolations would often be near the predictions, so the results would be little changed. We thus have:

Problem Statement 1: Predict the difference between the functions $A_A(t)$ and $A_B(t)$ for the period between the end of the study and the time horizon, t_H, with a measure of uncertainty. (This difference may be one or more summaries of a difference function like $A_A(t) - A_B(t)$ or $A_A(t)/A_B(t)$, etc.)

Chance

Measuring uncertainty.—When using a prediction or estimate to make decisions, we need criteria for judging its reliability and comparing it to other estimates. One criterion might be the size of its error, but the error is unknown. An alternative is to measure the reliability of the estimating *method* when applied to data of the type we observe: Are its errors "usually" or "likely to be" small? The method will give varying answers because the data will vary—the observed data are treated as only one instance of data we *could* have observed, and it is to these possibilities that "usually" and "likely" refer.

Probability provides the units to measure uncertainty or reliability. It is hard to define, but ignoring its definition can lead to meaningless calculations. We use the "frequentist" definition, in which "chance processes" can be repeated under "identical" conditions and lead to different results. The probability of any set of possible results is the limit, as $N \rightarrow \infty$, of its relative frequency in N identical repetitions, "independent" in the sense that no set of repetitions affects any other set. Thus it is a property of the limit of many repetitions, even though in fact we usually perform (or observe) only one.

This definition agrees with the intuitive idea of probability as (number of favorable cases) ÷ (total number of cases) and is consistent with the theory underlying other types of measure (e.g., standard axioms lead to the laws of large numbers)—but consistency and agreement with intuition do not establish a relationship with the real world. Exactly identical conditions are both impossible and undesirable: they would lead to identical results rather than a distribution, except possibly for some aspects of particle behavior. "Apparently identical" seems preferable, but may introduce dependence on an observer. Some scientists reject frequentism, and define probability in terms of personal degree of belief. The resulting Bayesian methods may have much to offer in assessment (e.g., West and Harrison 1989, Raftery et al. 1995, Crome et al. 1996, Ellison 1996, Wolfson et al. 1996) and some of this paper applies to them too, but they introduce new problems so we do not consider them explicitly. Other scientists may accept frequentist definitions as intuitively reasonable—and put them out of mind: after all, "mass," "point," and even "life" are hard to define, too, though clearly useful. Putting them out of mind can lead to trouble.

Inferential "probability:" sources and credibilities.—Claims of randomness and calculations of confidences, *P* values and other frequentist conclusions refer to long-run results from some repeatable multi-outcome process. If we cannot specify the process and the source of the "chance," these claims and calculations have no meaning. When we can, their credibility depends on how well the idealized process matches the real process producing the data. Some categorizations may help assess this. We ignore the credibility problem of misleading description, e.g., when only the "favorable" experiments, summaries, or tests are reported.

Device-based probabilities result from deliberate randomization, using devices like coins or random-number tables to choose individuals ("units") for observation or to assign them to treatments. Nature-based probabilities treat each unit's value as an outcome of a natural process (movement, mate choice, cell division, environmental variation, etc.) that yields different results from apparently identical conditions.

The distinction is rough: devices are "natural" too. If anything, they are the more deterministic, but our ignorance allows us to construct setups that seem to us identical in all relevant ways yet do not lead to identical results. Their credibility is greater because their simple assumptions (independent trials; equal chances of 0, 1, . . . , 9) can be checked, and have been supported, by experiments with coins, random-number generators, etc., at many times and places. (For this reason, we think of "randomly thrown quadrats" as Nature based.) Even so, most of us will randomize again if the randomizing device gives a strong "nonrandom" pattern in a particular case—the probabilities might be "true" but still not reflect anyone's uncertainty well.

Design-based probabilities refer to hypothetical repetitions of the "choice" process by which the unit values observed came to be selected, or assigned to a particular treatment. This group contains all device-based probabilities, like experiments with units assigned randomly to treatments or samples of quadrats chosen from random-number tables. It also contains some Nature-based probabilities, where no explicit randomizing device is used but the natural process determining unit values is assumed unrelated to the choice process. We call this "as if random" sampling by Nature. Organisms from traps, mist nets, commercial suppliers, or "grab samples" in the laboratory are examples, as often are higher-order terms in random-effects ANOVA. Model-based probabilities are calculated from models of the process by which units obtained their values. Measurement error, time series, spatial statistics, and analyses of survival or reliability are examples.

This distinction is also rough. Nature-based, design-based probabilities often model the distribution from which the units were chosen, or the form of relations with covariates. Some model-based methods use implicit models: e.g., assumption of measurement error Normality arises from convenience, experience, and perhaps the idea that these errors are roughly sums of many small errors to which the central-limit theorem should apply. There are cases where device-based and model-based probabilities giving different answers can both be defended; e.g., we could use randomly chosen points to estimate percent cover of algae over an area, but base the estimate and its uncertainty on a model allowing for correlation between points.

The credibility of nature-based assumptions varies. "As if random" observations may come from a subpopulation that is smaller than the target population but seems likely to reflect it in all relevant ways; e.g., medical trials select patients from those available at the time, but the results are used mainly for future patients. There may be little or no device-based selection and units may differ in ways suspected to be relevant but hoped to be small or to "balance out": distributions of potential confounding variables in the samples being compared are assumed similar enough to have arisen by genuine random sampling from a common population, and thus to be treatable as part of the error. This can be risky; e.g., the Salk vaccine trials could have used children whose parents refused permission as "Controls," hoping that permission and vulnerability were unrelated, but these children were mainly from lower socio-economic classes, where immunity due to mild forms of early childhood polio was more common (Freedman et al. 1998). Implicit models can lead to meaningless, misunderstood, or arbitrary conclusions; e.g., Breiman (1995) argues that a standard inference of sex bias in employment uses an imaginary replication of something nonreplicable (a particular corporation) and therefore "makes no sense." Similarly, the Impact site in assessment is nonreplicable (see BACI: Response to comments, below). Thus, when the "chance" in inference statements is to be attributed to nature, it is important to say where the "chance" comes from: What are the "repeatable" processes on which it is based, and how credible are models used to describe them?

Time series

Introducing chance: the need for a model.—We first consider the problem of predicting the censused abundance, $A_B(t)$, at a time after startup, having observed $A_B(t)$ exactly and continuously during the Before period. I.e., we assume all Before times are observed, with no sampling error. We do so partly to simplify the problem but also to make a point: since this prediction cannot be guaranteed, the "chance" in our real problem cannot lie only in design choices, either of units to sample (sampling error) or of Before-period sampling times.

The extra chance must lie either in the observed Before-period values, or in the future value we want

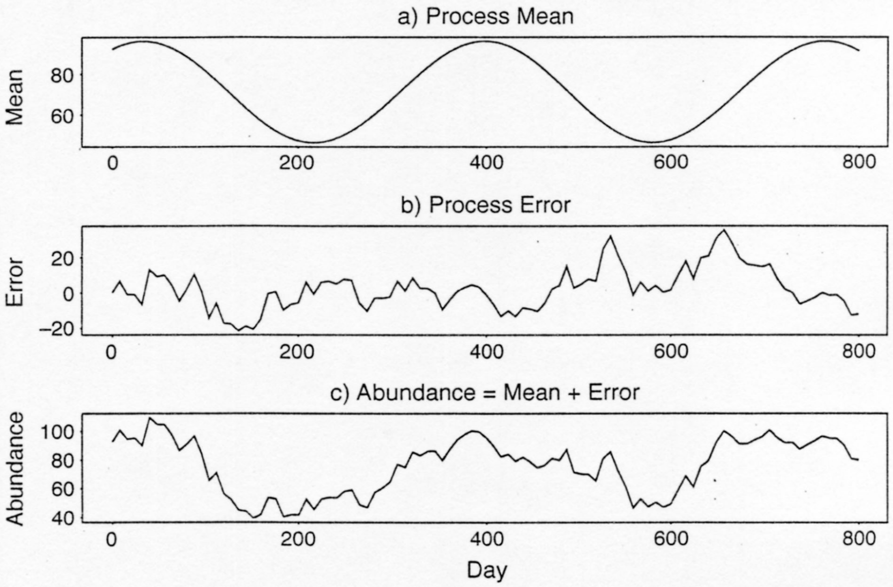


FIG. 1. Simulated example of a time-series process adding process mean and process error. (a) The deterministic part (process mean) is $F(t) = 70 + 25 \sin(1 + 2\pi t/365)$, where t is in days. (b) The stochastic part (process error) is the AR(1) model $\varepsilon(t) = \rho\varepsilon(t - 1) + a(t)$, with $\rho = 0.985$ and $\mathrm{Var}(a) = 0.6$. (c) The sum of (a) and (b).

to predict, or both. At first it may seem that it should lie only in the future value: we are not uncertain about past values. But if the observed $A_B(t)$ values do not involve chance, then we cannot use them to assess the chance error in the future $A_B(t)$ value, and thus the reliability of our prediction. We cannot make the prediction at all unless we assume some relationship between the past of $A_B(t)$ and its future.

A model-based way to solve such problems is to treat the entire function, $A_B(t)$, both observed past and predicted future, as arising from a natural process, involving deterministic and stochastic parts. The deterministic part usually has known (guessed) form with unknown parameters, e.g., a seasonal sine wave plus a linear trend,

$$F(t) = \mu + \alpha \sin(2\pi t) + \beta \cos(2\pi t) + \gamma t \qquad (2)$$

where t is in years. The stochastic part is a single function, e.g., $A_B(t) - F(t)$, but we cannot anticipate the sizes of future chance errors unless they are the results of processes that have been repeated multiple times in the observed data. We thus need to model the stochastic part as a function of multiple independent "errors." A discrete-time example is the autoregressive process AR(1)

$$\varepsilon(t) = \rho\varepsilon(t - 1) + a(t) \qquad (3)$$

where ρ is a fixed number, the first-order autocorrelation. (The continuous-time version, the Ornstein-Uhlenbeck process, is more complicated; see Arnold 1974 or Priestley 1981:158–168.) The $\varepsilon(t)$'s are not independent (each contains a residue of the last) but are functions of the independent perturbations $a(t)$, $a(t - 1)$, ... Because of this structure, we can estimate distributional properties of the $\varepsilon(t)$'s from observations of them or of $F(t) + \varepsilon(t)$. Eq. 3 can be extended to an AR(k) process (Table 1) and in other ways.

The deterministic and stochastic parts can be combined additively to give

$$A_B(t) = F(t) + \varepsilon(t) \qquad (4)$$

as in Fig. 1 (with $\gamma = 0$), or multiplicatively to give

$$A_B(t) = F(t)e^{\varepsilon(t)} \qquad (5)$$

or in many other ways, e.g., by making $A_B(t)$ a transformation of one of these. Deciding the appropriate model forms for F, ε, and their combination is often difficult and uncertain. A linear trend is usually inappropriate for abundance, though it might approximate a long cycle or slow recovery over a limited time range, or a declining population might be represented by $A_B(t) = \exp\{F(t) + \varepsilon(t)\}$ if γ is negative. Seasonal variation in abundance may not be like a sine wave, and it is easy to imagine alternatives to the implicit assumption here that the chance disturbances, $a(t)$, have additive effects which decline exponentially over time. However, the combination of a deterministic forcing function and a chance function that can be decomposed into independent parts can allow for much complexity of behavior.

The main point is that introducing chance into a time-series problem requires a model whose form is usually uncertain. Under it, every value of the outcome function, $A_B(t)$, both past and present, involves chance: the entire function can be regarded as a random choice (a "realization") from a collection of possible functions. The random values, $A_B(t)$, for $-\infty < t < \infty$, are not usually independent nor identically distributed. For $t \neq u$, the means and variances of the random values $A_B(t)$ and $A_B(u)$ may be different, so both the "mean," $\mu_B(t) = E\{A_B(t)\}$, and the "variance," $\sigma_B^2(t) = \text{Var}\{A_B(t)\}$, of the process A_B are functions of t. Neither is necessarily related to the average or the variation of $A_B(t)$ over any range of time values, although most models imply such relationships, so these parameters can be estimated. If $A_B(t)$ is a future value, and $\hat{A}_B(t)$ is its prediction based on the Before values, then the "chance" in the prediction error, $\hat{A}_B(t) - A_B(t)$, comes from both terms.

Interpreting "chance."—"Nature" is the source of the chance, which arises from events and cycles like those of Table 5. The chance part of $A_B(t)$ arises because a particular sequence of these chance events occurs; if a different sequence had occurred, $A_B(t)$ would have been different. Treating such events as "chance" seems reasonable in many cases, since they are numerous and our ability to predict or measure them or their effects is poor. It also links observed values to future ones: both are outputs of the same processes, both chance and deterministic.

Since Nature is also the source of the forcing function, the distinction between random and deterministic variation is partly arbitrary: goals, predictability (regularity), and mathematical convenience all play a role. In Table 5, sources 5–7 will usually be "chance," as will source 4, except in special cases. For source 3, seasonal variation seems regular and well-understood enough to be "systematic," but can be treated as random (e.g., Box and Jenkins 1976), while ENSO is likely to be "random" until our ability to predict its occurrence and effects improves.

In time series, more than one chance model is likely to be credible. Nature's randomizing is usually repeatable only in imagination. It may be plausibly represented as combining many independently replicable events, some of which can be observed or checked (e.g., from weather records, experiments, or "replicates" made up of separated segments of the time-series record). Goodness of fit can be checked for any particular model. Still, the "chance" represents our ignorance as much as objective reality. Choices of what features to model, which ones to treat as deterministic, model forms for both deterministic and stochastic parts, and how to combine them, are often based as much on plausibility, flexibility, and tractability as on known mechanisms or goodness of fit. There will usually be many models that fit about equally well, especially when data are sparse as is common in assessment. Even if one fits better than others, this is not proof of reality (if any model can be "real"). It will often be important to present results from a range of models, which are compared for both substantive plausibility and formal fit to the data.

IA Problem statement 2: prediction and parameter estimation

We have presented the assessment task as predicting the differences between future values under "alteration" and "no alteration" conditions, or summaries of these differences. We now argue that it can be redefined as estimating the difference between parameters (usually means or parameters describing mean functions) of "alteration" and "no alteration" models. This requires justification.

Most predictions are parameter estimates. E.g., from observations of $A_B(t_i) = F(t_i) + \varepsilon(t_i)$ in Eqs. 2 and 3, we might obtain estimates $\hat{\mu}$, $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$, and thus predict a future value as $\hat{A}_B(t) = \hat{\mu} + \hat{\alpha}\sin(2\pi t) + \hat{\beta}\cos(2\pi t) + \hat{\gamma}t$. Usually, the uncertainty in prediction is greater than that in estimation: the former, estimate−future value $= \hat{A}_B(t) - A_B(t)$, has two parts contributing independent variation: (1) estimation error = estimate−parameter value $= \hat{A}_B(t) - F(t)$, and future variation = parameter value−future value $= F(t) - A_B(t)$, where $F(t) = \mu + \alpha\sin(2\pi t) + \beta\cos(2\pi t) + \gamma t$. We must show that this is not a problem here. When our arguments do not apply, substituting parameter estimation for prediction may need more justification than it usually gets.

One argument is that our "effect" is the difference between future values, not the values themselves, so:

$$\text{Prediction error} = [\hat{A}_A(t) - \hat{A}_B(t)] - [A_A(t) - A_B(t)].$$

If abundance is the sum of a forcing function and a stochastic term, as in $A_B(t) = F_B(t) + \varepsilon_B(t)$ and $A_A(t) = F_A(t) + \varepsilon_A(t)$, then

$$A_A(t) - A_B(t) = [F_A(t) - F_B(t)] + [\varepsilon_A(t) - \varepsilon_B(t)]$$

i.e., parameter − [parameter−future value]. But if the alteration affects only the forcing function, so that $\varepsilon_A(t) = \varepsilon_B(t)$, then the second term, which is often the larger part of the uncertainty in prediction, is zero.

When the alteration affects the chance component (e.g., changing the response to storms), or the stochastic and deterministic components are not added (as in the model $A(t) = F(t)\exp\{\varepsilon(t)\}$), or the "effect" is not absolute change, $A_A(t) - A_B(t)$, but fractional change, $1 - A_A(t)/A_B(t)$, or another measure, cancellation is likely to be imperfect. Some cancellation will still occur unless the alteration reverses responses (e.g., previously harmful perturbations become beneficial), but the error in predicting a particular future value may be significantly greater than the error in estimating its mean in these cases.

This additional error may be reduced by averaging. Judgments of the damage done by an alteration will usually depend more on effect averages over time (e.g., the average summer effect) than on the effect at any particular time. The average of independent random values converges to its mean as the sample size increases. For time series, the average of the actual series over a time period converges to the average (over time) of the mean function, $\mu(t)$, as the time period increases. There are exceptions, e.g., the natural history of Earth may have been redirected when a chance asteroid led to the extinction of the dinosaurs and the rise of mammals, but it would be hard to justify using such "nonergodic" processes for making decisions. Most standard models and almost all plausible models of natural series like abundances are ergodic (Breiman 1968: chapter 6). Thus the averages of both the "alteration" and "no alteration" future values are likely to be close to their means if the averaging is over a long period, like 100 years or the planned life of the alteration.

This may not be enough. Averages of time series may converge to their means only slowly. If serial correlation is large, and we are concerned mainly with the effect of the alteration over only a few years, then the uncertainty in the predicted difference could be nonnegligibly greater than the uncertainty in the estimated difference of the means, unless the chances largely cancel as above. If the time horizon is short, predictions (forecasts and interpolations) based on both the time-series model and the recent values (e.g., Box and Jenkins 1976) may be better than estimates of means.

Thus the mean squared error of the estimated difference of means is often, but not always, a reasonable approximation to the mean squared error of the predicted difference of future values. When this is true we obtain:

Problem Statement 2: Estimate the difference between the means of functions $A_B(t)$ and $A_A(t)$ for the period between the end of the study and the time horizon, t_H, with a measure of uncertainty.

As before, the "mean" of $A_B(t) = \mu_B(t)$, a function of time, so the difference between the means is also a function of time. In practice, we usually will not need this function for each time point, but only summaries. For example, for most models, $\mu_A(t) - \mu_B(t)$ will not depend on the year but only on the time of year, so averages over a year or a season will be useful.

Variation, correlation, and sparse data

Two error sources.—We now drop the assumptions of continuous and exact observation of censused abundances, $A_B(t)$ or $A_A(t)$. Instead, we observe (or estimate) abundances at discrete sets of times:

$$I_{Bi} = A_B(t_{Bi}) + \xi_{Bi} \qquad I_{Ai} = A_A(t_{Ai}) + \xi_{Ai} \qquad (6)$$

where the ξ's are independent sampling errors. This is a new source of error, additional to the "process error" in the stochastic models of $A_B(t)$ and $A_A(t)$ (e.g., the $\varepsilon(t)$ of Eq. 4). They do not change the problem, though they complicate it, e.g., adding independent sampling errors to an AR(k) process for the censused abundances gives an autoregressive moving average process ARMA(k, k) for the estimated abundances (Cox 1981; Table 1). Their variances and other parameters can be estimated if several samples are taken at each sampling time, but are usually of less interest; the targets are parameters of the distributions of $A_B(t)$ and $A_A(t)$.

Errors from both sampling and temporal fluctuations are often large—standard deviations may be as great as the abundance itself. Sampling errors are large because of spatial patchiness, sampling difficulty, etc. They can be reduced, e.g., the variance of the estimated censused abundance can be made arbitrarily close to zero by taking enough samples or sampling a large enough fraction of the Impact site on each visit.

Correlated temporal errors.—"Errors" due to temporal variation in the censused abundances are harder to reduce. Events and cycles like those in Table 5 can have long-lasting effects on abundance even when their environmental effects quickly disappear (e.g., source 5). Usually, a chance excess or deficit is likely to remain for some time: abundances close in time are likely to be close in value, so both censused abundances, A, and observed abundances, I, will be serially correlated.

Increasing the number of sampling times without increasing the length of the study period decreases the gaps between samples. This increases the correlation between adjacent values—much of the information from the additional samples is redundant. Fig. 1b was generated from Eq. 3, using a correlation of 0.985 between values 1 d apart; one could observe it for months without seeing most of the variability. Fig. 2 shows more formally that, for Eq. 3, the variance of the estimated mean does not shrink to zero as sample size increases within a fixed study period. Its limit (an infinite set of sampling times, or continuous sampling) can still be too large for effective decision making if the study period is short or the correlation is large.

These features also apply in more realistic models: estimation errors can be large because natural variation is high, correlation declines only slowly over time, and study periods are short. The Before period, especially, cannot usually be extended to permit better estimation. Eq. 3 may be optimistic; e.g., recovery or decline from a crash or peak caused by source 5 (Table 5) may be slower than exponential.

High correlation might also not be detected in a short period. If the first and last observations within a period are highly correlated, a tendency for observations closer in time to be closer in value may be obscured. This can arise in ecological time series when occasional rapid changes in abundance are followed by periods of relative constancy. For example, a crash or boom occurring in or near the "Interim" period (between Before and After) might last through much of the After period and hide or mimic an alteration effect.

High temporal variability and correlation make estimation inaccurate: they force us to make a large allowance for uncertainty. Perhaps worse, long-term correlation can conceal the inaccuracy, so our allowance for uncertainty is too small—our conclusions are in-

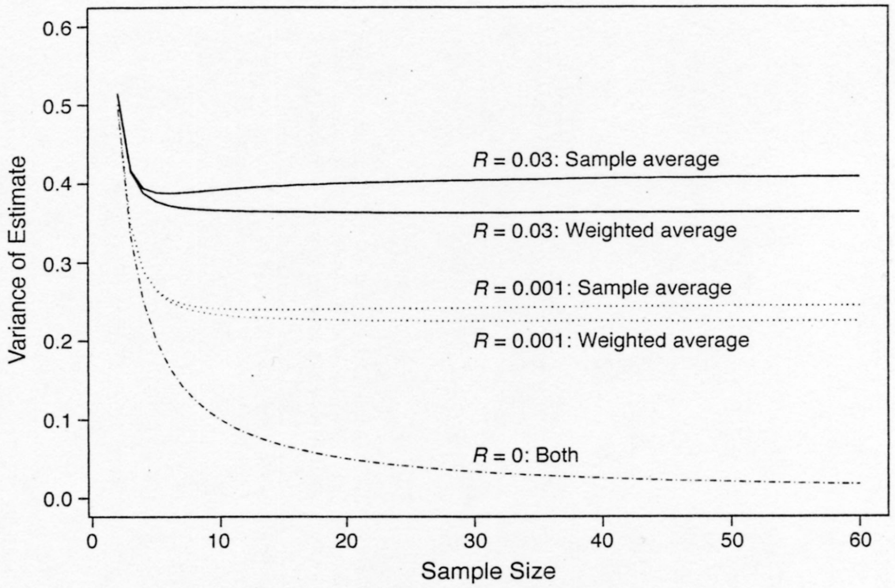


FIG. 2. The effect of increasing sampling frequency on the variance of the estimated mean (sample average or best weighted average) when the errors follow the AR(1) model. The study period is fixed at 1 unit (e.g., 1 yr), so increasing the frequency decreases the time between samples and increases serial correlation. A single observation has variance Var(x) = 1. R is the correlation between the first and last observations; the correlation between adjacent values is $R^{1/N}$, where $N + 1$ = sample size. The most efficient linear estimate of the mean is not the sample average, $x_.$, but the weighted average, $x_{w.}$ = $[\rho(x_0 + x_N) + (1 - \rho)(N + 1)x_.]/[2\rho + (1 - \rho)(N + 1)]$. The first and last observations are more informative than the rest, so increasing the sample can sometimes cause Var($x_.$) to increase. As $N \to \infty$, Var($x_.$) $\to 2[R - 1 - \log(R)]/[\log(R)]^2$; Var($x_{w.}$) $\to (4 - 2\log(R))/(2 - \log(R))^2$.

valid. For example, if we run repeated trials of our methods, using "Before" and "After" data from places that have not in fact been altered, our confidence intervals may contain the true "effect," zero, less often than claimed, unless observations are taken over longer periods than are usually available. Reducing sampling error or increasing sampling frequency will not prevent either. Methods to reduce temporal variation and correlation (especially long-term correlation) are often essential.

BACI: using comparison sites and covariates to reduce temporal variation and correlation

In experiments, covariates or concomitant variables "predict to some degree the . . . response . . . on the unit" (Snedecor and Cochran 1989:374; also Cochran 1957). The prediction is possible because some sources of natural variation affect both covariate and response. Ideally, using the covariate will remove this common variation from the prediction error. If treatments do not affect the covariate, then changes in the prediction can be used to estimate treatment effects; e.g., instead of (predicted value given treatment 1) − (predicted value given treatment 2), we can use (predicted value given treatment 1 and temperature = 15°) − (predicted value given treatment 2 and temperature = 15°), or the average of this difference over temperatures. These estimates will be more precise than estimates based on response values alone if the variation removed (that common to both the covariate and the response) is greater than the variation added (that affecting only the covariate).

In observational studies, covariates are ways to "remove the effects of disturbing variables" (Cochran 1957:262) or "adjust for sources of bias" (Snedecor and Cochran 1989:375). In effect, we compare predictions of the values that would have been obtained if the disturbing variables had the same values for all units.

Both goals apply to the disturbing effect of variation over time in impact assessment. Formally, this does not lead to bias in IA, because variance due to temporal fluctuations in the censused abundance is estimated and allowed for. From this viewpoint, the main aim is to get more accurate effect estimates by reducing this variance. On the other hand, if the autocorrelation is too strong, the full range of natural temporal variation will not be seen in a short Before or After time series, so its variance will be underestimated. It becomes a source of natural difference between Before and After that is not fully accounted for—i.e., a source of bias. From this viewpoint, the main aim is to get more reliable estimates of temporal variance, by reducing this autocorrelation. The BACI approach attempts to achieve both goals by using unaffected sites to help predict Impact values.

A simple comparison site model.—To motivate the simplest BACI method, the Group 2(a) analysis (see *Intervention analysis and BACI: Group 2. . . ,* above), imagine a region consisting of a patchwork of sites, one of which is the Impact site. Assume that the observed abundance at any site, S, and time, t, is the sum of (1) an average value determined by permanent features of the site, given by m_S; (2) systematic variation, $f(t)$, and broad-scale stochastic variation, $R(t)$, which are the same for all sites in a region surrounding the Impact site; (3) local stochastic natural variation, which is different at different sites and is described by $\varepsilon_S(t)$; (4) sampling error, also different at different sites and described by $\zeta_S(t)$; and (5) a constant alteration effect, Δ, which affects only the Impact site, only in the After period. Formally, we observe at the Impact site:

$$I_{\mathrm{B}i} = m_\mathrm{I} + f(t_{\mathrm{B}i}) + R(t_{\mathrm{B}i}) + \varepsilon_\mathrm{I}(t_{\mathrm{B}i}) + \zeta_\mathrm{I}(t_{\mathrm{B}i}) \quad \text{for } i = 1, 2, \ldots, T_\mathrm{B} \qquad (7)$$

$$I_{\mathrm{A}i} = m_\mathrm{I} + f(t_{\mathrm{A}i}) + R(t_{\mathrm{A}i}) + \varepsilon_\mathrm{I}(t_{\mathrm{A}i}) + \zeta_\mathrm{I}(t_{\mathrm{A}i}) + \Delta \quad \text{for } i = 1, 2, \ldots, T_\mathrm{A}. \qquad (8)$$

With only the $I_{\mathrm{P}i}$ as data, an intervention analysis approach would be to make parametric models of f, R, and ε, and fit them to predict the future values, $A_\mathrm{B}(t_f) = m_\mathrm{I} + f(t_f) + R(t_f) + \varepsilon_\mathrm{I}(t_f)$ and $A_\mathrm{A}(t_f) =$ the same $+ \Delta$, or just to estimate Δ directly. The results would be highly uncertain, because of the uncertain model forms for f and R, and because R may involve long-term temporal correlation stemming from major, region-wide, environmental variation.

Now suppose we also have observations on abundance at the same times from a comparison site, in the same region but far enough away not to be affected by the alteration. We obtain two more data sets, given by the same Eqs. as 7 and 8 except that "C" replaces "I" and "Δ" is absent. The difference between the Impact and comparison forms of Eq. 7 gives

$$I_{\mathrm{B}i} - C_{\mathrm{B}i} = m_\mathrm{I} - m_\mathrm{C} + \varepsilon_\mathrm{I}(t_{\mathrm{B}i}) - \varepsilon_\mathrm{C}(t_{\mathrm{B}i}) + \zeta_\mathrm{I}(t_{\mathrm{B}i}) - \zeta_\mathrm{C}(t_{\mathrm{B}i}) \quad \text{for } i = 1, 2, \ldots, T_\mathrm{B} \qquad (9)$$

and

$$I_{\mathrm{A}i} - C_{\mathrm{A}i} = m_\mathrm{I} - m_\mathrm{C} + \varepsilon_\mathrm{I}(t_{\mathrm{A}i}) - \varepsilon_\mathrm{C}(t_{\mathrm{A}i}) + \zeta_\mathrm{I}(t_{\mathrm{A}i}) - \zeta_\mathrm{C}(t_{\mathrm{A}i}) + \Delta \quad \text{for } i = 1, 2, \ldots, T_\mathrm{A}. \qquad (10)$$

The important change is the disappearance of R and f. In their places are additional errors from sampling and local temporal variation. The uncertainty in the estimate of Δ is reduced if these additional errors are smaller, less complicated, and less strongly correlated over time than were f and R. This may often be plausible. The sampling errors, ζ, are usually independent and can be reduced as discussed earlier. Local temporal perturbations may be smaller and fade faster than region-wide ones. In some cases they may fade fast enough for the correlation between successive differences to be negligible. Fig. 3 illustrates this: the original site abundances in Fig. 3a are strongly correlated but not seasonal; the Impact−Control differences of Fig. 3b are weakly correlated, and show the alteration effect more clearly.

If approximate independence of the differences over time is plausible, and satisfies tests and other data checks, then the analysis of Group 2(a) is reasonable, if simplistic. This could be based on means or more robust estimates, and perhaps should avoid assuming equal Before and After variances if the magnitudes of population fluctuations or sampling errors seem likely to vary naturally over time (see Stewart-Oaten et al. 1992). If the differences are correlated over time, as in Fig. 3, the analysis is messier, but prediction accuracy will still be much better than without the comparison site—and will itself be estimated more accurately—if this correlation is weak.

Covariates and "Controls" in assessment.—The comparison site "covariate" in the previous section has sometimes been called a "Control," e.g., by Green (1979) and Stewart-Oaten et al. (1986). The distinction is blurred. Controls and covariates both remove potentially confounding extraneous variation to highlight treatment effects. In the case of controls, this variation might be natural or a treatment artifact (e.g., an effect of caging or transplanting); without controls, it might be ignored, leading to bias. In the case of covariates, the variation is usually natural; without the covariates, it would still usually be allowed for, leading to greater explicit uncertainty. One distinction is that variation among Controls is often used to help estimate error, while variation among covariates rarely is. It usually makes no sense because the covariates measure different types of variables, e.g., rainfall and temperature.

The use of the term "Control" in assessment may be unfortunate. It could cause confusion with experimental controls, an error in the IVRS approach, which we discuss in that section (see *IVRS (impact vs. reference sites): Design-based justification. . . ,* below). It could also lead to reduced flexibility in analysis, by limiting the covariate relationship to models like Eqs. 9 and 10. This is simpler than the typical covariate relationship, although it is common when the covariate (C) is of the same type (abundance at a site) as the dependent variable. It is discussed in this connection, as an alternative to covariance, by Fisher (1960: chapter IX), Cox (1957, and 1958: chapter 4), and Snedecor and Cochran (1989: chapter 18).

Analysis 2(b) arises from the usual covariate relationship (Fig. 4a), the linear model

$$I_{\mathrm{B}i} = \beta C_{\mathrm{B}i} + \alpha + \text{error}$$

$$I_{\mathrm{A}i} = \beta C_{\mathrm{A}i} + \alpha + \Delta + \text{error} \qquad (11)$$

where β, α, and Δ are all unknown. This would correspond to Eqs. 7 and 8 if f and R varied among sites

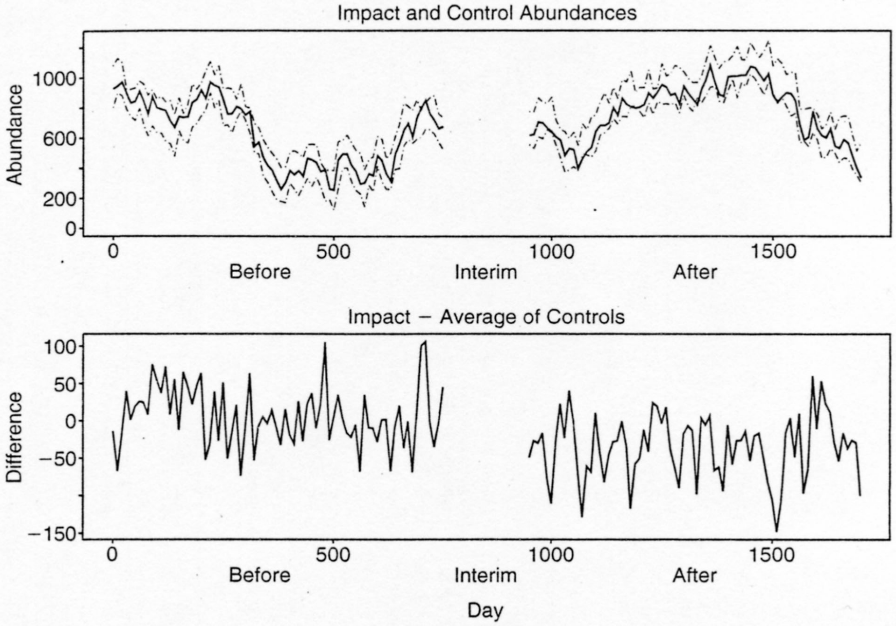


FIG. 3. Simulated example of the simplest BACI, Eqs. 7–10, but with no systematic variation, f (season or trend), and no sampling error, ζ. (a) Impact (solid line) = $580 + R(t) + \varepsilon_I(t) + \Delta$, where t is in days, $R(t)$ is given by Eq. (3) with $\rho = 0.995$, and $\mathrm{SD}\{a\} = 20$; $\varepsilon_I(t)$ is given by Eq. (3) with $\rho = 0.8$, and $\mathrm{SD}\{a\} = 20$; and $\Delta = 0$ in the Before period and $\Delta = -40$ in the After period. Control 1 = $680 + R(t) + \varepsilon_{C1}(t)$, where ε_{C1} is generated like ε_I but independently of it. Control 2 = $460 + R(t) + \varepsilon_{C2}(t)$. Plots record censuses taken every 10 days. (b) Difference between Impact and average of Controls. This line is the solid line in (a) minus the average of the broken lines in (a).

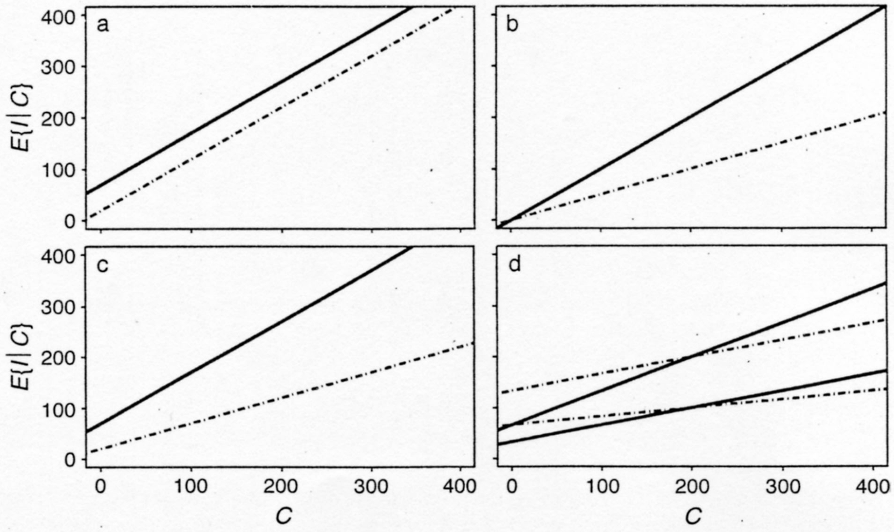


FIG. 4. Possible relationships between the observed value at control, C, and the conditional expected value at Impact, $E\{I \mid C\}$. For panels a–c, solid line = Before, broken line = After. (a) $E\{I \mid C\}$ vs. C based on Eq. 11 (cf. Mathur et al. 1980), with slope $\beta = 1$ and intercept $\alpha = 70$. The alteration reduces $E\{I \mid C\}$ by the same amount for each value of C, by reducing the intercept: $\Delta = -50$. (b) $E\{I \mid C\}$ vs. C based on Eq. 12. The effect is a reduction proportional to "natural" abundance: $\Delta = -0.5$, so the slope changes from $\beta = 1$ to $\beta + \Delta = 0.5$. (c) $E\{I \mid C\}$ vs. C based on Eq. 14. The alteration reduces both the slope ($\beta_B = 1$ to $\beta_A = 0.5$) and the intercept ($\alpha_B = 70$ to $\alpha_A = 20$); both act to lower abundance. (d) How measurement and process error influence the estimate of a purely multiplicative effect using the covariate model. Abundances are assumed to be generated by $I = \beta_P Q + \nu$ and $C = Q + \xi$, where Q, ν and ξ are independent random variables with means 200, 0, and 0, respectively. The alteration changes $\beta_B = 1$ to $\beta_A = 0.5$. The pairs of lines show the Before (higher line) and After relationships between $E\{I \mid C\}$ and C. Solid lines, $\sigma_\xi^2/\sigma_Q^2 = 0.5$; broken lines, $\sigma_\xi^2/\sigma_Q^2 = 2$. (Panel b shows $\sigma_\xi^2/\sigma_Q^2 = 0$.)

but with $f_S(t) = \beta_S f(t)$ and $R_S(t) = \beta_S R(t)$; e.g., the sites may respond to regional changes not identically but proportionately. As far as we know, Mathur et al. (1980) were the first to use this model in assessment, and to stress that the "control" is a predictor of the Impact site. They present their approach as an alternative to using ratios (cf. Eberhardt 1976), which corresponds to the covariate relationship

$$I_{Bi} = \beta C_{Bi} + \text{error}$$
$$I_{Ai} = (\beta + \Delta)C_{Ai} + \text{error} \qquad (12)$$

with error variances roughly proportional to the means (Fig. 4b).

Multiple covariates and nonlinear models.—As Mathur et al. (1980) remark, this approach can use multiple covariates, which need not be comparison-site abundances. They also use temperature and river flow, and comment that, with more than one control, they would have used the average or a multiple regression. A simple extension of Eqs. 9 and 10 is to replace C_{Pi} by $C_{\cdot Pi}$, the average over Control sites. An extension of the regression model of Eq. 11 is

$$I_{Bi} = \sum_j \beta_j X_{jBi} + \alpha + \text{error}$$
$$I_{Ai} = \sum_j \beta_j X_{jAi} + \alpha + \Delta + \text{error} \qquad (13)$$

where the X's include both multiple-comparison site abundances (or average abundances over groups of sites) and environmental variables like temperature. Other extensions are possible by transforming any of the variables, or by nonlinear regression relationships, though we suspect that real data will usually be too sparse, and errors too large, to justify complex models or allow numerous parameters to be estimated.

Some advantages of comparison sites as covariates.—Any variable unaffected by the alteration can be a covariate. Physical, chemical, or biological variables are attractive candidates. They directly affect abundances and may lead to a better understanding of mechanisms. However, they may require complicated models with time lags and many unknown parameters. Both problems also arise if too many covariates are used.

Comparison-site abundances have several potential advantages. They may reflect much of the natural variation of the physical, chemical, and biological factors affecting the Impact-site abundance. They can reflect only widespread variation, since they must be far enough away from the Impact site to be unaffected or negligibly affected by the alteration. But this may include much of the variation we most want to remove: large changes that reduce precision, and long-lasting fluctuations that threaten validity because the Before and After periods may be too short to observe their extent. When sources of variation (Table 5) are strong enough to cause large, long-lasting changes at a site, they may tend to be widespread. Recovery from a boom or crash, by immigration or dispersion, can be fast if it is local but slower if neighboring sites are similarly affected.

Another potential advantage is simplicity of the prediction formula. The same disturbance should affect similar, nearby sites in roughly the same way at roughly the same time. Simple (e.g., linear) models without time delays have a good chance of removing significant variation and correlation. This is important when the number of sampling times is small, since each parameter to be estimated adds uncertainty and reduces the degrees of freedom available to measure it.

Conditional estimates of non-additive effects: BACI problem statement 3.—Relatively simple functions, like Eqs. 11–13 with only a few C's or X's, suit the usually small number of sampling times. However, all these functions assume that the effect of the alteration is a constant. Instead, the effect could vary with season, current direction, and other environmental variables, or appear as a trend through some or all of the After period, reaching a maximum late in the period or after it.

Effect size is also likely to vary as regional abundance varies, both for natural reasons and because we will want to know effects in different scales—e.g., absolute effects for commercial reasons but proportional effects for understanding mechanisms or planning mitigation. Effects that are constant in one scale will be variable in others. This variation could be built into the covariate model, by allowing coefficients, as well as the intercept, to change for the After series. Eq. 12 already does this; the change at Impact is proportional to what the value would have been without the alteration—both are proportional to the abundance at the Control. More generally, Eq. 11 might become

$$I_{Bi} = \beta_B C_{Bi} + \alpha_B + \text{error}$$
$$I_{Ai} = \beta_A C_{Ai} + \alpha_A + \text{error} \qquad (14)$$

(Fig. 4c), to allow non-additive effects. Bence et al. (1996) considered this model in detail, and contrasted it with the model of Eqs. 9 and 10.

Such a model may change the target of the assessment. So far, it has been the difference between a future value and what it would have been without the alteration, or between the averages of these variables over some time periods. It may be useful to describe varying effects by their dependence on future conditions, especially when these are predictable, e.g., seasons. When the conditions cannot be well predicted, as when the effect appears to depend on the comparison-site value (as a proxy for overall abundance or for what the Impact abundance would have been), conditional descriptions may be preferable for some purposes and inferior for others.

When the effect varies, conditional estimates may be easier to produce. Eq. 14 estimates the "effect" of the alteration at time t by $(\hat{\beta}_A - \hat{\beta}_B)C(t) + \hat{\alpha}_A - \hat{\alpha}_B$, where the $\hat{\beta}$'s and $\hat{\alpha}$'s are estimates of the β's and α's and $C(t)$ is the future "Control" value.

Unconditional estimates may be harder, but still are often of central interest. A reliable estimate that an alteration will decrease a fish population by 40% when future abundances are high and 20% when future abundances are low would be useful, but we might also want to know how many fish are lost on average. We are led to:

Problem Statement 3: Estimate the difference between the means of functions $A_B(t)$ and $A_A(t)$ for the period between the end of the study and the time horizon, t_H, conditionally on other variables that seem important, with uncertainty measures. Use these results to estimate the unconditional difference.

The covariate and difference models assess varying effects differently. For the difference model we seek a transformation (scale) for which the effect is constant. For example, Eqs. 9 and 10 can handle a constant multiplicative effect by interpreting "I" and "C" as the logs of the abundances, so $\hat{\Delta}$ estimates $\log(I_{\text{with}}) - \log(I_{\text{without}})$. Effects on the untransformed scale are conditional on the Impact value. For example, for multiplicative effects, if we observe I in the After period then $Ie^{-\hat{\Delta}}$ is an estimate of the "no effect" value from which we can estimate the absolute change. Conversely, if a future "no effect" value is I, then the future "with effect" value can be estimated as $Ie^{\hat{\Delta}}$. Similar conversions can be made for other transformations, and confidence bounds can also be treated this way.

In contrast, the covariate model expresses effects in terms of future Control values, and can handle both effects that can be expressed as constant on some known scale and more complicated ones. For example, a multiplicative effect could be modeled by Eq. 12 or by Eq. 14. (We would prefer Eq. 14 because of the "errors in variables" problem described below). Eq. 14 can also deal with effects that cannot be expressed as constant, e.g., a change that is partly multiplicative and partly additive, and its "I" and "C" can be transformed as well. Its estimated future effects, $(\hat{\beta}_A - \hat{\beta}_B)C + \hat{\alpha}_A - \hat{\alpha}_B$, are in terms of future values of C (rather than I), which may be easier to predict since both Before and After observations can be used. Formal inferences may also be clearer, since error terms are handled more transparently: it is estimating a future observed Impact in terms of a future observed Control, while the additive model's $\hat{\Delta}$ is estimating the difference between the means of the logs of I_{with} and I_{without}.

"Errors in variables".—In Eq. 14 "C" is a future observed Control value: it includes both process error and sampling error. As for the "errors in variables" problem in regression (Fuller 1987), the values of the β's and α's would change if the variances of the errors changed—e.g., if the sampling error were reduced. The predicted "effect," $(\hat{\beta}_A - \hat{\beta}_B)C + \hat{\alpha}_A - \hat{\alpha}_B$, can be displayed by plotting past or typical Control values, and comparing the corresponding Before and After Impact predictions. However, the larger the error in "C," the smaller the absolute value of "$\hat{\beta}_A - \hat{\beta}_B$" is likely to be.

For a simplified example, suppose $I = \phi Q + \nu$, where ν = error and Q is the "true" Control abundance without either process error or sampling error; e.g., Q/ϕ satisfies Eq. 7 with ε and ζ both zero. A large Q then indicates a large I. But if $C = Q + \xi$, a large C may indicate either a large Q and hence I, or a large error, ξ, which says nothing about I. Thus, the regression of I on C, say $I = \eta C$ + error, will be flatter: $|\eta| < |\phi|$. The fractional reduction depends on the variances of Q and ξ, but as $\text{Var}\{\xi\} \to \infty$, $\eta \to 0$ because C tells nothing about Q and hence nothing about I.

The covariate method estimates η, not ϕ, since it fits I to C directly. It does the same for the generalization of Eq. 13 to

$$I_{Bi} = \sum_j \beta_{Bj} X_{jBi} + \alpha_B + \text{error}$$

$$I_{Ai} = \sum_j \beta_{Aj} X_{jAi} + \alpha_A + \text{error} \qquad (15)$$

where the alteration can affect any of the β's as well as α.

Converting the difference model, Eqs. 9 and 10, to the form of the covariate model given by Eq. 11 provides another example. The result is not Eq. 11 with slope (β) equal to 1. Formally, "$I - C = \alpha + \varepsilon$" can be written as $I = C + \alpha + \varepsilon$, but this is misleading because C and ε are both subject to chance, and are not independent. Assuming Normality, $E\{I \mid C\} = \lambda C + (1 - \lambda)\mu + \alpha$, where μ is the mean of Q (and may vary over time) and $\lambda = \sigma_Q^2/(\sigma_Q^2 + \sigma_\xi^2)$. That is, both the slope and intercept are altered by amounts depending on σ_ξ^2.

Interpretation of the difference and covariate models is similar if alteration effects are additive (in the scale of I and C in both models) and variances do not change between periods. For example, both Eq. 11 and Eqs. 9 and 10 can be used to estimate the effect, Δ. Although σ_ξ^2 alters both the intercept and slope of the relationship between I and C, these alterations are the same in both Before and After, and the effect is estimated by the change in intercept; in the example above, the Before relationship is $E\{I \mid C\} = \lambda C + \mu(1 - \lambda) + \alpha$ and the After relationship is $E\{I \mid C\} = \lambda C + \mu(1 - \lambda) + \alpha + \Delta$. The difference in interpretation is larger when effects depend on natural abundance. E.g., if $I = \beta_P Q + \nu$ and $C = Q + \xi$, where β_P is the multiplier for period P, then the predicted future Impact values with and without the alteration are $\beta_P \lambda C + \beta_P(1 - \lambda)\mu$, where μ and λ are as above. The purely multiplicative effect in the difference model "$I - \beta_P C = \varepsilon$" (or $\log(I) = \log(C) + \alpha_P + \varepsilon$) changes both the slope and the intercept in the derived covariate model (Fig. 4d). It is for this reason that we prefer Eq. 14 to Eq. 12 even when we believe natural variation and alteration effects are all entirely multiplicative.

Feasibility and model uncertainty.—The BACI approach requires comparison sites, and possibly other covariates, which can remove significant amounts of the variation in Impact-site abundances, especially variation with strong temporal correlation. We do not know how often such sites exist, or when. Are there particular life-history characteristics that make this approach more feasible? If so, how should the comparison sites be chosen? Data on abundances from multiple sites within a roughly homogeneous region, and extending over many years, would be helpful in exploring this.

We note that the reduction in extraneous temporal variance may not be obvious from the data in a particular case if the temporal correlation is strong or the Before and After periods are short. If the regional population fluctuates strongly but slowly, the full range of natural temporal variation may not appear during a Period. The error estimate from IA might be smaller than that from BACI, because the observed variation within a period is due mainly to sampling error and local, within-site fluctuations. The relevant error for the effect estimate, however, is over the full study period—Before, Interim, and After. The IA error estimate would be biased low. If the BACI model is correct, then its effect estimate is not contaminated by the regional variation. Thus its true error will be smaller than the IA error (assuming the regional fluctuation is larger than the local variation and sampling error), and is much more accurately estimated. However, this case also presents a potential problem for BACI: without much regional variation in the Before period, it will be difficult to find and fit a model that can accurately predict Impact-site values from Control values over the full, more variable, study period.

All solutions to the assessment problem depend on models. Problem statements 2 and 3 incorporate models by targeting the mean or some similar parameter, which has no meaning except in the context of a model. These models are not reality but represent a combination of reality and human ignorance of natural laws and initial conditions. Thus, in addition to model-dependent formal measures of the uncertainty of effect estimates, there is uncertainty about the accuracy of the model's representation of reality. Measuring model uncertainty formally is a difficult problem (Chatfield 1995, Draper 1995, Buckland et al. 1997, Mallows 1998), but repeating the assessment using several different plausible models will usually provide insight, especially when combined with measures of model fit (Burnham and Anderson 1992, 1998). Descriptive (e.g., graphical) methods for including these as part of the assessment results would be useful. We expect that, in many cases, acceptable models will give similar answers (e.g., Bence et al. 1996).

BACI: Response to Comments

In this section, we respond to some of the objections that have been raised to the BACI approach, and comment on some of the proposed corrections and improvements.

"Multiple Controls are needed"

Underwood (1992:147, 1994:152) asserts that the use of a single Control site arose "for reasons that are completely illogical," and that multiple Controls are needed to "solve problems caused by the lack of spatial replication." This "$N = 1$" criticism is wrong.

Variation among "Control" sites is irrelevant to the assessment problem, because the goal concerns a change at a particular nonrandom place. Intervention analysis, Group 1 (see *Intervention analysis and BACI: Group 1 . . .*, above), needs no control at all ("$N = 0$?"). It constructs, checks, and fits plausible models, and uses them to predict future values or long-term averages. This basic approach underlies almost all time-series analysis. The appropriate measure of uncertainty for IA (intervention analysis) estimates and predictions is the variation in prediction errors over time.

The same is true for BACI (Before–After, Control–Impact analysis). It is merely a special case of IA in which one or more unaffected sites are used as covariates, to reduce unexplained temporal variation and correlation. BACI Controls are not experimental controls. They are not chosen randomly nor to be independent of the Impact site, but deliberately chosen to be highly correlated with it so they will be useful covariates. Variation among them is no more used to estimate variances of the effect estimates than is the meaningless variation between such covariates as temperature and rainfall.

The IVRS (impact vs. reference sites) approach does use spatial variation to estimate error variance. This is usually invalid, as explained in the *IVRS* section, below. However, there are reasons for including multiple controls in some impact designs. We now discuss some of these.

Better prediction.—One possibility is to include multiple Controls as covariates in the analysis. Like other covariates, Controls can improve prediction, but are not guaranteed to do so. Additional Control sites are useful if they reflect different sources of variability acting on the Impact site, i.e., if they are highly correlated with the Impact site but less so with other Controls. For example, Controls might be selected on opposite sides of the Impact site. Too many predictors can spoil the prediction (e.g., Hocking 1976, Belsley et al. 1980:186–191, Miller 1990), especially if the predictors are highly correlated, as Controls are likely to be. To be useful, a Control must share common sources of temporal variation with the Impact site, but then multiple Controls are likely to share sources with each other.

Another approach is to combine multiple Controls into a few averages or medians representing different aspects of variability (Upcoast/Downcoast, Open/Sheltered, Sandy/Rocky) for the main analysis. This could reduce extraneous variability by averaging both local temporal variability and sampling error, but again is not guaranteed to improve prediction. A large number of comparison sites may have a large fraction of "bad" ones, which do not reflect the natural variation affecting Impact, and add extraneous variance of their own.

Scale and causality.—If a single Control is itself affected by the alteration, the effect at the Impact site will be underestimated or completely missed. A "gradient" design, the Group 2(d) analysis, with a series of sites at varying distances from the Impact site, could reveal the alteration's effect as a function of distance and direction. These sites would not be replicates, however. Their positions would be important for assessing scale and cause. The spatial pattern of estimated effects could be compared with that expected from the alteration. A match would increase confidence that the alteration was responsible.

Schroeter et al. (1993) used a similar approach, with a "Near Impact" site, a "Far Impact" site, and a "Control." Using the Control as a covariate, they showed that several related species declined at the Near Impact site after the alteration (a power plant). There were also smaller declines at the Far Impact site. These were evidence that the plant caused the changes, and also helped indicate their extent. Conversely, the BACI analysis showed an increase in the abundance of two species at the Near Impact site compared to the Control, although no mechanism involving the power plant seemed plausible and other explanations were available. These alternatives were supported when the BACI analysis showed a still greater increase at the Far Impact site compared to the Control.

Insurance.—The most obvious argument for multiple controls may be the strongest: something could go wrong with a single Control. Commercial harvesting, dumping of waste, or an accident occurring roughly between Before and After could make it useless for prediction. With several Controls, a site with known problems can be dropped.

If a problem is suspected but not known, then the average (or a robust estimate like the median) of a group of Controls could reduce its effect. Multiple Controls can also be used to check for such "confounding" disturbances. Note that a "pseudo-assessment," in which one Control plays the role of an "Impact" site and others the role of "Controls," should not indicate a significant change occurring at the time of the alteration (Carpenter et al. 1989).

"Different sites have different abundance paths"

Underwood argues that "there is . . . no reason to expect two sites to have the same time-course of changes in mean abundance" (1991:1575), "two arbitrarily chosen sites may very well differ in their changes through time, regardless of whether there has been" human interference (1992:146), and "[BACI cannot] detect impacts in populations that have spatial and temporal interactions in their abundances" (1994:5). These assertions do not distinguish between the censused abundance, $A(t)$, and its mean, $E\{A(t)\}$, which is defined by a time-series model. The standard ANOVA test for Time × Site interaction tests parallelism of the censused abundances, $A_I(t_{Bi})$ and $A_C(t_{Bi})$. BACI does not require these to follow parallel paths at Impact and Control. Each is expected to exhibit some variability that the other does not, from local natural temporal fluctuations and differences in responses to the same broad natural fluctuations.

It is the Impact and Control mean functions that are assumed to be parallel—and then only for Group 2(a) analysis and Eqs. 7–10 . This is no more discredited when two random paths are not parallel than is a coin's fairness when a single toss gives a tail. The problem described as "non-additivity" by Stewart-Oaten et al. (1986) arises in Group 2(a) analysis if the Before differences, $I_{Bi} - C_{Bi}$, are consistently larger when the sum is larger, or steadily decrease over time, or display other patterns that support plausible alternatives to assumptions about mean functions (e.g., parallelism) or errors (e.g., independence or a simple correlation structure). Checking for such patterns, and judging plausibility, are important parts of assessing model uncertainty.

If the model $I - C = \mu + \varepsilon$, underlying Analysis 2(a), is implausible or fits the data poorly, the abundances can be transformed, e.g., to logs, other covariates can be used, or μ can be replaced by $\mu + \beta\sin(\phi + 2\pi t)$ or another systematic function of time. Analyses 2(b) and 2(c) are still more flexible and may be more interpretable. In all these models, the distinction between "chance" and "systematic" variation is one of judgment and model convenience. The modeler can choose to regard all variation as due to "chance" (as do Box and Jenkins [1976], with seasonal variation), so the model cannot easily be discredited until the errors are modeled too.

No model is exact; we explicitly ignored slow "geological" changes and discussed the problems of modeling "chance." IA and BACI model "error" in terms of independent values from distributions representing sampling error and local perturbations. The model must assume enough of these values occur during each period (Before and After) for the variances of effect estimates to be reliably estimated. Study periods could be too short for this to be credible; consider that some types of perturbation (e.g., epidemics) may occur too rarely to be allowed for but have effects that last through the period. This possibility motivates BACI and other covariate methods (an epidemic is likely to hit neighboring sites similarly). Even for reasonable study periods, the special features that caused the Impact site to be chosen for the alteration may make it too different from nearby sites for BACI methods to help. Cycles (Table 5 source 4) and other local perturbations and interactions may have longer-lasting effects in closed populations (e.g., lakes) than in open ones. Special care is needed, e.g., in checking alternative explanations, if BACI methods are used in these situations.

But it seems extreme to claim none of these models is ever credible. It suggests that prediction is rarely possible in ecology, unless from studies longer than assessments, and that generalization over regions and habitats are less feasible still. Experiments using matched pairs of sites or halves of lakes seem pointless if there is no more reason to expect similarity from these pairs than from any other matching. Do any experiments have value if the results are uninformative about any other place or time? Efforts to determine how environments affect species may also be pointless if there is no sense in which, to the species, some pairs of sites are more "similar" than others.

Whether and when covariate sites can usefully reduce the uncertainty in effect estimates and improve its measurement, which are most likely to do so, what model forms are best, and the determination of appropriate study period length, are questions we regard as open. Whether such models have been valid and useful in the past might also be regarded as open, though we believe they have (e.g., Murdoch et al. 1989, Bence et al. 1996). To rule them out a priori seems unreasonable.

BACI "estimates effects on the mean only"

Underwood (1991:569) describes the aim of assessing changes on location parameters like means or medians as "oversimplistic and based on poor logic."

The logic.—The BACI section derives this aim from more basic ones. An "effect" is the difference between what will happen after the alteration and what would have happened without it. Predicting this difference can be almost the same task as estimating the difference between the means. Thus estimated effects on means can be justified as predictions of change in future values, or sets of future values.

Second, assessments are undertaken to help decision makers weigh biological effects with economic, social, aesthetic, and other environmental effects. BACI's predicted change in future values addresses this need. Weighing a 1% loss in local employment against a predicted 30% loss in fodder fish (or a 40% loss in summer and a 20% gain in winter) is hard, but this is because different kinds of effects do not have an agreed-on common currency. Weighing the employment loss against a 40% change in a parameter like variance is harder because such parameters represent more complicated concepts and their definition often directly involves our ignorance: Is it a bad effect of the alteration if we become less able to predict future values? Weighing the loss against a *P* value is yet more difficult.

Third, means are easier to estimate than other distributional summaries, and more is known about the properties of the estimates. In particular, problems of robust estimation, bias, and correlated data have received more attention in the context of estimating means, location parameters, and future values than in others.

A fourth point is that covariates can be helpful in the estimation of means, but are less so in the estimation of other parameters. In particular, Control sites may reduce the uncertainty in estimating a change for the Impact site mean but not for other parameters.

We do not claim that effects on other parameters are of no interest. An increase in the amplitude of population fluctuations might be important, and interpretable enough to use in decision making. Even here, however, it may be better to focus on estimating the mean. This need not be constant over time; e.g., it could be modeled as a periodic function. The "effect" could be a change in the amplitude from Before to After or the fraction of time the mean is below some critical level.

We illustrate these general points by considering two specific distributional parameters, extinction risk and variance.

Extinction risk.—Local impacts rarely threaten global extinction of a species, but local extinction is of concern, if unlikely to be quickly overcome by immigration. Thus a potentially useful goal is to predict whether local extinction will occur before the time horizon, with and without the alteration. Since there is almost always some chance of persistence, there seems no alternative to assessing extinction by computing its risk. However, extinction-risk estimates tend to be highly model sensitive. Extinction may depend on events that are observed too rarely for reliable estimates of their frequencies. Models that behave similarly when abundances are near mean levels may behave differently when abundances are extreme—conditions observed too rarely to guide model choice. Extinction risk is difficult to define as a parameter mainly describing nature rather than our level of ignorance. Estimating the change in it may be less useful than estimating the fraction of time the abundance will be below a threshold level.

The variance.—Underwood (1991) suggests assessing "the variance." This needs definition; as Table 5 shows, temporal variation of observed abundances or differences includes predictable and unpredictable trends, cycles, and irregular events, as well as sampling error. Some of these are better seen as part of the mean function, and estimated from it. Others may themselves vary in frequency or size over time, requiring frequent sampling for reliable estimation.

Variance results may be hard to interpret. Underwood (1991) argues that variance is an indicator of extinction risk, but we have seen that this risk is itself hard to interpret. Also, its relationship with variance is vague, controversial (McArdle et al. 1990), and unquantified, even when the "variance" is corrected for sampling error (Stewart-Oaten et al. 1995).

Inference for variances is more difficult than for means. It is so much affected by nonnormality that Moore and McCabe (1993:557) advise non-experts not to do it. The serial correlation, variable variances, and uncertain models of time series are even more serious problems than for means. It is also harder to use Control sites or other covariates to reduce these problems.

We do not claim that variance effects should be ignored but rather that they are hard to estimate reliably and interpret clearly, and that some aspects, like the amplitude of regular fluctuations, may be better estimated from models of the mean function.

A variance change proposal.—Underwood (1991) proposes a test to assess variance change. However, without unrealistic simplifying assumptions, the "variance" tested is hard to interpret and the test is likely to be invalid. We describe several versions of this test in the Appendix, but all are simple elaborations of the first, Analysis 6(a), a two-sided F test comparing Before and After temporal variation. The ratio is $s^2_{\mathrm{DA}}/s^2_{\mathrm{DB}}$ if there are Controls, and $s^2_{\mathrm{IA}}/s^2_{\mathrm{IB}}$ if not. Its validity and interpretation depend on the distributions (especially the means) of the s^2's (s^2_{DA}, etc.). These are functions of random vectors like $(I_{\mathrm{B1}}, I_{\mathrm{B2}}, \ldots, I_{\mathrm{B}T_{\mathrm{B}}})$, whose means, variances, and covariances can vary among components; e.g., I_{B1} and I_{B2} can have different means and variances, and cov$(I_{\mathrm{B1}}, I_{\mathrm{B2}})$ may not equal cov$(I_{\mathrm{B2}}, I_{\mathrm{B3}})$. Thus the s^2's estimate mixtures of systematic and chance temporal variation, sampling error, and temporal autocorrelation.

For example, the mean of s^2_{IB} is

$$E\{s^2_{\mathrm{IB}}\} = \sum (\mu_{\mathrm{B}i} - \mu_{\mathrm{B.}})^2/(T_{\mathrm{B}} - 1) + \sum [\sigma^2_{\mathrm{B}i} + \mathrm{var}\{I_{\mathrm{B}i}\}]/T_{\mathrm{B}} - 2 \sum \sum_{i<j} \mathrm{cov}\{I_{\mathrm{B}i}, I_{\mathrm{B}j}\}/T_{\mathrm{B}}(T_{\mathrm{B}} - 1). \quad (16)$$

The first term is the sample variance of the means of the abundance process at times t_{B1}, t_{B2}, etc. The second is the averages over these times of the process variances and the sampling-error variances. The third is the average of the covariance of censused abundances over pairs of times, $(t_{\mathrm{B}i}, t_{\mathrm{B}j})$.

The means of s^2_{DB} and s^2_{DA} are just as complex in general. They may be simpler if some BACI-type assumptions hold, but still mix temporal and sampling variance. The simplest form of BACI Analysis 2(a) assumes that the successive differences, $D_{\mathrm{B}i}$, have the same mean and negligible correlation. If so, then

$$E\{s^2_{\mathrm{DB}}\} = \sum [\sigma^2_{\mathrm{DB}i} + \mathrm{var}\{I_{\mathrm{B}i}\} + \mathrm{var}\{C_{\mathrm{B}i}\}]/T_{\mathrm{B}} \quad (17)$$

where $\sigma^2_{\mathrm{DB}i}$ is the variance of the difference between the Impact and Control censused abundances at time $t_{\mathrm{B}i}$ and the Var's are sampling-error variances.

Underwood's (1991) variance test compares an estimate of the right side of Eq. 16 or 17 with its "After" equivalent. Even if the test is valid, it is not clear what a change in these expressions would mean, or how much change would be cause for concern. Eq. 17 is the simpler, but an "effect" could arise if either temporal variation or sampling error is proportional to overall abundance (at both sites), which undergoes a natural Before–After change—the kind of long-lasting variation that BACI attempts to reduce. The validity of the test requires $s^2_{\mathrm{DA}}/s^2_{\mathrm{DB}}$ or $s^2_{\mathrm{IA}}/s^2_{\mathrm{IB}}$ to have an F distribution under the null hypothesis. In standard cases like ANOVA, this is derived from strong assumptions, like constant means, variances, and sampling errors, zero covariances, and Normality. When these do not hold, as is likely here, the F distribution may be far from the truth.

"Samples must be simultaneous"

The BACI design has been described as consisting of paired samples, "in the sense that the Control and Impact sites are sampled simultaneously (as near as possible)" (Stewart-Oaten et al. 1992:1397). This unnecessarily constrains the design, and leads Underwood (1991, 1994) to suggest that the design cannot be employed if logistics prevent simultaneous sampling.

The pairing is needed only to allow the Control value to help predict the Impact value. Sampling close in time will often do this best, but simultaneity is not needed. Sampling earlier at Control might even be better if organisms or environmental changes move mainly from the Control to the Impact site (e.g., down a river).

As an alternative, Underwood (1991, 1994) suggests sampling at times chosen randomly and independently at Impact and Control, for analysis by two-way ANOVA. This misses the point of collecting Control data in a BACI design: the reduction of temporal variation to get useful effect estimates and the reduction of autocorrelation to allow the variation to be measurable by data from within a period. Without these, we are no better off than in the Intervention Analysis case, where we have only Impact data.

"Sampling times should be random"

Underwood (1991) advocates sampling at random times. Stewart-Oaten et al. (1986) suggested this because a fixed sampling interval might lead to bias by coinciding with the period of a cycle, but they also noted drawbacks to random times. Here we argue that the statistical and logistical advantages of regular sampling times will almost always override the small chance of bias removal offered by random sampling times. Underwood (1991:577) attributes regular sampling times to "history, routine, fashion or lack of imagination" or (p. 571) to contractors' payment schedules. In our view, regular times have obvious logistical advantages and are statistically more efficient and easier to analyze.

For a fixed number of sampling times, equal spacing usually comes close to maximizing precision. Except for small "end effects" (cf. Fig. 2), it minimizes the variance of the estimate of the mean for series with variance constant and correlation a positive, convex decreasing function of the time gap (e.g., an exponential decrease with time)—both standard assumptions. Random times are usually inefficient in this sense, sampling some time periods intensively but others sparsely (see Bellhouse 1988). Equal spacing is often preferable for checking and fitting a model for a time-varying mean function with irregular cycles—least likely to miss some major feature, and most likely to give reliable function estimation by smoothing. For such reasons, much of the theory of time-series analysis is based on equally spaced times. For biological populations, "equal spacing" may need broad interpretation, as the rates and frequencies of processes like growth or disturbance may vary seasonally. One possibility is to view time as moving faster in some periods (like "degree days" in some population models), and to sample more frequently at these times, albeit still on a regular schedule.

The type of bias addressed by random sampling times may arise only rarely. It is limited mainly to high-frequency cycles whose period (or an integer multiple of it) is equal to the interval that regular sampling would use—usually less than one year. Except for diurnal cycles, this seems unlikely, especially for long-lived species which are often the focus of monitoring programs. Random times will reduce this bias by a factor of about $1/\sqrt{T_P}$ (on average, over randomizations) and give a more realistic variance estimate, but the estimates must take explicit account of the times actually chosen, and of their spacing: they would not usually be the "$\bar{X}$", "s^2" and "s^2/n" of random sampling. Random times cannot usually exchange "Before" and "After" (see Granelli et al. [1990], for an experiment of this kind), so bias due to a slow cycle (e.g., a single peak in the Before period and a single trough in the After) cannot be prevented this way. If regular samples would be taken at several fixed dates per year, the biasing cycle must shift phase, having its After peaks at about the times of its Before troughs. To adopt a complex, inefficient sampling program in case it reduces such an unlikely bias seems a mistake.

BACI "cannot assess causes"

Underwood claims BACI's "lack of replicated control sites provides insufficient evidence for an impact being due to the development" (1992:175; see also pp. 148 and 151; 1993:102, and 1996:152–154). His claims assume a single Control, which BACI does not, but apply to all BACI methods, since none of them treat Controls as replicates. IVRS methods do, so Underwood (e.g., 1992:176) claims they "provide better evidence of causal links". This claim assumes that assessments can be analyzed as experiments—e.g., "assessment of environmental impact will only become scientific when impacts are themselves treated as experiments" (Underwood 1992:176) and "It is a pity that previous planned environmental disturbances have not been evaluated properly as experiments" (Underwood 1993:111).

This is mistaken: real assessments cannot be treated as experiments. The reason is fundamental: the *Impact site is not randomly chosen*, either by humans or by Nature, either from a population or from the sites used in the study. We expand on this in the *IVRS* section, below. Random choice forms the foundation of experimental inference. Even when the units are not randomly selected from a larger population (e.g., most laboratory organisms), random assignment allows the probability of any given result arising in the absence of treatment effects to be calculated on the basis of the collection of all possible assignments that *could* have been obtained.

Lacking random choice of Impact site, an assessment is closer to an observational study than to an experiment. There are ways to strengthen causal inferences from observational studies (Jeffreys 1961, Hill 1965, Campbell 1969, Rubin 1974, Cook and Campbell 1979, Rosenbaum 1984, 1987, 1995, Cox 1992), but no assessment design or analysis can provide causal evidence as strong as an experiment could. There are at least three types of uncertainty associated with an effect estimate: the formal uncertainty calculated from a model, uncertainty about the model, and uncertainty about cause. In a sense, uncertainty about cause is already included in the other two; if our model were correct, it would allow for all sources of variation, so any difference not due to chance natural effects (whose size limits are given by our confidence interval) *must* be due to the alteration. But no model will be "correct" in this broad sense, e.g., allowing for such "natural" events as tsunamis or waste dumping. Most will estimate natural variability only from variation within the Before or After period, so will not allow for sources of variation that do not affect these data. Analysis of experiments can allow for such confounding factors, by taking the data as fixed and basing chance only on the randomized assignment. Nonrandom selection of the Impact site prevents this in assessment.

IVRS (IMPACT VS. REFERENCE SITES)

IVRS uses observations at the Impact site and several "Controls," at one or more times in the Before and After periods. In essence it summarizes all the data from a site into a single number (e.g., average After value − average Before value) and bases uncertainty measurement on variation among Control sites, rather than over time.

We first attempt an unambiguous formulation with an appropriate definition of an "effect." We then discuss four possible justifications for the IVRS method. Three are design-based, and, in the terms of our *Chance* section earlier, (see *Intervention analysis and BACI: Chance: Inferential "probability"*. . . , above) they assume device-based random site selection, "as if random" selection by Nature, and approximate "as if random" selection. The fourth is model-based, treating site values as a realization of a stochastic spatial process.

In brief, we conclude that the first two are invalid: device-based random site selection is known to be false and "as if random" selection for this problem would require all Controls to be identical to Impact. The third can justify rough "expert opinion" effect estimates but not formal inferences. The fourth leads to analyses more complex than the ANOVAs of Group 3. It shares features with time-series approaches such as IA and BACI, but it has stronger assumptions, less opportunity to check them, and less flexibility in modeling or effect description. However, it may offer a way to assess an impact without Before data, a situation where IA and BACI are unavailable.

Design-based justification for IVRS

IVRS problem statements.—The most detailed account of how IVRS is to be set up appears in Underwood (1992:152). We have added the markers "(A)" to "(E)" for reference.

"There should be a series of sites, randomiy chosen out of a set of possible sites that have similar features to those of the one where the development is being proposed. (A) The only constraint on random choice of sites is that the one planned to be impacted must be included in the sample. (B) This is not nearly as difficult as it seems; the sites do not have to be identical . . . (C) They simply have to follow the normal requirements that they come from a population of apparently similar sites. . . . (D) sites should be independently arranged so that there is no great spatial autocorrelation among them . . . sufficiently widely spaced that they are not correlated by processes of recruitment or disturbances. (E) The logic of the design is that an impact in one site should cause the mean abundance of animals there to change more than expected on average in undisturbed sites. . . . Impacts are those disturbances that cause mean abundance in a site to change more than is found on average."

Sentence (D) highlights a major difference between IVRS and BACI, whose aim is to choose sites that *are* correlated (with the Impact site and thus likely with each other) "by processes of recruitment or disturbances." Sentences (E) appear to regard any Impact change greater than the Control mean change as the "effect," but the chance of such a change is about 50% without any alteration at all, if all sites are random. The idea of randomly chosen sites suggests that the intention is to treat the observed Impact and Control sites as samples from "populations" of such sites, with the "effect" defined as a difference between the populations, and estimated by the corresponding difference between the samples. This is reinforced by other passages elsewhere (Underwood 1992, 1993, 1994). We are led to the following problem statement.

IVRS Problem Statement A: The observed Impact and Control sites are random samples from two hypothetical populations of sites, "Impact" and "Control," which would be identical without the alteration. The alteration's effect is the difference between the population means. Estimate it, with an uncertainty measure.

The "effect" could also be defined in terms of parameters other than means, e.g., variances, but we discuss only means here.

The "effect" in the problem statement does not correspond to the usual assessment target. Unlike most scientific studies, the assessment task concerns "a *particular* impact in a *particular* place from a *particular* facility" not the "average or 'usual' effect of an intervention over a large population of possible instances" (Stewart-Oaten et al. 1986:930). There is no "Impact population." Assessment is undertaken to provide information useful for decisions about a particular facility, e.g., to require or permit design or operation changes, impose penalties, etc. Knowing what this type of intervention or alteration would do somewhere else can be useful to help design a study in a particular place or support a judgment about what has happened there, but is otherwise irrelevant to these decisions. We therefore reformulate Problem Statement A to address the following assessment question.

IVRS Problem Statement B: The observed Impact and Control sites are randomly chosen from the same population of sites. The effect of the alteration is the difference between the Impact site abundance After the alteration and the abundance it would have had without the alteration. Estimate it, with an uncertainty measure.

The basis for inference.—As discussed in the "Chance" section (see *Intervention analysis and BACI: Chance: Inferential "probability"*. . . , above), design-based probabilities arise from the process by which units are chosen for study or assigned to treatments. Ideally, an artificial randomizing device makes the choices, but sometimes "as if random" selection by Nature is assumed. Either way, each possible sample or set of assignments should have an equal chance. Model-based inference does not base chances on the choice process. Instead, a chance model involving unpredictable natural events is assumed to determine the actual values of the units.

The BACI approach is model based: neither sites nor times are randomly chosen. Some random selection may be used in sampling within a given site at a given time, but this is not the basis of inference. The basis of IVRS quoted above seems to be design-based random site selection, although passage (D)'s effort to make sites independent, and comments elsewhere in

Underwood (1992, 1993, 1994) about temporal independence, suggest some model-based thinking.

The randomization argument.—Design-based inferences are based on imaginary repetitions of the unit (site) selection process, either random sampling or random assignment. Random sampling assumes units for each treatment are randomly chosen from the same large population; inferences compare the hypothetical populations that would be obtained by giving each member the same treatment. Random assignment refers only to the units actually used in the study, and bases probabilities on the assignments that could have been made. The former is equivalent to first randomly selecting all the experimental units, then randomly assigning them to treatments. In this sense, random assignment is a weaker assumption. Some authors (e.g., Kempthorne 1975:322) justify standard analyses like ANOVA only as approximations to randomization analyses, seeing the "random selection" assumption as "usually completely ludicrous." Others may accept this assumption as an approximation justifying the use of Normal theory, but concede that it is often "difficult to credit" (Mead 1988:230) so also view random assignment as necessary.

To illustrate the argument, suppose we have one Impact site and 39 Controls. Each site has a value, e.g., Before average − After average. The Control values are 2, 4, . . . , 76, 78, and the Impact value is 77. If the alteration has no effect at all, the chance that the Impact site has a value of 77 or more is the chance that the site chosen for the alteration was the "77" or "78" site, i.e., 2/40. This is the *P* value for a one-sided test of the null hypothesis of no effect. Doubling it gives a two-sided *P* value. A 95% confidence interval for the change would be all values of δ for which the null hypothesis "the alteration lowers abundance by δ" is accepted by a two-sided test. For example, if the null hypothesis is "the alteration lowers abundance by 6," and is true, then the Impact site would have been 71 without the alteration; the "true" site values were 2, 4, . . . , 78, and 71, so the chance of an Impact value of 77 or more was the chance of choosing one of the sites 71, 72, 74, 76, and 78, i.e., 5/40 (or 10/40 for the two-sided test), and the null hypothesis is accepted. Here, the 95% confidence interval is (−1, 75). The *P* value is 2/40 for $\delta = -1$ (i.e., the alteration causes an increase of 1) and 1/40 ("significant") for any smaller value; and is 2/40 for $\delta = 75$ (so the natural value of the Impact site was 2, equal smallest) and 1/40 for any larger value.

The only assumption made here is that each of the 40 sites had an equal chance to be "Impact." There are some subtleties. The probabilities treat the values as fixed: e.g., "$P\{\text{Impact} \geq 77\}$" is taken as "the probability that Impact is ≥77 given that the 40 site values are 2, 4, . . . , 77, 78." This is inappropriate if the alteration could increase sampling error without affecting the "true" value. Further complications arise in assessment experiments with multiple Impact sites, and in experiments generally; e.g., the null hypothesis may not include an effect that is positive at some sites, negative at others, but zero on average (the "unit-treatment additivity" assumption), and blocking, interactions, and covariates need care. These and other knotty points are explored by Kempthorne (1955), Cox (1958: chapter 2), Scheffé (1959:chapter 9), Edgington (1987), Welch (1990), Welch and Fahey (1994), and Manly (1997), among others.

The IVRS ANOVA Analysis 3(a) needs a further assumption, approximate Normality of (Impact − average of Controls). This is justifiable by an elaboration of the central-limit theorem in an experiment when all treatments (e.g., Impact) have large numbers of units, and in some cases under weaker assumptions. There seems no way to justify it for a few Controls and only one Impact site except by basing inferences on stronger assumptions, like random sampling from Normally distributed site values; in particular, the use of averages over time does not justify Normality unless the variation in these averages is due mainly to temporal error rather than site differences.

Random sites.—This randomization argument is valid in experiments (e.g., Lewis 1997), where assignment of sites to "treatment" or "control" is under the experimenter's control. It is invalid for the assessment problem because Impact sites for power plants, oil platforms, sewage outfalls, breakwaters, developments, etc., are virtually never chosen randomly. This is fatal, not a mere "constraint," as sentence (A) (see *IVRS problem statements*, above) asserts. The argument depends strictly on the process by which the sites were chosen and assigned to treatments: its "chances" (*P* values, etc.) are based entirely on imaginary repetitions of this process. "Apparently similar" (sentence C) has nothing to do with it. With site assignment by a repeatable chance process, the argument is valid even if the sites are scattered on land and sea all over the world, or if a different variable is measured at each site, or if some "sites" are not sites at all. Without the process, the "chance" in the argument is meaningless, no matter how similar the sites seem to be.

Thus IVRS is not an experiment but an observational study where the Impact site is compared to a group of sites the investigator thinks similar. BACI is an observational study too, but compares the Impact site to its "no alteration" self: the Controls merely help in this.

If the Impact site is not random, there is no population of sites from which it was randomly chosen. The investigator must subjectively specify a population of sites for "random" selection of Controls. This may not be easy. BACI involves only a few sites which can each be defined uniquely, but a "population" needs a generic definition or a long list. A crude marine example would be to take each point on a stretch of coast to represent the site obtained by going a kilometer upcoast, downcoast, and offshore. The numerous objections to this show how difficult the definition is likely to be.

Given the population, random selection is easy in principle (in the crude example, randomly select the representative points) but the "constraints" of sentence (D) above make it impossible in practice. True random choices cannot depend on the results of previous choices, but (D) requires that a choice too close to a previously chosen site must be rejected. Rare rejections, as for sampling without replacement, might not matter, or be allowed for in formulae, but no case has been made that they would be rare. Thus IVRS "controls" are nonrandom choices from a subjective population.

If an eccentric power company chooses its development site randomly from some collection of sites, a randomization test for an "effect" would be valid if it used the non-chosen sites as "controls," no matter how different they were. This is no guarantee of a good assessment. Unless the sites are similar, power will be low—the chance that the Impact site will be "extreme" with respect to a given species will be much the same with or without the alteration. Also, the multiple testing problem will be worse than usual: the Impact site is likely to be "extreme" with respect to at least one species, with or without the alteration.

Representative sites.—Nonrandom, subjective controls need not be worthless. Many medical studies compare study groups to such controls. The survival of a group of smokers might be compared to that of a group of non-smokers, or the level of factor X among patients with disease Y might be compared with the level among a reference group without it. These groups might contain all students in a school, or all patients in a given hospital: no choice at all, except that some students or patients might be excluded from the reference group as being (subjectively) too dissimilar (e.g., in age) from the study group. This method has obvious risks, despite refinements like multiple reference groups, stratification, matching (e.g., Mantel and Haenszel 1959) and meta-analyses, but it is clearly useful.

For these studies, sentences (B) and (C) are correct. The groups do not need to match each other exactly in every respect. They need only be "as if random" choices by Nature from populations with the same distributions of potentially confounding variables, those which are not affected by the treatment (smoking or disease Y) but are associated with the response (survival or factor X). But "only" is misleading. The need for "apparent similarity" with respect to these distributions is a major change from genuine random selection. Full satisfaction seems impossible. For example, a "significant" test for a difference in one of the distributions can cast doubt on the "as if random" assumption. The difference need not be in the mean: a confounding variable could have a nonlinear effect. There are usually many possible confounding variables and the important distribution may be joint, in two or more of them. It would be unusual if some of these did not show significant differences. There may also be other variables we do not know about. Thus the assumption of "as if random" sampling is always subject to doubt that is not reflected in formal uncertainty measures. Still, it can be a basis for useful inference when it is plausible and satisfies reasonable checks.

However, IVRS is qualitatively different because there is no "Impact population" other than the Impact site itself. The distribution of any potential confounding variable in this population allows only one value. The distribution in the Control population can match it only if all Control sites also have this value. Sentence (B) is now wrong: the sites do have to be identical for every potential confounding variable. Sentence (C)'s "apparently similar" must mean that any observed difference in one of these variables must be due only to within-site measurement error, not between-site variation. We think this is never credible.

Almost-representative sites.—Representative controls can be useful even if confounding variable distributions are not exactly identical. A medical study is not discredited when a test finds a significant difference in one of these, if the difference is thought unimportant. This involves subject-matter judgment about the confounding variable's likely effects, but can also be assessed statistically, by comparing between-group variation to within-group variation (Are the confounding-variable values for the two groups well interspersed?), or using the confounding variable as a covariate or to define strata. If it is unimportant, then estimates and even inferences (*P* values, standard errors, and confidence intervals for the effect or factor under study) can be regarded as good approximations.

IVRS results could be useful for the same reason, but the justification is more difficult. It is needed for all confounding variables, since all will have Control distributions that are different from the Impact distribution. It must also be mainly one of subjective judgment. A developer might claim, in an "After only" study, to have chosen an Impact site "well known" to be sparser than nearby "apparently similar" sites; or, in a "Before–After" study, that the Impact site was "well known" to cycle slowly and be due for a crash. Without other evidence, we might accept the first claim but not the second, based on our judgment and experience about the likelihood of such cases and the ability of casual observers to identify them. We can check that Impact values of quantitative confounding variables are within the ranges of Control values, but this does not account for possible interactions or nonlinear effects. Using confounding variables as covariates or for strata may provide only weak evidence, as they are likely to outnumber the sites.

The different types of variation in the Impact and Control "populations" makes bias checking hard and formal inference meaningless. The variation among sites in the Impact population is zero, while the Control variation depends on the investigator's subjective determination of "apparently similar." If another investigator determined it differently, obtaining a very different population, there is no statistical way to judge between them except range checking, which favors the larger population. Inferences about effects (P values, variances of effect estimates, confidence intervals, power calculations) will depend mainly on the number of "Control" sites and the variation among them, i.e., not on inherent natural variability but on the investigator's judgment about how close "apparently similar" should be.

An exception arises when effects are diffused among a variety of sites, as occurs with oil spills and other accidents. These may need some form of IVRS, since the Before data may be inadequate for IA or BACI. The affected region consists of many sites, of which only a sample is seen. An Impact "distribution" of a confounding variable is now nontrivial; in fact, the Impact region might be stratified by some suspected confounding variables, and sets of Control regions chosen to match them (Peterson et al. 2001). The "almost representative" justification could apply to these Controls, although it would not cover variables (e.g., currents) that caused a site to be affected (e.g., oiled).

Repeated measures.—Repeated-measures methods are applied to data taken on the same units at a series of times. Although samples of units can be compared by collapsing the data over time, to yield one value per unit, methods that retain the information in the separate sampling times can sometimes be more efficient (powerful), at the cost of assumptions about temporal correlation.

Assessment data, with Impact and Control samples of units (sites), may appear to fit this description. Green (1993) has suggested repeated-measures ANOVA (Crowder and Hand 1990) as a suitable method. This requires a "sphericity" assumption (roughly, correlations not to depend on the time gaps between observations) that is unlikely to hold, or an approximate F test with reduced degrees of freedom. However, there are also more general methods for "longitudinal data" (Diggle et al. 1994).

While both deterministic and stochastic variation within units (sites) is often modeled, the inferences on treatment effects in all these methods are design based. The units (sites) must be independent and randomly assigned to treatments—the "random sites" assumption discussed previously. The methods can also be applied to "as if random" treatment groups, but the assumption that confounding variables have the same distributions in the parent populations is still needed and still false in most assessments.

Longitudinal data methods are appropriate for analyzing an *experiment,* e.g., where a set of reefs is randomly divided into treatment (oil) and control groups, as may have been intended by Green (1993). Provided the oiling of one reef has no effect on others, the randomized assignment can justify ignoring the spatial correlation. In some cases the "as if random" argument might justify these methods for the diffuse impacts of the previous section.

Model-based justification for IVRS

Treating the IVRS "Controls" as quasi-random representative sites can justify little more than approximate effect estimates with subjective measures of uncertainty, yet something like IVRS may be the only choice when Before data suitable for BACI are unavailable. Glasby (1997) discusses this case (with multiple Impact sites) thoughtfully, although the above objections to IVRS apply also to his use of it. In this section, we sketch some aspects of using modeling to get more reliable conclusions.

Spatial models.—As before, the main idea is to compare the Impact site after the alteration to an estimate of what it would have been like without the alteration. The Controls are used to help estimate the latter. In the IVRS version, each observed site yields a single number, like the average of "After only" data, or After average − Before average. The effect of the alteration is the difference between the Impact value, which is affected by the alteration, and its prediction based on the Controls, which are not.

In spatial prediction and interpolation there is a value associated with every point on a map. We know the values at some points (Controls), and want to estimate the value at another point (the "no alteration" Impact value). The IVRS approach uses the average of the Control values as the prediction, with error estimated by their variance. This is a candidate, but it treats all sites as equally likely to be accurate. It ignores position as well as any special characteristics of the sites.

More often, the estimate is derived from a model including a chance term arising from natural processes and sampling error. Unlike design-based inference, imaginary repetition of the site-selection process plays no role in this chance term. Instead, the actual positions and other features of the sites are used as fixed terms in the model. The entire map's set of values, of which we observe some points, is treated as a "random function," one of a population of possible sets, with correlations between site values being functions of their positions (e.g., sites nearer in space are likely to be nearer in value). Under assumptions similar to those for time series, variances and covariances of site values can then be estimated from the data (Ripley 1981, Isaaks and Srivastava 1989, Cressie 1991, Aubry and Debouzie 2000).

There is a difference between interpolation and prediction like that between the censused abundance and the underlying process or between estimating a mean and predicting a future value. The interpolated value and the prediction are often the same, but the latter is less accurate because there is an extra layer of uncertainty in its target. Aubry and Debouzie (2000) explain the distinction as between conditional and unconditional estimation, and give ways to compute the uncertainties (variances). For assessment, prediction (unconditional estimation) is needed. The interpolation estimates the difference between the "alteration" and "no alteration" censused abundances at a particular time, or the average (or other summary) over the sampled times during the study period. This is itself only an estimate of the "true effect," which averages the difference over the period to the time horizon of interest, most or all of which is unobserved.

Selecting a reference group.—In model-based IVRS, as in BACI, the reference sites estimate what the Impact site would have been like. To do this, they should reflect the natural variation at the Impact site as closely as possible, without being affected by the alteration. Thus sentence (D) at the start of this section (see *Design-based justification for IVRS: IVRS problem statements*, above) is what we do *not* want in the model-based approach. Instead, we want unaffected sites that are near enough to experience the same conditions, and similar enough to respond to them the same way. For ocean species that recruit from the plankton, we want the recruitment process between Impact and Control sites to be as highly correlated as possible. The same applies to disturbances. There may be small alteration effects on the Control sites, e.g., if the Impact site contributes to their recruitment pool, or predators move to them when prey become scarce at the Impact site. Such "effects" are usually negligible.

"Similarity" of sites is hard to define, since many possible confounding variables could cause response to natural events to be different. In marine settings, depth, slope, aspect, relief, substrate, grain size, runoff, turbulence, current, and upwelling patterns are all candidates. Human use is often another. Some key variables may be hard to compare without thorough search, e.g., presence of major predators like lobsters or octopus.

Some problems in model-based IVRS.—Kriging and other spatial-prediction methods can be highly sensitive to model assumptions, which may be based more on tractability than realism. Most models assume stationarity and isotropy. The first means that patterns among sites depend only on their relative positions; e.g., the model correlation between any pair of sites depends only on the length and direction of the line joining them. Isotropy means the direction can also be ignored. These are often implausible for biological variables. Two sites in the same bay (or indentation) will often be more similar than sites in different bays, even if the latter are closer together. Along-coast and offshore variation often differ, perhaps too much for standard anisotropic models (e.g., Isaaks and Srivastava 1989). Journel (Isaaks and Srivastava 1989:xi) notes widely different conclusions drawn from the same data set by leading geostatisticans: "the illusion of objectivity can be maintained only by . . . scientific bullying in which laymen are dismissed as incapable of understanding the theory and . . . disqualified from questioning the universal expertise written into some cryptic software package."

This comment applies to other statistical methods, including time series. It is why avoiding statistical jargon and complex mathematics in discussing concepts is important scientifically, and why it is useful to compare effect estimates from several plausible models, in spatial prediction as well as in BACI. There is no necessary cause for alarm when different "experts" make different predictions, but there is cause when the experts' predictions (and the actual values being predicted) are far outside each other's allowances for uncertainty.

A variety of models can be produced by fine-tuning such details as functional forms for variograms. (These give the average squared difference between the values at pairs of sites distance *d* apart as a function of *d*.) However, we suspect that prediction errors will often stem more from differences between sites in their history (e.g., "founder effects") and special features (depth, substrate, etc.) than from the details of spatial models. Thus a useful form of multiple models is multiple sets of reference sites. We suggest two types. One uses stratified sets, i.e., sites are similar within sets but dissimilar between sets, based on potential confounding variables. Dissimilar predictions might suggest the use of covariates. The other is cross-validation, using roughly similar sets, e.g., chosen from the full collection of Controls by restricted randomization, to ensure a mixture of site types and some spatial interspersion. Predictions made by these sets should be within each-other's allowances for uncertainty. The predictions could also be treated as a sample to help establish the uncertainty allowance. A related way to check model assumptions is to use some Controls to predict others.

IA vs. IVRS: a tale of two series

We have shown that model-based IVRS can lead to meaningful inferences. It is sensitive to some strong assumptions, but IA and BACI also use models based on assumptions. Some direct comparison seems appropriate.

The series.—We consider the special case of IVRS where sites differ only in their along-coast positions—e.g., they are all the same distance offshore, covering the same range of depths. This removes the need for the isotropic assumption. If each site's observations are summarized in a single value, like After mean − Before mean, the data will have the appearance of a time series, with "time" being site position. The main difference is that the ordering of spatial position (e.g., increasing as we go upcoast from Impact) does not correspond to the order of cause and effect. IVRS compares the value at the Impact distance with the prediction based on the reference distances.

We compare this with IA, treating BACI as a special case of IA with covariates. IA compares the After times at the Impact site with predictions based on the Before times. IA and IVRS use different data, so represent alternative study designs rather than analyses. We compare the usefulness and reliability of results, but they also differ in cost and feasibility of data collection. In their minimal versions, IA needs only one site but samples it many times, while IVRS needs only one sampling time but samples many sites. IA is restricted by the length of the Before period, and IVRS by the distance from the Impact site at which potential "Controls" become uninformative.

Similarities and differences.—Both approaches require chance models, so that observations at each set of points (times or distances) have a specified joint distribution. Both use assumptions based on subject-matter consensus, simple approximate models of mechanisms, and mathematical convenience. However, there are differences. For IA, we have tried to describe the kinds of events and natural fluctuations that make up the "chance" and to give examples. These descriptions are used to establish the plausibility of simple models. This is harder for the distance series. A chance event occurring at a given place may affect sites on either side of it, while an event at a given time affects only later times, so building a model from a sequence of general cause–effect mechanisms is easier for a time series than for a spatial "series."

Abrupt changes, which separate the series into distinct parts, seem more common for distance series than for time series. Times have few distinguishing characteristics except for position, which can be adequately accounted for, often by the correlation alone and nearly always by correlation, a linear or quadratic trend, and one or two cyclic functions with known periods, like a day or a year. (Y2K may be an example of a "special" time.) Distances (sites) have many distinguishing features, such as aspect, substrate, etc. This makes stationarity assumptions more plausible for IA than for IVRS.

In particular, a time-series model that accurately describes a long-enough Before period at the Impact site should also accurately describe the After period, if there were no alteration. No brief Interim is likely to coincide with a natural division of the history of the Impact site into distinct periods and, while not chosen randomly, the Interim usually arises haphazardly from events in the planning, financing, permitting, or construction processes. But a model that describes the Control sites could easily fail to describe the Impact site, which is usually chosen carefully and may be unlike any other site. Similarly, all observable times are relevant for IA (with modifications for seasonal organisms); i.e., all can be incorporated into a single model with a small number of unknown parameters. Only sites satisfying guessed reference-set criteria are relevant for IVRS; these guesses could be wrong, and there may even be no site satisfying the right criteria except Impact itself.

Both approaches can use physical and chemical covariates to help predict what the Impact site would have been without the alteration. IA uses the Before times to construct and fit the prediction equations. IVRS can use the Control sites for this. Again, a model built by fitting the Before times is more likely to fit the After times than is a model built by fitting the Control sites to fit the Impact site. IA can also use other sites as covariates (BACI); there seems no simple analog for IVRS to use times as covariates.

Models can be checked against the data more easily and thoroughly in IA than in IVRS. Both can check assumptions by predicting some subsets of the data from others—IVRS by comparing subsets of unaltered sites, IA by comparing subsets of Before times. BACI can do both. These comparisons can include mock assessments, since there is known to be no alteration effect on either subset. IA can also use subsets of the After period to check assumptions or aid description of varying effects. IVRS cannot "subset" the Impact site. In part, IA has more model checks because it has more models to check, due to its greater flexibility, but this gives more realistic estimates of uncertainty, and the consistency of "effect" estimates across models is itself a check.

IA can estimate summaries of the After period, like the long-term mean, and compare them with the actual values after the alteration. These values have the advantages of averaging: smaller variation from both temporal correlation and sampling error, a more tractable distribution (e.g., Normal) for the errors, and perhaps rejection of outliers. IVRS compares the value at a single point, the Impact site, to its prediction from the Control points, so lacks these advantages. IA can also focus on means for given seasons or conditions, aiming at a more detailed description of an alteration effect that varies. IVRS could also be applied to values obtained from different subsets of Before and After times to describe a varying effect. However, IA is based on a model describing temporal variation, so its predictions (or estimates of the mean function) for future After times can use data from all times, with the usual gains in precision and tractability. IVRS must use sets of "similar" times in isolation, unless its model is elaborated to include temporal variation and correlation—i.e., unless it becomes more like IA and BACI.

Discussion

We have compared two approaches to impact assessment. One is Intervention Analysis (IA, Group 1), which compares Before and After time series, perhaps with the help of covariates. Before–After, Control–Impact (BACI, Group 2) is a special case of IA, where the covariates are similar unaffected sites. The other approach is impact vs. reference sites (IVRS, Group 3), which compares an Impact site value, such as the

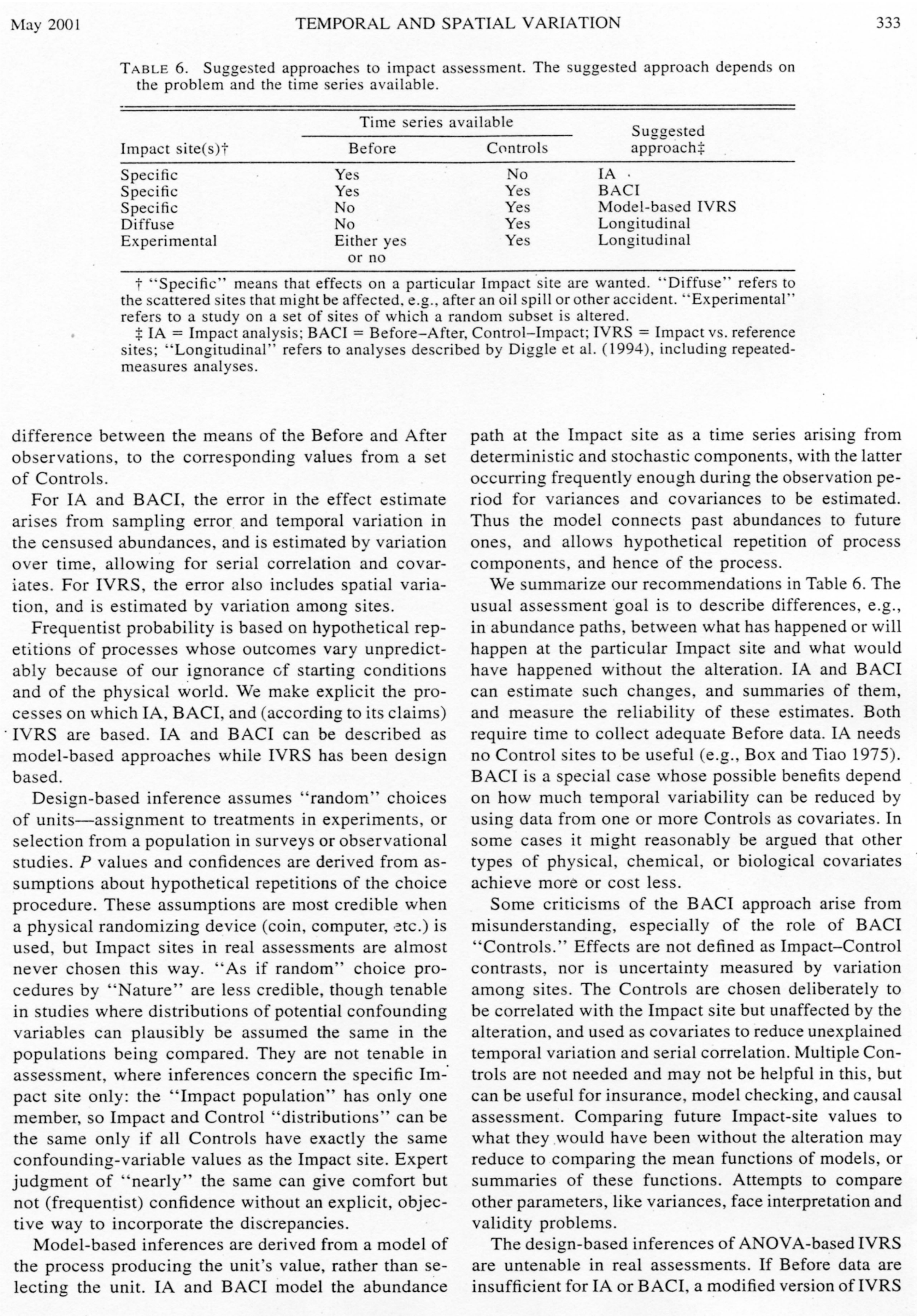


TABLE 6. Suggested approaches to impact assessment. The suggested approach depends on the problem and the time series available.

Impact site(s)†	Time series available Before	 Controls	Suggested approach‡
Specific	Yes	No	IA
Specific	Yes	Yes	BACI
Specific	No	Yes	Model-based IVRS
Diffuse	No	Yes	Longitudinal
Experimental	Either yes or no	Yes	Longitudinal

† "Specific" means that effects on a particular Impact site are wanted. "Diffuse" refers to the scattered sites that might be affected, e.g., after an oil spill or other accident. "Experimental" refers to a study on a set of sites of which a random subset is altered.

‡ IA = Impact analysis; BACI = Before–After, Control–Impact; IVRS = Impact vs. reference sites; "Longitudinal" refers to analyses described by Diggle et al. (1994), including repeated-measures analyses.

difference between the means of the Before and After observations, to the corresponding values from a set of Controls.

For IA and BACI, the error in the effect estimate arises from sampling error and temporal variation in the censused abundances, and is estimated by variation over time, allowing for serial correlation and covariates. For IVRS, the error also includes spatial variation, and is estimated by variation among sites.

Frequentist probability is based on hypothetical repetitions of processes whose outcomes vary unpredictably because of our ignorance of starting conditions and of the physical world. We make explicit the processes on which IA, BACI, and (according to its claims) IVRS are based. IA and BACI can be described as model-based approaches while IVRS has been design based.

Design-based inference assumes "random" choices of units—assignment to treatments in experiments, or selection from a population in surveys or observational studies. *P* values and confidences are derived from assumptions about hypothetical repetitions of the choice procedure. These assumptions are most credible when a physical randomizing device (coin, computer, etc.) is used, but Impact sites in real assessments are almost never chosen this way. "As if random" choice procedures by "Nature" are less credible, though tenable in studies where distributions of potential confounding variables can plausibly be assumed the same in the populations being compared. They are not tenable in assessment, where inferences concern the specific Impact site only: the "Impact population" has only one member, so Impact and Control "distributions" can be the same only if all Controls have exactly the same confounding-variable values as the Impact site. Expert judgment of "nearly" the same can give comfort but not (frequentist) confidence without an explicit, objective way to incorporate the discrepancies.

Model-based inferences are derived from a model of the process producing the unit's value, rather than selecting the unit. IA and BACI model the abundance path at the Impact site as a time series arising from deterministic and stochastic components, with the latter occurring frequently enough during the observation period for variances and covariances to be estimated. Thus the model connects past abundances to future ones, and allows hypothetical repetition of process components, and hence of the process.

We summarize our recommendations in Table 6. The usual assessment goal is to describe differences, e.g., in abundance paths, between what has happened or will happen at the particular Impact site and what would have happened without the alteration. IA and BACI can estimate such changes, and summaries of them, and measure the reliability of these estimates. Both require time to collect adequate Before data. IA needs no Control sites to be useful (e.g., Box and Tiao 1975). BACI is a special case whose possible benefits depend on how much temporal variability can be reduced by using data from one or more Controls as covariates. In some cases it might reasonably be argued that other types of physical, chemical, or biological covariates achieve more or cost less.

Some criticisms of the BACI approach arise from misunderstanding, especially of the role of BACI "Controls." Effects are not defined as Impact–Control contrasts, nor is uncertainty measured by variation among sites. The Controls are chosen deliberately to be correlated with the Impact site but unaffected by the alteration, and used as covariates to reduce unexplained temporal variation and serial correlation. Multiple Controls are not needed and may not be helpful in this, but can be useful for insurance, model checking, and causal assessment. Comparing future Impact-site values to what they would have been without the alteration may reduce to comparing the mean functions of models, or summaries of these functions. Attempts to compare other parameters, like variances, face interpretation and validity problems.

The design-based inferences of ANOVA-based IVRS are untenable in real assessments. If Before data are insufficient for IA or BACI, a modified version of IVRS could base inferences on a spatial model that uses the Controls to predict "no effect" values at the Impact site. The model and analysis are likely to need strong assumptions and be more complex than IVRS ANOVA. IVRS, especially longitudinal analyses, can be used in experiments where "impact" and "control" treatments are randomly assigned to sites. Similar analyses can be used when there is little Before data and impact sites are scattered and mixed with "controls;" but this assumes "as if random" sampling, so is misleading if the impact sites have been "selected" (e.g., in an oil spill) because of special features not shared with the controls.

Inferences in real assessments will be model based, and all models use assumptions. They can be checked against biological knowledge and the data, but no model is "correct" and several plausible models may give adequate fits. This adds uncertainty. Model uncertainty is not usually covered by formal inference and it is not clear how to quantify it (but see Burnham and Anderson 1998). It may be better kept separate, since the basis of quantification in formal inference (relative frequency in independent trials) may be unsuitable for model uncertainty. There may be ways to bound model uncertainty, e.g., it should be smaller if different models give similar answers than if only similar models do. But unambiguous definitions of "similarity" among models, or of whether a group of models "surrounds" the true model, remain a challenge.

We have avoided details of specific models, and advocate a flexible approach to modeling temporal and spatial variability. However, the data will rarely be sufficient to justify elaborate prediction functions, *f*, in models of the form Impact = *f*(Controls) + error. Models derived from simple mechanistic arguments, linear models with Controls combined into subgroups, or obtained from these by transforming all abundances (e.g., logs), may be all the data will bear. The same applies to models of the error; e.g., ARMA (autoregressive moving-average) models have simple structure, make intuitive sense, and are nested, allowing the strength and complexity of serial correlation to be examined, within the limits imposed by the length of the study.

A subtext of this paper is that mistakes arise from rote use of standard methods. ANOVA is a common villain. Rote use reduces efforts to assess assumptions, inquire where the "chance" comes from, check models, explore sensitivity, and interpret parameters and results. Rote methods can also answer the wrong question: ANOVA is often treated as only an *F* test for a "difference," but effect sizes and patterns may be needed, with measurements of reliability and model uncertainty (Stewart-Oaten 1996*b*).

Acknowledgments

We thank Bob Black, John Dixon, Bill Murdoch, Craig Osenberg, Steve Schroeter, and John Steinbeck for comments on earlier drafts. Comments and suggestions from Phil Dixon and two anonymous reviewers greatly increased clarity and focus, and provided valuable references. A. Stewart-Oaten's work was supported in part by the Minerals Management Service, U.S. Department of the Interior, under MMS Agreement Number 14-35-0001-30471 (The Southern California Educational Initiative). J. R. Bence's work was partially supported by the Michigan Department of Natural Resources and by the Michigan Sea Grant College Program. The views and interpretation contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either express or implied, of the U.S. Government.

Literature Cited

Arnold, L. 1974. Stochastic differential equations: theory and applications. John Wiley & Sons, New York, New York, USA.

Aubry, P., and D. Debouzie. 2000. Geostatistical estimation variance for the spatial mean in two-dimensional systematic sampling. Ecology **81**:543–553.

Bellhouse, D. R. 1988. Systematic sampling. Chapter 6 pages 125–145 *in* P. R. Krishnaiah and C. R. Rao, editors. Handbook of statistics. Volume 6. Elsevier, Amsterdam, The Netherlands.

Belsley, D. A., E. Kuh, and R. E. Welsch. 1980. Regression diagnostics: identifying influential data and sources of collinearity. John Wiley & Sons, New York, New York, USA.

Bence, J. R., A. Stewart-Oaten, and S. C. Schroeter. 1996. Estimating the size of an effect from a before–after-control–impact paired series design: the predictive approach applied to a power plant study. Pages 133–149 *in* R. J. Schmitt and C. W. Osenberg, editors. Detecting ecological impacts: concepts and applications in coastal habitats, Academic Press, San Diego, California, USA.

Box, G. E. P., and G. M. Jenkins. 1976. Time series analysis: forecasting and control. Holden-Day, San Francisco, California, USA.

Box, G. E. P., and G. C. Tiao. 1965. A change in level of a non-stationary time series. Biometrika **52**:181–192.

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.

Breiman, L. 1968. Probability. Addison-Wesley, Menlo Park, California, USA.

Breiman, L. 1995. Discussion of the paper by Chatfield (1995). Journal of the Royal Statistical Society **A 158**:455.

Buckland, S. T., K. P. Burnham, and N. H. Augustin. 1997. Model selection: an integral part of inference. Biometrics **53**:603–618.

Burnham, K. P., and D. R. Anderson. 1992. Data-based selection of an appropriate biological model: a key to modern data analysis. Pages 16–30 *in* D. R. McCullough and R. H. Barrett, editors. Wildlife 2001: populations. Elsevier, London, UK.

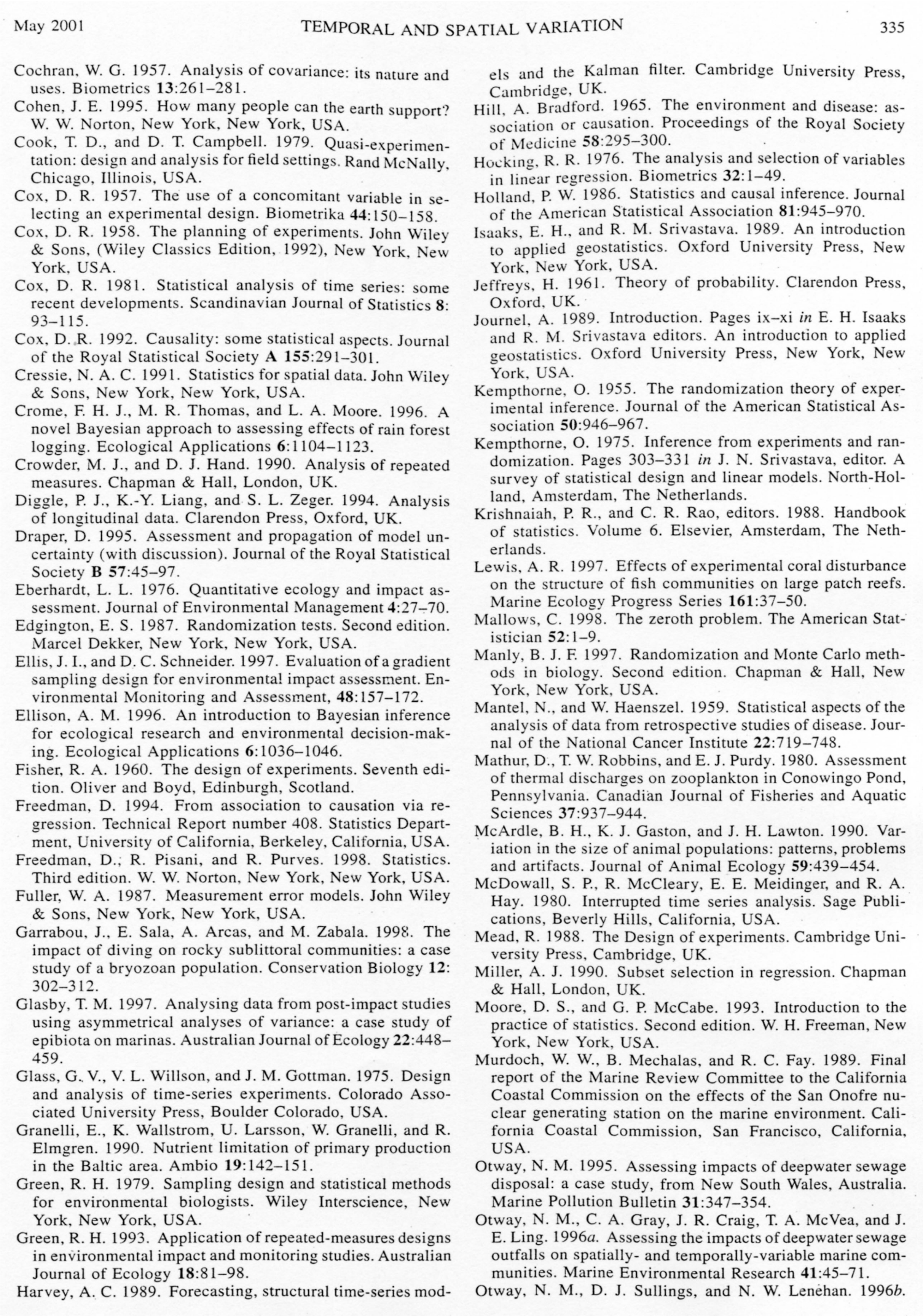
Burnham, K. P., and D. R. Anderson. 1998. Model selection and inference: a practical information theoretic approach. Springer-Verlag, New York, New York, USA.

Campbell, D. T. 1969. Prospective: artifact and control. Chapter 8 *in* R. Rosenthal and R. Rosnow, editors. Artifact and behavioral research. Academic Press, New York, New York, USA.

Campbell, D. T., and J. C. Stanley. 1966. Experimental and quasi-experimental designs for research. Rand McNally, Chicago, Illinois, USA.

Carpenter, S. R., T. M. Frost, D. Heisey, and T. K. Kratz. 1989. Randomized intervention analysis and the interpretation of whole-ecosystem experiments. Ecology **70**:1142–1152.

Chatfield, C. 1995. Model uncertainty, data mining, and statistical inference. Journal of the Royal Statistical Society, Series A **158**:419–466.



Cochran, W. G. 1957. Analysis of covariance: its nature and uses. Biometrics **13**:261–281.

Cohen, J. E. 1995. How many people can the earth support? W. W. Norton, New York, New York, USA.

Cook, T. D., and D. T. Campbell. 1979. Quasi-experimentation: design and analysis for field settings. Rand McNally, Chicago, Illinois, USA.

Cox, D. R. 1957. The use of a concomitant variable in selecting an experimental design. Biometrika **44**:150–158.

Cox, D. R. 1958. The planning of experiments. John Wiley & Sons, (Wiley Classics Edition, 1992), New York, New York, USA.

Cox, D. R. 1981. Statistical analysis of time series: some recent developments. Scandinavian Journal of Statistics **8**: 93–115.

Cox, D. R. 1992. Causality: some statistical aspects. Journal of the Royal Statistical Society **A 155**:291–301.

Cressie, N. A. C. 1991. Statistics for spatial data. John Wiley & Sons, New York, New York, USA.

Crome, F. H. J., M. R. Thomas, and L. A. Moore. 1996. A novel Bayesian approach to assessing effects of rain forest logging. Ecological Applications **6**:1104–1123.

Crowder, M. J., and D. J. Hand. 1990. Analysis of repeated measures. Chapman & Hall, London, UK.

Diggle, P. J., K.-Y. Liang, and S. L. Zeger. 1994. Analysis of longitudinal data. Clarendon Press, Oxford, UK.

Draper, D. 1995. Assessment and propagation of model uncertainty (with discussion). Journal of the Royal Statistical Society **B 57**:45–97.

Eberhardt, L. L. 1976. Quantitative ecology and impact assessment. Journal of Environmental Management **4**:27–70.

Edgington, E. S. 1987. Randomization tests. Second edition. Marcel Dekker, New York, New York, USA.

Ellis, J. I., and D. C. Schneider. 1997. Evaluation of a gradient sampling design for environmental impact assessment. Environmental Monitoring and Assessment, **48**:157–172.

Ellison, A. M. 1996. An introduction to Bayesian inference for ecological research and environmental decision-making. Ecological Applications **6**:1036–1046.

Fisher, R. A. 1960. The design of experiments. Seventh edition. Oliver and Boyd, Edinburgh, Scotland.

Freedman, D. 1994. From association to causation via regression. Technical Report number 408. Statistics Department, University of California, Berkeley, California, USA.

Freedman, D., R. Pisani, and R. Purves. 1998. Statistics. Third edition. W. W. Norton, New York, New York, USA.

Fuller, W. A. 1987. Measurement error models. John Wiley & Sons, New York, New York, USA.

Garrabou, J., E. Sala, A. Arcas, and M. Zabala. 1998. The impact of diving on rocky sublittoral communities: a case study of a bryozoan population. Conservation Biology **12**: 302–312.

Glasby, T. M. 1997. Analysing data from post-impact studies using asymmetrical analyses of variance: a case study of epibiota on marinas. Australian Journal of Ecology **22**:448–459.

Glass, G. V., V. L. Willson, and J. M. Gottman. 1975. Design and analysis of time-series experiments. Colorado Associated University Press, Boulder Colorado, USA.

Granelli, E., K. Wallstrom, U. Larsson, W. Granelli, and R. Elmgren. 1990. Nutrient limitation of primary production in the Baltic area. Ambio **19**:142–151.

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

Green, R. H. 1993. Application of repeated-measures designs in environmental impact and monitoring studies. Australian Journal of Ecology **18**:81–98.

Harvey, A. C. 1989. Forecasting, structural time-series models and the Kalman filter. Cambridge University Press, Cambridge, UK.

Hill, A. Bradford. 1965. The environment and disease: association or causation. Proceedings of the Royal Society of Medicine **58**:295–300.

Hocking, R. R. 1976. The analysis and selection of variables in linear regression. Biometrics **32**:1–49.

Holland, P. W. 1986. Statistics and causal inference. Journal of the American Statistical Association **81**:945–970.

Isaaks, E. H., and R. M. Srivastava. 1989. An introduction to applied geostatistics. Oxford University Press, New York, New York, USA.

Jeffreys, H. 1961. Theory of probability. Clarendon Press, Oxford, UK.

Journel, A. 1989. Introduction. Pages ix–xi *in* E. H. Isaaks and R. M. Srivastava editors. An introduction to applied geostatistics. Oxford University Press, New York, New York, USA.

Kempthorne, O. 1955. The randomization theory of experimental inference. Journal of the American Statistical Association **50**:946–967.

Kempthorne, O. 1975. Inference from experiments and randomization. Pages 303–331 *in* J. N. Srivastava, editor. A survey of statistical design and linear models. North-Holland, Amsterdam, The Netherlands.

Krishnaiah, P. R., and C. R. Rao, editors. 1988. Handbook of statistics. Volume 6. Elsevier, Amsterdam, The Netherlands.

Lewis, A. R. 1997. Effects of experimental coral disturbance on the structure of fish communities on large patch reefs. Marine Ecology Progress Series **161**:37–50.

Mallows, C. 1998. The zeroth problem. The American Statistician **52**:1–9.

Manly, B. J. F. 1997. Randomization and Monte Carlo methods in biology. Second edition. Chapman & Hall, New York, New York, USA.

Mantel, N., and W. Haenszel. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute **22**:719–748.

Mathur, D., T. W. Robbins, and E. J. Purdy. 1980. Assessment of thermal discharges on zooplankton in Conowingo Pond, Pennsylvania. Canadian Journal of Fisheries and Aquatic Sciences **37**:937–944.

McArdle, B. H., K. J. Gaston, and J. H. Lawton. 1990. Variation in the size of animal populations: patterns, problems and artifacts. Journal of Animal Ecology **59**:439–454.

McDowall, S. P., R. McCleary, E. E. Meidinger, and R. A. Hay. 1980. Interrupted time series analysis. Sage Publications, Beverly Hills, California, USA.

Mead, R. 1988. The Design of experiments. Cambridge University Press, Cambridge, UK.

Miller, A. J. 1990. Subset selection in regression. Chapman & Hall, London, UK.

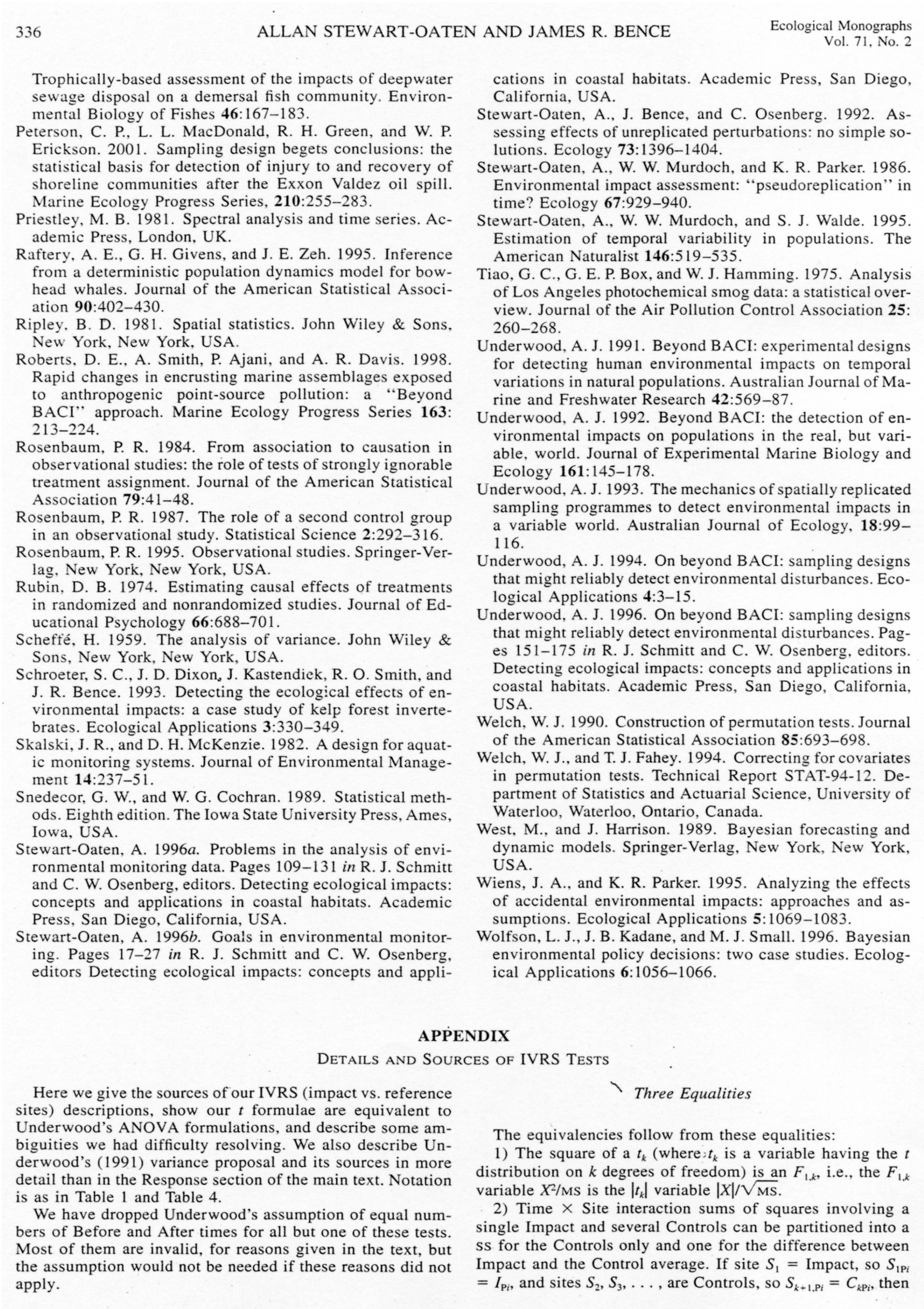
Moore, D. S., and G. P. McCabe. 1993. Introduction to the practice of statistics. Second edition. W. H. Freeman, New York, New York, USA.

Murdoch, W. W., B. Mechalas, and R. C. Fay. 1989. Final report of the Marine Review Committee to the California Coastal Commission on the effects of the San Onofre nuclear generating station on the marine environment. California Coastal Commission, San Francisco, California, USA.

Otway, N. M. 1995. Assessing impacts of deepwater sewage disposal: a case study, from New South Wales, Australia. Marine Pollution Bulletin **31**:347–354.

Otway, N. M., C. A. Gray, J. R. Craig, T. A. McVea, and J. E. Ling. 1996*a*. Assessing the impacts of deepwater sewage outfalls on spatially- and temporally-variable marine communities. Marine Environmental Research **41**:45–71.

Otway, N. M., D. J. Sullings, and N. W. Lenehan. 1996*b*.



Trophically-based assessment of the impacts of deepwater sewage disposal on a demersal fish community. Environmental Biology of Fishes **46**:167–183.

Peterson, C. P., L. L. MacDonald, R. H. Green, and W. P. Erickson. 2001. Sampling design begets conclusions: the statistical basis for detection of injury to and recovery of shoreline communities after the Exxon Valdez oil spill. Marine Ecology Progress Series, **210**:255–283.

Priestley, M. B. 1981. Spectral analysis and time series. Academic Press, London, UK.

Raftery, A. E., G. H. Givens, and J. E. Zeh. 1995. Inference from a deterministic population dynamics model for bowhead whales. Journal of the American Statistical Association **90**:402–430.

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

Roberts, D. E., A. Smith, P. Ajani, and A. R. Davis. 1998. Rapid changes in encrusting marine assemblages exposed to anthropogenic point-source pollution: a "Beyond BACI" approach. Marine Ecology Progress Series **163**: 213–224.

Rosenbaum, P. R. 1984. From association to causation in observational studies: the role of tests of strongly ignorable treatment assignment. Journal of the American Statistical Association **79**:41–48.

Rosenbaum, P. R. 1987. The role of a second control group in an observational study. Statistical Science **2**:292–316.

Rosenbaum, P. R. 1995. Observational studies. Springer-Verlag, New York, New York, USA.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology **66**:688–701.

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

Schroeter, S. C., J. D. Dixon, J. Kastendiek, R. O. Smith, and J. R. Bence. 1993. Detecting the ecological effects of environmental impacts: a case study of kelp forest invertebrates. Ecological Applications **3**:330–349.

Skalski, J. R., and D. H. McKenzie. 1982. A design for aquatic monitoring systems. Journal of Environmental Management **14**:237–51.

Snedecor, G. W., and W. G. Cochran. 1989. Statistical methods. Eighth edition. The Iowa State University Press, Ames, Iowa, USA.

Stewart-Oaten, A. 1996*a*. Problems in the analysis of environmental monitoring data. Pages 109–131 *in* R. J. Schmitt and C. W. Osenberg, editors. Detecting ecological impacts: concepts and applications in coastal habitats. Academic Press, San Diego, California, USA.

Stewart-Oaten, A. 1996*b*. Goals in environmental monitoring. Pages 17–27 *in* R. J. Schmitt and C. W. Osenberg, editors Detecting ecological impacts: concepts and applications in coastal habitats. Academic Press, San Diego, California, USA.

Stewart-Oaten, A., J. Bence, and C. Osenberg. 1992. Assessing effects of unreplicated perturbations: no simple solutions. Ecology **73**:1396–1404.

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

Stewart-Oaten, A., W. W. Murdoch, and S. J. Walde. 1995. Estimation of temporal variability in populations. The American Naturalist **146**:519–535.

Tiao, G. C., G. E. P. Box, and W. J. Hamming. 1975. Analysis of Los Angeles photochemical smog data: a statistical overview. Journal of the Air Pollution Control Association **25**: 260–268.

Underwood, A. J. 1991. Beyond BACI: experimental designs for detecting human environmental impacts on temporal variations in natural populations. Australian Journal of Marine and Freshwater Research **42**:569–87.

Underwood, A. J. 1992. Beyond BACI: the detection of environmental impacts on populations in the real, but variable, world. Journal of Experimental Marine Biology and Ecology **161**:145–178.

Underwood, A. J. 1993. The mechanics of spatially replicated sampling programmes to detect environmental impacts in a variable world. Australian Journal of Ecology, **18**:99–116.

Underwood, A. J. 1994. On beyond BACI: sampling designs that might reliably detect environmental disturbances. Ecological Applications **4**:3–15.

Underwood, A. J. 1996. On beyond BACI: sampling designs that might reliably detect environmental disturbances. Pages 151–175 *in* R. J. Schmitt and C. W. Osenberg, editors. Detecting ecological impacts: concepts and applications in coastal habitats. Academic Press, San Diego, California, USA.

Welch, W. J. 1990. Construction of permutation tests. Journal of the American Statistical Association **85**:693–698.

Welch, W. J., and T. J. Fahey. 1994. Correcting for covariates in permutation tests. Technical Report STAT-94-12. Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada.

West, M., and J. Harrison. 1989. Bayesian forecasting and dynamic models. Springer-Verlag, New York, New York, USA.

Wiens, J. A., and K. R. Parker. 1995. Analyzing the effects of accidental environmental impacts: approaches and assumptions. Ecological Applications **5**:1069–1083.

Wolfson, L. J., J. B. Kadane, and M. J. Small. 1996. Bayesian environmental policy decisions: two case studies. Ecological Applications **6**:1056–1066.

APPENDIX

Details and Sources of IVRS Tests

Here we give the sources of our IVRS (impact vs. reference sites) descriptions, show our t formulae are equivalent to Underwood's ANOVA formulations, and describe some ambiguities we had difficulty resolving. We also describe Underwood's (1991) variance proposal and its sources in more detail than in the Response section of the main text. Notation is as in Table 1 and Table 4.

We have dropped Underwood's assumption of equal numbers of Before and After times for all but one of these tests. Most of them are invalid, for reasons given in the text, but the assumption would not be needed if these reasons did not apply.

Three Equalities

The equivalencies follow from these equalities:

1) The square of a t_k (where t_k is a variable having the t distribution on k degrees of freedom) is an $F_{1,k}$, i.e., the $F_{1,k}$ variable X^2/MS is the $|t_k|$ variable $|X|/\sqrt{\text{MS}}$.

2) Time × Site interaction sums of squares involving a single Impact and several Controls can be partitioned into a SS for the Controls only and one for the difference between Impact and the Control average. If site S_1 = Impact, so $S_{1Pi} = I_{Pi}$, and sites $S_2, S_3, \ldots$, are Controls, so $S_{k+1,Pi} = C_{kPi}$, then

$$\sum_i \sum_k (S_{k\mathrm{B}i} - S_{k\mathrm{B}\cdot} - S_{\cdot\mathrm{B}i} + S_{\cdot\mathrm{B}\cdot})^2$$
$$= (T_\mathrm{B} - 1)s^2_\mathrm{DB}/(1 + 1/N_\mathrm{C})$$
$$+ (T_\mathrm{B} - 1)(N_\mathrm{C} - 1)s^2_\mathrm{CB}. \qquad \text{(A.1)}$$

Dividing each term on the right by its df (degrees of freedom) and taking the ratio gives

$$F_\mathrm{DC} = s^2_\mathrm{DB}/(1 + 1/N_\mathrm{C})s^2_\mathrm{CB}. \qquad \text{(A.2)}$$

With independence, equal variances, and Normality—each of which we dispute—this has the F distribution on $T_\mathrm{B} - 1$ and $(T_\mathrm{B} - 1)(N_\mathrm{C} - 1)$ df. The After period partitions similarly.

3) Period × Site interactions can also be partitioned. With the same notation for sites,

$$\sum_k \sum_\mathrm{P} (S_{k\mathrm{P}\cdot} - S_{k\cdot\cdot} - S_{\cdot\mathrm{P}\cdot} + S_{\cdot\cdot\cdot})^2$$
$$= f[(d_\mathrm{I} - d_{\mathrm{C}\cdot})^2/(1 + 1/N_\mathrm{C}) + \sum (d_{\mathrm{C}k} - d_{\mathrm{C}\cdot})^2] \qquad \text{(A.3)}$$

where $f = (T^2_\mathrm{B} + T^2_\mathrm{A})/(T_\mathrm{B} + T_\mathrm{A})^2$. The F ratio of the two parts of this partition is:

$$F_{d\mathrm{C}} = (d_\mathrm{I} - d_{\mathrm{C}\cdot})^2/(1 + 1/N_\mathrm{C})s^2_\mathrm{d} \qquad \text{(A.4)}$$

on 1 and $(N_\mathrm{C} - 1)$ df. This is the square of the t statistic $(d_\mathrm{I} - d_{\mathrm{C}\cdot})/s_\mathrm{d}\sqrt{(1 + 1/N_\mathrm{C})}$. It has the t distribution (and $F_{d\mathrm{C}}$ has the F) if the d's are independent and Normal, with equal variances—each of which we dispute.

Group 3. Changes in mean, using spatial variation for error

The following letters refer to the same letters in *Intervention analysis and BACI: ANOVA-based analysis: Group 3.*

a) The sources of our descriptions of the IVRS procedures are: for one time per period: Underwood (1992: Table I, footnote b); for multiple times per period: Underwood 1992: Table II, footnote e, and presumably Table V; 1993: Tables 5–8; and 1994: Table 4 and presumably Table 2). Underwood (1992: Table III) gives this test for different times (though equal numbers of them) at different sites.

In all cases, Underwood's "B × I" and "B × C" are converted to a t ratio as in Eqs. A.3 and A.4.

b) The source is Underwood (1992: Table VII). To describe this, we imagine an "Impact bay" and a set of "Control bays," with the Impact bay containing an Impact and a set of Control sites. The first level compares the Impact site to the Control sites in the Impact bay; the second compares the Impact bay to the Control bays.

For the first level, the source is Item 1b(2), line 3 (in Table VII). This is scrambled: it should read "MS B × S_1(Im) vs. O(Im)/MS B × O(Im)." These are the two terms of Eqs. A.3 and A.4 using only the sites in the Impact bay. This test is carried out only if

(1) $rs^2_\mathrm{d}/(1/T_\mathrm{B} + 1/T_\mathrm{A})s^2_\mathrm{R}$ is significant (Item 1b(2), line 2; see Group 4 (b), below),

(2) the variance tests in Group 6(c) 1: First level, below, are nonsignificant (Item 1a);

(3) $s^2_\mathrm{d}/(\Sigma_k\, s^2_k/N_\mathrm{E})$ is nonsignificant, where s^2_k = the variance of After−Before differences for the $N_\mathrm{C} + 1$ sites in Control bay k (= 1, 2, . . . , N_E, the number of these "external" bays). This is Item 1b(2), line 4. With $\mathrm{E}_{km\mathrm{B}i}$ = Before observation i at site m in Control bay k and $d_{km} = E_{km\mathrm{A}\cdot} - E_{km\mathrm{B}\cdot}$, it comes from partitioning SS "B × S(L)," which uses $\Sigma_k\, \Sigma_m\, (d_{km} - d_{k\cdot})^2 = \Sigma_k\, N_\mathrm{C} S^2_k$. "B × S(Im)" uses only the first term, corresponding to the Impact bay ($k = 0$), and "B × S(C)" uses the others, for the Control bays. "B × O(Im)" is the second term of the partitioning of "B × S(Im)" as in Eq. A.3.

For the second level the source is Item 2b(2), line 3: "B × I" and "B × C" come from "B × L" as in Eqs. A.3 and A.4. This is like the first-level test, but replaces $I_{\mathrm{P}i}$, the Impact site value, by $S_{\mathrm{P}i} = (I_{\mathrm{P}i} + \Sigma_\mathrm{m}\, C_{\mathrm{mP}i})/(N_\mathrm{C} + 1)$, the average over sites in the Impact bay, and the N_C Control site values, $\mathrm{C}_{k\mathrm{P}i}$, by the N_E Control bay averages, $E_{k\cdot\mathrm{P}i}$. The test is carried out only if all first-level tests and all variance tests in 6(c) below (item 2a) are nonsignificant, and $(N_\mathrm{C} + 1)rs^2_\mathrm{Ed}/(1/T_\mathrm{B} + 1/T_\mathrm{A})s^2_\mathrm{R}$ is significant (item 2b(2), line 2), where s^2_Ed is the sample variance of $\{d_{\mathrm{E}k} = E_{k\cdot\mathrm{A}\cdot} - E_{k\cdot\mathrm{B}\cdot}\}$ over Control bays.

c) The source is the "B × L" line of Table 1b of Underwood (1994: Table 1b).

d) "Repeated-measures" ANOVA is mentioned by Underwood (e.g., 1993: Table 1), and outlined by Green (1993).

Group 4. Changes in mean, using residual variation for error

a) For a single Before and single After time at a single Control, the source is Table 1b of Underwood (1991). The confidence interval can be written as $D_\mathrm{A} - D_\mathrm{B} + 2ts_\mathrm{R}/\sqrt{r}$, where the D's and s^2_R are given in Tables 1 and 4, with $T_\mathrm{B} = T_\mathrm{A} = N_\mathrm{C} = 1$. For multiple Controls, the sources are footnote b of Table I (with $T_\mathrm{B} = T_\mathrm{A} = 1$), footnote d of Table II, and "B × I" of Table V of Underwood (1992); and "B × I" of Tables 5–8 of Underwood (1993) and of Table 2 of Underwood (1994). The "between vs. within sites" restriction is "B × C vs. Residual" in the first two 1992 Tables (the second of which also requires nonsignificant tests for variance change), and may be intended in the other Tables also.

b) For the first level (the Impact site in the terminology used for Group 3(b) above), the source is Underwood (1992: Table VII: Item 1b(1)) and for the second level (bays), (1992: Table VII: Item 2b(1)). The numerators use the first term of Eq. A.3 in partitioning the sums of squares for

$$\text{“B × S(Im)”} = \sum \sum (U_{k\mathrm{P}\cdot} - U_{k\cdot\cdot} - U_{\cdot\mathrm{P}\cdot} + U_{\cdot\cdot\cdot})^2$$

where the U's are the site values in the Impact bay: $k = 1$ gives the Impact site; and

$$\text{“B × L”} = \sum \sum (V_{k\mathrm{P}\cdot} - V_{k\cdot\cdot} - V_{\cdot\mathrm{P}\cdot} + V_{\cdot\cdot\cdot})^2$$

where the V's are the bay averages. In the notation introduced in Group 3(b) above, $k = 1$ gives the Impact bay ($V_{1\mathrm{P}i} = S_{\mathrm{P}i}$), and the rest give the external bays ($V_{(k+1)\mathrm{P}i} = E_{k\cdot\mathrm{P}i}$, for k = 1, 2, . . . , N_E).

The F ratios are $r(d_\mathrm{I} - d_{\mathrm{C}\cdot})^2/(1/T_\mathrm{B} + 1/T_\mathrm{A})(1 + 1/N_\mathrm{C})s^2_\mathrm{R}$ for "SS B × O(Im)" and $r(N_\mathrm{C} + 1)(S_{\mathrm{A}\cdot} - S_{\mathrm{B}\cdot} - (E_{\cdot\cdot\mathrm{A}\cdot} - E_{\cdot\cdot\mathrm{B}\cdot}))^2/(1/T_\mathrm{B} + 1/T_\mathrm{A})(1 + 1/N_\mathrm{E})\mathrm{s}^2_\mathrm{R}$ for "SS B × I." These divide squared "effect" estimates, $d_\mathrm{I} - d_{\mathrm{C}\cdot}$ or $S_{\mathrm{A}\cdot} - S_{\mathrm{B}\cdot} - (E_{\cdot\cdot\mathrm{A}\cdot} - E_{\cdot\cdot\mathrm{B}\cdot})$, by their variances, $(1/T_\mathrm{B} + 1/T_\mathrm{A})(1 + 1/N_\mathrm{C})s^2_\mathrm{R}/r$ or $(1/T_\mathrm{B} + 1/T_\mathrm{A})(1 + 1/N_\mathrm{E})S^2_\mathrm{R}/r(N_\mathrm{C} + 1)$. For example, $S_{\mathrm{A}\cdot}$ averages T_A times, each the average of $N_\mathrm{C} + 1$ sites, each the average of r replicates, so $S_{\mathrm{A}\cdot}$ has variance $\sigma^2/rT_\mathrm{A}(N_\mathrm{C} + 1)$. $E_{\cdot\cdot\mathrm{A}\cdot}$ averages N_E such averages, so has variance $\sigma^2/rT_\mathrm{A}N_\mathrm{E}(N_\mathrm{C} + 1)$. The Before variances are similar, so the variance estimate of $S_{\mathrm{A}\cdot} - S_{\mathrm{B}\cdot} - (E_{\cdot\cdot\mathrm{A}\cdot} - E_{\cdot\cdot\mathrm{B}\cdot})$ is $(1/T_\mathrm{B} + 1/T_\mathrm{A})(1 + 1/N_\mathrm{E})s^2_\mathrm{R}/r(N_\mathrm{C} + 1)$.

The first-level test is carried out only if the first-level variance tests in Group 6(c) below, and the test of $rs^2_\mathrm{d}/(1/T_\mathrm{B} + 1/T_\mathrm{A})S^2_\mathrm{R}$ are nonsignificant. The latter is "MS B × O(Im)/MS Residual." Its numerator SS is the second term of the partition in Eq. A.3 for the interaction between periods and sites in the Impact bay. The second-level test is carried out only if all first-level tests, the second-level variance tests in Group 6(c), and $r(N_\mathrm{C} + 1)S^2_\mathrm{Ed}/S^2_\mathrm{R}(1/T_\mathrm{B} + 1/T_\mathrm{A})$ are all nonsignificant. The last uses "MS B × C/MS Residual:" its numerator SS is the second term of the partition in Eq. A.3 for the interaction between periods and bays.

Group 5. Changes in mean, using temporal variation for error

a) The sources are: for Impact site data only, Underwood (1991: Tables 1a and 3); for Impact and one Control, Underwood (1991: Table 1c, and 1994: Table 1a); and for Impact and several Controls, Underwood (1994: Table 3).

b) For a single Control, the source is Underwood (1991: Table 1d). The "B × L" MS is $(\Sigma\Sigma 1/T_{\mathrm{SP}})(I_{\mathrm{A}\cdot} - I_{\mathrm{B}\cdot} - C_{\mathrm{A}\cdot} + C_{\mathrm{B}\cdot})^2$ and the "Times (B × L)" MS is $[\Sigma\Sigma(I_{\mathrm{P}i} - I_{\mathrm{P}\cdot})^2 + \Sigma\Sigma(C_{1\mathrm{P}i} - C_{1\mathrm{P}\cdot})^2]/[\Sigma\Sigma T_{\mathrm{SP}} - 4]$, where $S = I$ or C and T_{SP} is the number of times site S ($= I$ or C) was sampled in period P ($=$ B or A). The T_{SP}'s do not need to be equal: the validity of the test depends mainly on the (unlikely) absence of serial correlation.

For multiple Controls, the source is Underwood (1992: Table III: note b). "B × I" is the first term in Eq. A.3. It may be that "*T*(B × C)" is a misprint for "*T*(B × L)." If so, s^2 becomes $\mathrm{s}_{\mathrm{L}}^2 = [\Sigma\, \Sigma\, (I_{\mathrm{P}i} - I_{\mathrm{P}\cdot})^2 + \Sigma\, \Sigma\, \Sigma\, (C_{k\mathrm{P}i} - C_{k\mathrm{P}\cdot})^2]/(\Sigma\, T_{\mathrm{IP}} + \Sigma\, \Sigma\, T_{k\mathrm{P}} - 2N_{\mathrm{C}} - 2)$, i.e., it uses the Impact site as well. Equal T_{IP}'s and $T_{k\mathrm{P}}$'s are unnecessary except for this alternative. The test is not carried out if $\mathrm{s}_{\mathrm{d}}^2/\mathrm{s}_{\mathrm{L}}^2 =$ "MS B × *C*/MS T(B × L)" is significant; Group 3(a) is used instead (Underwood 1992: Table III: note c). "SS B × C" is the second term in the partition of "SS B × L" as in Eq. A.3.

Group 6. Tests for changes in temporal variance

These tests do not appear in the *Analyses* section, but are discussed in the *BACI: Response to comments* section. Here we give each test or set of tests followed by its source.

a) A two-sided *F* test comparing Before and After temporal variation. The ratio is $s_{\mathrm{DA}}^2/s_{\mathrm{DB}}^2$ if there are Controls, and $s_{\mathrm{IA}}^2/s_{\mathrm{IB}}^2$ if not. With multiple Controls, "detection" may be declared only if such changes do not also occur among them: $s_{\mathrm{DA}}^2/(1 + 1/N_{\mathrm{C}})s_{\mathrm{CA}}^2$ is significant and $s_{\mathrm{CA}}^2/s_{\mathrm{CB}}^2$ is not, where $s_{\mathrm{CP}}^2 = \Sigma\, \Sigma\, (C_{\mathrm{kPi}} - C_{\mathrm{kP}\cdot} - C_{\cdot\mathrm{Pi}} + C_{\cdot\mathrm{P}\cdot})^2/(N_{\mathrm{C}} - 1)(T_{\mathrm{B}} - 1)$, the Time × Site interaction for Control sites in period P.

For Impact data only, the sources are Underwood (1991: Fig. 2c and Table 3: T(P Aft)/T(P Bef) (but with only one "Period"). For a single Control, the source is Underwood (1991: Table 1c: footnote B). "C × T(B)" should read "L × T(B)," and uses $\Sigma_{\mathrm{P}}\, \Sigma_i\, (\mathrm{I}_{\mathrm{P}i} - C_{\mathrm{P}i} - I_{\mathrm{P}\cdot} + C_{\mathrm{P}\cdot})^2 = \Sigma_{\mathrm{P}}\, \Sigma_i\, (D_{\mathrm{Pi}} - D_{\mathrm{P}\cdot})^2$. For multiple Controls, the sources are: Underwood (1992: Table II: footnote c(ii) and Table V (second last line)); Underwood (1993: Tables 5 and 6 (top, second left box)); and Underwood (1994: Table 4, second to last line). "T(Bef) × I" uses $\Sigma_i(I_{\mathrm{B}i} - C_{\cdot\mathrm{B}i} - I_{\mathrm{B}\cdot} + C_{\cdot\mathrm{B}\cdot})^2$. The requirements on $s_{\mathrm{DA}}^2/(1 + 1/N_{\mathrm{C}})s_{\mathrm{CA}}^2$ and $s_{\mathrm{CA}}^2/s_{\mathrm{CB}}^2$ are in Underwood (1992: Table II: footnotes c(i) and c(ii)) and Table V: last line, 1993: Tables 5 and 6 (top two left boxes) and 1994: Table 4: last and third last lines). "T(Aft) × I" and "T(Aft) × C" are obtained as in Eq. A.1.

b) An *F* test comparing temporal variation in the After period to sampling variance, s_{R}^2. If there are no Controls, the *F* ratio is $rs_{\mathrm{IA}}^2/s_{\mathrm{R}}^2$; otherwise it is $rs_{\mathrm{DA}}^2/(1 + 1/N_{\mathrm{C}})s_{\mathrm{R}}^2$. The first presumably "detects" an effect only if the test on $rs_{\mathrm{IB}}^2/s_{\mathrm{R}}^2$ was not significant. The second may also require one or more of (1) $rs_{\mathrm{DB}}^2/(1 + 1/N_{\mathrm{C}})s_{\mathrm{R}}^2$ and the Control ratios (2) $rs_{\mathrm{CA}}^2/s_{\mathrm{R}}^2$ and (3) $s_{\mathrm{CA}}^2/s_{\mathrm{CB}}^2$ to be not significant and possibly (4) $s_{\mathrm{DA}}^2/s_{\mathrm{DB}}^2$ to be significant.

The source for $rs_{\mathrm{IA}}^2/s_{\mathrm{R}}^2$ is Underwood 1991: Table 3 (the middle "Residual" entry under "*F* ratio versus" when there is only one "period"). The $rs_{\mathrm{IB}}^2/s_{\mathrm{R}}^2$ requirement is our guess. The sources for $rs_{\mathrm{DA}}^2/(1 + 1/N_{\mathrm{C}})s_{\mathrm{R}}^2$ are: for one Control site, Underwood (1991: Table 1c [the tests using "Residual"]); and for multiple Controls Underwood (1992: Table II: footnote b(i) and Table V, 1993: Tables 5 and 6, and 1994: Tables 2 and 3). The "T(Aft) × I" of these tables is related to s_{DA}^2 as the first term of the partition in Eq. A.1 is related to s_{DB}^2. For the restrictions (1)–(4), the sources are: for (1) possibly Underwood 1991: Table 1c; for (2), Underwood 1992: Table II: footnote b, and the *opposite* of (4) (presumably a slip: footnote b(ii) contradicts footnote c(ii)); for (1), (2), and possibly (4), Underwood 1992: Table V; for (2), (3), and (4) Underwood 1993: Tables 5 and (6); and for (1) and (2), possibly Underwood 1994: Tables 2 and 3. In these Tables, "T(Bef) × I" and "T(Bef) × C" partition "T(Bef) × L" as shown in Eqs. A.1 and A.2, and the "After" partitions are similar. Our "possibly" means that the tests are indicated in the Table but their connection with an Impact is not made explicit.

c) Variance tests like (a) and (b) above, but at two effect levels, like the analyses in Group 3(b) and Group 4(b). "Detection" is declared if one of several combinations of "significance" and "NS" are found. For the example of 3(b), an effect at the (within bay) Impact site is detected if $s_{\mathrm{DA}}^2/s_{\mathrm{DB}}^2$ and $rs_{\mathrm{DA}}^2/(1 + 1/N_{\mathrm{C}})s_{\mathrm{R}}^2$ are significant but $rs_{\mathrm{CA}}^2/s_{\mathrm{R}}^2$, $s_{\mathrm{CA}}^2/s_{\mathrm{CB}}^2$ and $s_{\mathrm{EA}}^2/s_{\mathrm{EB}}^2$ are not; or if $s_{\mathrm{DA}}^2/s_{\mathrm{DB}}^2$ and $s_{\mathrm{DA}}^2/(1 + 1/N_{\mathrm{C}})s_{\mathrm{CA}}^2$ are significant but $s_{\mathrm{CA}}^2/s_{\mathrm{CB}}^2$ and $s_{\mathrm{EA}}^2/s_{\mathrm{EB}}^2$ are not. Here, $s_{\mathrm{EA}}^2/s_{\mathrm{EB}}^2$ tests for a Before–After change in Time × Site interactions in the external bays: s_{EP}^2 is defined like s_{CP}^2 in Group 6(a), above, but with *E* (average over sites in an external bay) replacing *C*. Similar combinations are tested at the second level if no first-level changes in either mean (Group 3(b) and Group 4(b)) or variance are detected.

For the first level, the sources are Underwood (1992: Table VI and Table VII:Items 1a(1) and 1a(2). Underwood's "T(Bef) × S(C)" corresponds to our s_{EB}^2; his "*l*" is our $N_{\mathrm{E}} + 1$ and his "s" is our $N_{\mathrm{C}} + 1$; his degrees of freedom for s_{EB}^2 and s_{EA}^2 ("T(Bef) × S(C)" and "T(Aft) × S(C)") seem wrong: "*l*" should be "(l − 1)."

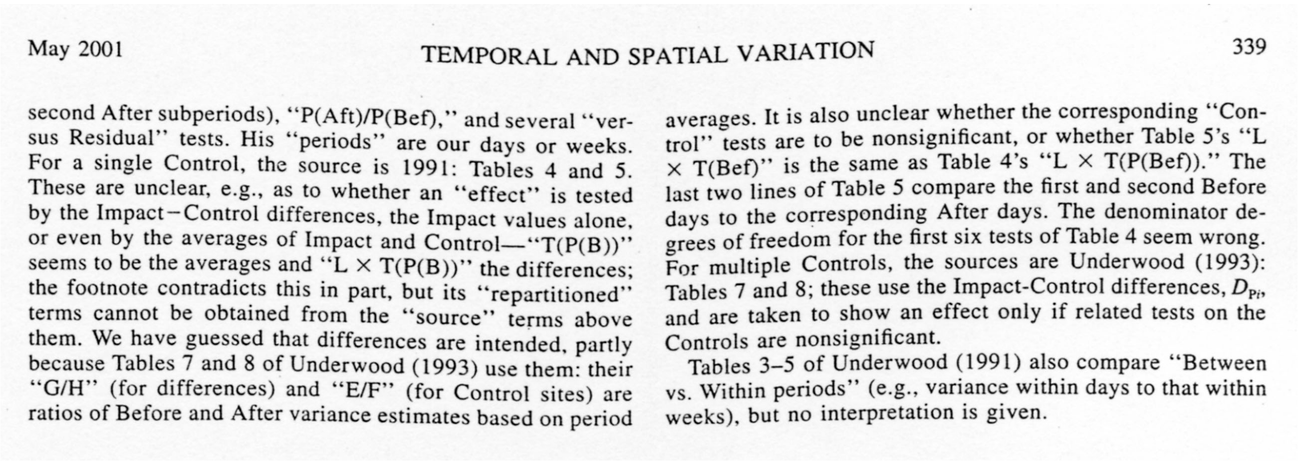
For the second level, using $I_{\mathrm{B}i}{}^*$, $E_{k\cdot\mathrm{B}i}$, etc., as for Group 3(b) in this *Appendix*, let $s_{2\mathrm{DB}}^2 = \Sigma\, (S_{\mathrm{B}i} - \mathrm{E}_{\cdot\cdot\mathrm{B}i})^2/(T_{\mathrm{B}} - 1)$, $s_{2\mathrm{CB}}^2 = \Sigma\, \Sigma\, (E_{k\cdot\mathrm{B}i} - E_{k\cdot\mathrm{B}\cdot} - E_{\cdot\cdot\mathrm{B}i} + E_{\cdot\cdot\mathrm{B}\cdot})^2/(N_{\mathrm{E}} - 1)(T_{\mathrm{B}} - 1)$, and $s_{2\mathrm{DA}}^2$ and $s_{2\mathrm{CA}}^2$ be similar. An effect is detected if $r(N_{\mathrm{C}} + 1)s_{2\mathrm{DA}}^2/(1 + 1/N_{\mathrm{E}})s_{\mathrm{R}}^2$ and $s_{2\mathrm{DA}}^2/s_{2\mathrm{DB}}^2$ are significant but $r(N_{\mathrm{C}} + 1)s_{2\mathrm{CA}}^2/s_{\mathrm{R}}^2$ and $s_{2\mathrm{CA}}^2/s_{2\mathrm{CB}}^2$ are not, or if $s_{2\mathrm{DA}}^2/s_{2\mathrm{DB}}^2$ and $s_{2\mathrm{DA}}^2/(1 + 1/N_{\mathrm{E}})s_{2\mathrm{CA}}^2$ are significant but $s_{2\mathrm{CA}}^2/s_{2\mathrm{CB}}^2$ is not. The sources are 1992: Table VI and Table VII: Items 2a(1) and 2a(2). Table VII says that "all" tests used at the first level of Control must be nonsignificant for the second to be studied, but it is unclear why this should apply to tests involving s_{R}^2. Underwood's "B × I" and "B × C" give our $s_{2\mathrm{DB}}^2$ and $s_{2\mathrm{CB}}^2$; "B × L" is partitioned as in Eqs. A.1 and A.2. For example, $s_{2\mathrm{DA}}^2$ is the sample variance of $S_{\mathrm{B}i} - E_{\cdot\cdot\mathrm{B}i}$; $S_{\mathrm{B}i}$ is the average of $N_{\mathrm{C}} + 1$ "site" values, each the average of *r* replicate samples; $E_{\cdot\cdot\mathrm{B}i}$ is the average of N_{E} values like $S_{\mathrm{B}i}$; thus the variance of $S_{\mathrm{B}i} - E_{\cdot\cdot\mathrm{B}i}$ is $\sigma^2(1 + 1/N_{\mathrm{E}})/r(N_{\mathrm{C}} + 1)$ if all observations are independent with variance σ^2.

d) N_{L} sites, all potentially affected. An *F* test using $s_{\mathrm{TL}}^2/s_{\mathrm{R}}^2$ with s_{TL}^2 as in Group 3(c). The motivation for this is unclear, since the numerator pools Before and After values, rather than contrasting them. Perhaps s_{TL}^2 is to be separated into its Before and After components, s_{BTL}^2 and s_{ATL}^2, with a Before–After variance change "detected" if $s_{\mathrm{ATL}}^2/s_{\mathrm{BTL}}^2$ is significant or if $s_{\mathrm{ATL}}^2/s_{\mathrm{R}}^2$ is significant but $s_{\mathrm{BTL}}^2/s_{\mathrm{R}}^2$ is not.

The source is Underwood 1994: Table 1b: the "T(B) × L" line.

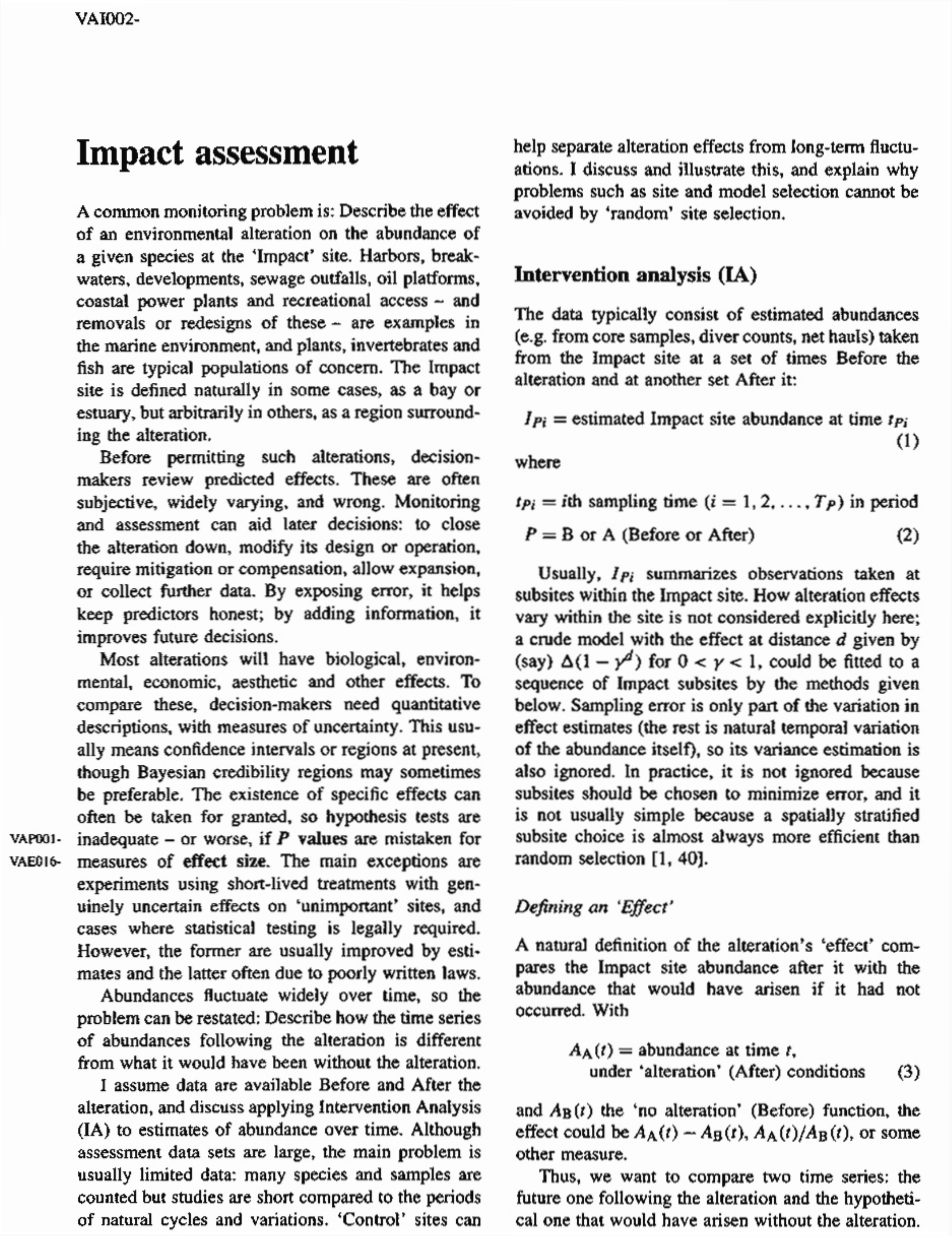
e) Multiple use of Group 6(a) and (b) when times within a period (Before or After) are not evenly spaced but are separated into subperiods. An example might be: (1) for each quarter (3-mo period), randomly choose four weeks; (2) for each chosen week, randomly choose three days; (3) for each chosen day, randomly choose four times of day. Then use 6(a) and 6(b) to test for changes in the within-day variance, the within-week variance (using the day averages), and the within-quarter variance (using the week averages). Some tests compare average Before and After variances, but others compare the *i*th Before day to the *i*th After day, or to the i_1th and i_2th After days. "Detection" may require nonsignificance of tests on the Controls or, when sampling error is used, of tests on the Before period. (We use "day" here for illustration: Underwood prescribes that samples be spaced sufficiently for "independence.") Some tests appear to compare "between subperiod" variance to "within subperiod," but no interpretation is given.

For "Impact only," the source is Underwood 1991: Table 3:"T(P Aft)/T(P Bef)," "T(P1 Aft)/T(P2 Aft)" (the first and



second After subperiods), "P(Aft)/P(Bef)," and several "versus Residual" tests. His "periods" are our days or weeks. For a single Control, the source is 1991: Tables 4 and 5. These are unclear, e.g., as to whether an "effect" is tested by the Impact−Control differences, the Impact values alone, or even by the averages of Impact and Control—"T(P(B))" seems to be the averages and "L × T(P(B))" the differences; the footnote contradicts this in part, but its "repartitioned" terms cannot be obtained from the "source" terms above them. We have guessed that differences are intended, partly because Tables 7 and 8 of Underwood (1993) use them: their "G/H" (for differences) and "E/F" (for Control sites) are ratios of Before and After variance estimates based on period averages. It is also unclear whether the corresponding "Control" tests are to be nonsignificant, or whether Table 5's "L × T(Bef)" is the same as Table 4's "L × T(P(Bef))." The last two lines of Table 5 compare the first and second Before days to the corresponding After days. The denominator degrees of freedom for the first six tests of Table 4 seem wrong. For multiple Controls, the sources are Underwood (1993): Tables 7 and 8; these use the Impact-Control differences, D_{Pi}, and are taken to show an effect only if related tests on the Controls are nonsignificant.

Tables 3–5 of Underwood (1991) also compare "Between vs. Within periods" (e.g., variance within days to that within weeks), but no interpretation is given.



VAI002-

Impact assessment

A common monitoring problem is: Describe the effect of an environmental alteration on the abundance of a given species at the 'Impact' site. Harbors, breakwaters, developments, sewage outfalls, oil platforms, coastal power plants and recreational access – and removals or redesigns of these – are examples in the marine environment, and plants, invertebrates and fish are typical populations of concern. The Impact site is defined naturally in some cases, as a bay or estuary, but arbitrarily in others, as a region surrounding the alteration.

Before permitting such alterations, decision-makers review predicted effects. These are often subjective, widely varying, and wrong. Monitoring and assessment can aid later decisions: to close the alteration down, modify its design or operation, require mitigation or compensation, allow expansion, or collect further data. By exposing error, it helps keep predictors honest; by adding information, it improves future decisions.

Most alterations will have biological, environmental, economic, aesthetic and other effects. To compare these, decision-makers need quantitative descriptions, with measures of uncertainty. This usually means confidence intervals or regions at present, though Bayesian credibility regions may sometimes be preferable. The existence of specific effects can often be taken for granted, so **hypothesis** tests are
VAP001- inadequate – or worse, if ***P* values** are mistaken for
VAE016- measures of **effect size**. The main exceptions are experiments using short-lived treatments with genuinely uncertain effects on 'unimportant' sites, and cases where statistical testing is legally required. However, the former are usually improved by estimates and the latter often due to poorly written laws.

Abundances fluctuate widely over time, so the problem can be restated: Describe how the time series of abundances following the alteration is different from what it would have been without the alteration.

I assume data are available Before and After the alteration, and discuss applying Intervention Analysis (IA) to estimates of abundance over time. Although assessment data sets are large, the main problem is usually limited data: many species and samples are counted but studies are short compared to the periods of natural cycles and variations. 'Control' sites can help separate alteration effects from long-term fluctuations. I discuss and illustrate this, and explain why problems such as site and model selection cannot be avoided by 'random' site selection.

Intervention analysis (IA)

The data typically consist of estimated abundances (e.g. from core samples, diver counts, net hauls) taken from the Impact site at a set of times Before the alteration and at another set After it:

$$I_{Pi} = \text{estimated Impact site abundance at time } t_{Pi} \tag{1}$$

where

$$t_{Pi} = i\text{th sampling time } (i = 1, 2, \ldots, T_P) \text{ in period}$$
$$P = \text{B or A (Before or After)} \tag{2}$$

Usually, I_{Pi} summarizes observations taken at subsites within the Impact site. How alteration effects vary within the site is not considered explicitly here; a crude model with the effect at distance d given by (say) $\Delta(1-\gamma^d)$ for $0 < \gamma < 1$, could be fitted to a sequence of Impact subsites by the methods given below. Sampling error is only part of the variation in effect estimates (the rest is natural temporal variation of the abundance itself), so its variance estimation is also ignored. In practice, it is not ignored because subsites should be chosen to minimize error, and it is not usually simple because a spatially stratified subsite choice is almost always more efficient than random selection [1, 40].

Defining an 'Effect'

A natural definition of the alteration's 'effect' compares the Impact site abundance after it with the abundance that would have arisen if it had not occurred. With

$$A_A(t) = \text{abundance at time } t, \quad \text{under 'alteration' (After) conditions} \tag{3}$$

and $A_B(t)$ the 'no alteration' (Before) function, the effect could be $A_A(t) - A_B(t)$, $A_A(t)/A_B(t)$, or some other measure.

Thus, we want to compare two time series: the future one following the alteration and the hypothetical one that would have arisen without the alteration.

2 Impact assessment

Only a limited future period is of interest, such as the expected lifetime of the alteration. Decisions will be based on summaries of these functions, such as the mean (calculated overall, or for particular seasons or conditions), not the full sets of continuous values. ('Chance of extinction' is also appealing but is hard to estimate or even define, because it reflects our ignorance as well as the population's fate.)

Models, Parameters and Estimates

The role of the data is to help predict $A_B(t)$ and $A_A(t)$ or the summaries over a future period. I assume this begins after the study ends: if it is within the study period, then we might use interpolation rather than prediction to guess $A_A(t)$.

Neither $\{I_{Bi}\}$ nor $\{I_{Ai}\}$ is from this period. Even if we knew the exact abundance continuously throughout the study, prediction of future values would be uncertain. To connect the observations to future values and provide a basis for calculating the uncertainty of predictions, we need to model abundance as the outcome of a process involving deterministic and stochastic parts. An example is

$$A_B(t) = f_B(t) + X(t) \qquad (4)$$

The deterministic part, $f_B(t)$, might be a seasonal sine wave

$$f_B(t) = \mu_B + \alpha_B \sin(2\pi t) + \beta_B \cos(2\pi t) \qquad (5)$$

if t is measured in years. The model for $A_A(t)$ could be similar with $\mu_A = \mu_B + \Delta$ or other parameter changes, or could allow for a gradual change to a new equilibrium, e.g. using a transition function such as $\mu_A(t) = \mu_B + \Delta(1 - \omega^t)$ for $0 < \omega < 1$.

Events causing 'chance' deviation from $f_B(t)$ can be brief (upwellings, predator visits), but still affect several sampling times: e.g. if $A_B(t) > f_B(t)$, then the additional population members will contribute to later abundances by survival and reproduction. Thus, the deviation at time t_i depends on past deviations and on chance events occurring, e.g.

$$X(t_i) = \sum_{j=1}^{p} \phi_j X(t_{i-j}) + U_i \qquad (6)$$

where the ϕs are constants to be estimated from the data. If the U_is are 'white noise' (uncorrelated with mean = 0 and equal variances), then this is an 'AR(p)' (pth order autoregressive (AR)) process [4]. However, part of U_i will be new effects of long-lasting events (storm runoff, epidemics, migrations, El Niño events) that have already affected previous sampling times. We might assume

$$U_i = V_i + \sum_{j=1}^{q} \theta_j V_{i-j} \qquad (7)$$

where the V_is are white noise. The U_is then form a 'MA(q)' (qth order moving average (MA)) process, and the X_is are then an autoregressive moving average (ARMA) 'ARMA(p, q)' process.

We do not observe $A_B = f_B + X$; we observe

$$I_{Bi} = A_B(t_{Bi}) + W(t_{Bi}) \qquad (8)$$

where W is sampling error. If W is white noise, independent of X, then $Z(t_i) = X(t_i) + W(t_i)$ forms an ARMA(p, Q) process, where $Q = \max(p, q)$ [4, p. 122]. A similar extension arises if the Ws are themselves an ARMA process.

If (4)–(8) are correct, then the assessment consists of estimating the parameters of f_B (5) and f_A, and comparing these estimates, or functions of them. The standard errors (SEs) of these comparisons depend on the parameters of X: $\sigma^2 = V\{V_i\}$, and the ϕs and θs. (Future values of $X(t)$ would be estimated for short-term forecasts of $A_B(t)$ and $A_A(t)$, but here the 'future period' of interest is much longer than the study.)

Maximum likelihood (ML) methods give consistent and approximately unbiased and normal estimates, and also estimate the covariance matrix of these estimates. Confidence intervals and tests can use the approximation (estimate − true value)/($\widehat{\text{SE}}$ of estimate) ~ N(0, 1) (or a t-distribution). For (5)–(7), ML estimates can be obtained iteratively: (i) estimate μ, α and β by ordinary least squares to obtain $\hat{f}_0$; (ii) use the residuals $I_{Bi} - \hat{f}_0(t_{Bi})$ to estimate the ARMA ϕs and θs; (iii) use the correlations implied by these estimates to re-estimate μ, α and β by generalized least squares to obtain $\hat{f}_1$; and (iv) repeat steps (ii) (using $\hat{f}_i$) and (iii) until estimates do not change.

See [5, 6], and [43] for IA, and [4, 7, 8, 13, 19, 50], and [70] for ARMA models.

Complications

These procedures were developed for long series with

VAI002-

Gaussian errors. The main results apply approximately to non-Gaussian distributions, and the algorithms are commonly applied in these cases. But virtually all theoretical results are asymptotic: Box and Jenkins [4, p. 33] recommend at least 50 observations. Impact assessment datasets are more likely to have 10–20 observations in a period, which is usually short relative to the persistence of the 'chance' effects listed below (6).

For short series, bias usually remains small, and may partly cancel out differences between Before and After estimates. Variance formulae and Normality are less reliable. One approach is computer-intensive Monte Carlo or bootstrap methods (e.g. [17, 22] and [39]. [*See* **Computer**
VAC042- **intensive methods**] Roughly, we fit a fully specified model, obtaining estimates $\hat{\mu}$, etc., $\hat{\phi}$s and $\hat{\theta}$s. For Monte Carlo, we use the fitted model to generate many random time series, apply our estimation methods to each, and use the error distribution of the resulting estimates as the basis for confidence intervals. For bootstrap, we use the fitted values of f (5) to estimate the Z_is following (8) by $\widehat{Z}_i$s. We use these and the $\hat{\phi}$s and $\hat{\theta}$s to estimate the iid innovations by $\{\widehat{V}_i\}$ (the first few $\widehat{V}_i$s cannot be estimated). We choose an initial set of V^*_js randomly from $\{\widehat{V}_i\}$, and simulate a subsequent ARMA process by plugging these, the $\hat{\phi}$s and $\hat{\theta}$s, and further random choices from $\{\widehat{V}_i\}$ into the defining (7) and (8). Davison and Hinkley [17, Chapter 8] suggest making this process longer than the actual series; the bootstrap sample uses only the last T_P terms, the others being used for 'burn-in' to approximate an equilibrium distribution for the first bootstrap value.

Further problems arise when (4)–(8) are implausible for observed abundances. ARMA models are flexible, but not all correlation structures are well approximated by (6) and (7) with low p and q. Sampling times may be unequally spaced. The V_is may vary more at some times of year than others. There may be a separate component of annual variation, e.g. for species with a short recruitment season. Errors may not be simply additive: e.g. (8) may be plausible while (4) is more plausible for $\log[A_B(t)]$.
VAS056- **State space models** [7, 18, 28, 33] are a promising approach in these cases, but may need still longer series.

Feasibility

Unreliable asymptotic distributions are not the only problem arising from the usually short series of observed I_{Pi}s.

'Chance' temporal fluctuations in abundance are high for many species, therefore predictions or estimates of model parameters may be too uncertain for decision-making. Increasing the sampling frequency can reduce uncertainty only to a limited degree because of temporal correlation.

Worse, this correlation, and thus the uncertainty itself, may be grossly underestimated. An El Niño event or disease could affect the whole of a short Before period, with little effect on the correlations among the I_{Bi}s. If the disease led to the sequence $\{D + I_{Bi}\}$ instead of $\{I_{Bi}\}$, the sample correlations would be unchanged. The chance of such a misleading event is greater if the Before and After periods are separated by a long 'Interim' period, e.g. for construction.

Before–After/Control Impact (BACI) – Using 'Control' Sites as Covariates VAB001-

Box and Tiao's [6] IA model includes exogenous variables other than the intervention itself. Such variables can reduce unexplained variation and may also reduce the role of long-term temporal correlation to negligible levels. Equation (5) makes time-of-year such a variable, and environmental variables such as temperature, salinity, nutrients, etc. might also play this role for abundance.

Abundances at other sites could be useful covariates if these sites are far enough away to be little affected by the alteration but near enough to experience the same major environmental fluctuations and similar enough to respond in the same way. Their relationship to Impact abundances may be simpler than those of other variables (e.g. no time delays), an important advantage for short series. Such sites can reflect only widespread variation, but this may be what we most want to remove: large, long-lasting fluctuations which threaten validity. References include [21, 57, 62], and [11] – a superb general account, presented in a social science context.

4 Impact assessment

Models

Observations C_{kPi} are taken at Control site k at the 'same' times (near enough for useful matching) as the I_{Pi}s are taken at Impact (1). The easiest model uses the differences, $I_{\mathrm{B}i} - C_{\bullet\mathrm{B}i}$, where '$\bullet$' indicates an average or other summary over the control sites. Suppose that the abundance at site S satisfies

$$S_{\mathrm{B}i} = \mu_{\mathrm{BS}} + f(t_{\mathrm{B}i}) + R(t_{\mathrm{B}i}) + \varepsilon_{\mathrm{S}}(t_{\mathrm{B}i}) + \zeta_{\mathrm{S}}(t_{\mathrm{B}i}) \quad \text{for } i = 1, 2, \ldots, T_{\mathrm{B}} \tag{9}$$

under Before (no alteration) conditions: μ_{BS} is the long-term mean at site S, f and R are fixed and random functions describing temporal variation in the region covering Impact and Control sites, ε_{S} is an ARMA process describing site S's deviation from the region, and ζ is sampling error. Then the Impact–Control difference satisfies

$$I_{\mathrm{B}i} - C_{\bullet\mathrm{B}i} = \mu_{\mathrm{BD}} + \varepsilon_{\mathrm{D}}(t_{\mathrm{B}i}) + \zeta_{\mathrm{D}}(t_{\mathrm{B}i}) \tag{10}$$

where the 'D' subscripts indicate differences, e.g. $\varepsilon_{\mathrm{D}} = \varepsilon_{\mathrm{I}} - \varepsilon_{\mathrm{C}\bullet}$. The hope is that the additional variation due to the $\varepsilon_{\mathrm{C}\bullet}$s and $\zeta_{\mathrm{C}\bullet}$s is less variable, less strongly correlated over time, and easier to model than the removed f and R. The variability added by the $\zeta_{\mathrm{C}\bullet}$s is often important but may be controllable by increasing sampling effort on each visit. The main assumption in (9) is that the regional variation, $f + R$, is additive. Transformation can allow other possibilities: for multiplicative variation, we can use $\log(I_{\mathrm{B}i}) - \log(C_{\bullet\mathrm{B}i})$ with an adjustment for zeros. A plot of the Before differences, $I_{\mathrm{B}i} - C_{\bullet\mathrm{B}i}$, against the sums helps examine this – a formal test for a nonzero slope corresponds to Tukey's [65] 'one degree of freedom' test [58, pp. 282–284].

Equation (10) allows estimation of the Before–After change in μ_{D}, as a constant or a function of time-of-year or other environmental variables. However, while it implies that the unconditional expectation, $\mathrm{E}\{I_{Pi} - C_{\bullet Pi}\}$, equals μ_{PD}, it does not imply that the conditional expectation, $\mathrm{E}\{I_{Pi}|C_{\bullet Pi}\}$, equals $C_{\bullet Pi} + \mu_{\mathrm{PD}}$. It does not allow us to compare estimates of what the Impact value would have been for a given Control value, under Before and After conditions.

To obtain such estimates, we need a model for $\mathrm{E}\{I_{Pi}|C_{\bullet Pi}\}$ or for $\mathrm{E}\{I_{Pi}|C_{\bullet Pi}$ and other predictors$\}$. This suggests the regression model

$$I_{\mathrm{B}i} = \alpha_{\mathrm{B}} + \beta_{\mathrm{B}} C_{\bullet\mathrm{B}i} + \varepsilon(t_{\mathrm{B}i}) \tag{11}$$

where ε is an ARMA process uncorrelated with $C_{\bullet}$ and accounting for both sampling error and fluctuations in the 'true' Impact and Control abundances. Equation (11) can be used to: (a) estimate effects under conditions of regional abundance or scarcity, as measured by $C_{\bullet}$; (b) estimate an average change, by $(\hat{}_{\mathrm{A}} - \hat{}_{\mathrm{B}})C + \hat{\alpha}_{\mathrm{A}} - \hat{\alpha}_{\mathrm{B}}$, where C is the average $C_{\bullet}$ value over both periods; (c) improve rough estimation of change on different scales (change in A, $\log(A)$, etc.) by using the average of $g(\hat{}\{I_{\mathrm{A}i}|C\}) - g(\hat{}\{I_{\mathrm{B}i}|C\})$ over all observed values of $C_{\bullet Pi}$ to estimate the average change in $g(A)$.

Unmeasured Uncertainty

The uncertainty expressed in SEs, confidence intervals, etc. does not include uncertainty about the model.

Equations (4)–(11) could apply to the abundances, I_{Pi} etc. or to their logs or other transformations. Rather than the sine wave of (5), seasonal variation could be treated by letting $Z_t = \gamma_t + R_t$, where γ_t is the mean of months corresponding to t and the differenced residuals, $R_t - R_{t-12}$ (for t in months), follow an ARMA model. This is more flexible, but requires more data and makes inference about changes in phase or amplitude harder. Other models might use a weighted average of the Controls (e.g. based on spatial position) or separate βs for different Controls. Some form of (9)'s 'value = F(site mean, regional variation, local variation, sampling error)' might yield a model for $\mathrm{E}\{I|Cs\}$, but no succinct one seems available. More mechanistic models, using information from the literature or supplementary experiments, may be more precise and better for management, corrective action, or distinguishing the effects of multiple alterations. Speed [59] and Raftery et al. [51] give excellent frequentist and Bayesian accounts, but the species' population dynamics are rarely known well enough. The possibilities are many but the practical ones often few because of sparse knowledge and short series.

One approach is to repeat the analysis with several models. When parameters of different models have different meanings, conversion is needed to compare effects of interest, e.g. absolute change, percent change, etc. These conversions seem easier from

(11) than from (10). A more formal approach uses weighted averages of results from different models (after conversion to common effect measures): methods are described by [10] and [29] from classical and Bayesian viewpoints, respectively.

Another source of uncertainty is changed conditions in the future. Defining the 'effect' as the difference between future values with and without the alteration over (say) 50 years lets us avoid taking geological and (perhaps) evolutionary change into account, but global warming and ozone depletion might affect both sets of future values by different amounts.

Examples

Numerical Example

Figure 1 shows data from the National Park Service's [45] annual samples (since 1982) of the Santa Barbara Channel Islands. There was no human alteration in this period, but I pretend that there was one at the North ('Impact') site between the 1990 and 1991 samples. The Impact data alone suggest a decrease, but the 'unaffected' site suggests there is correlation over a wider area, too long term to show up within a seven-year period.

With $x_i = 0$ or 1 for 'Before' (times $i =$ 1–8: 1983–1990) or 'After' (times 9–15), and I_i and C_i = Impact and Control averages at time i, I fitted five models:

$$I_i = \mu + \Delta x_i + \varepsilon_i \tag{12}$$

$$\log(I_i) = \mu + \Delta x_i + \varepsilon_i \tag{13}$$

$$I_i - C_i = \mu + \Delta x_i + \varepsilon_i \tag{14}$$

$$\log(I_i) - \log(C_i) = \mu + \Delta x_i + \varepsilon_i \tag{15}$$

$$\begin{aligned} I_i &= \alpha + \alpha_{\mathrm{A}} x_i + \beta C_i \\ &\quad + \beta_{\mathrm{A}} x_i C_i + \varepsilon_i \end{aligned} \tag{16}$$

Each model was fitted under six different assumptions about the correlation structure of the errors, ε_i: uncorrelated, AR(1), MA(1), AR(2), MA(2), and ARMA(1,1). The last 'Before' and first 'After' observations are correlated here. In many impact studies,

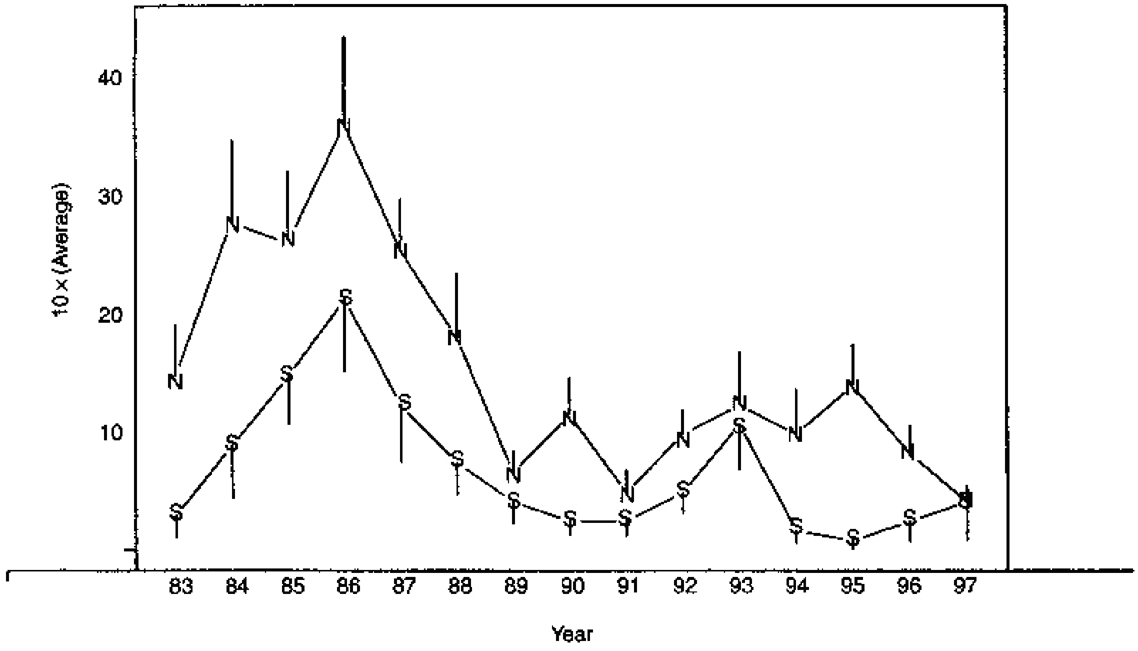


Figure 1 Average counts (×10) of the giant keyhole limpet, Megathura crenulata, on 60 m^2 bands at Johnson's Lee North (site 'N') and South ('S'), on Santa Rosa Island, California, from 1983 to 1997. The vertical lines show the estimated se due to that date's sampling error. (This is probably an overestimate: the samples were regular, not random)

the time gap between these two is much bigger than the other gaps, so this correlation is ignored. The Before and After values of σ^2 and the ϕs and θs might also be estimated separately.

S-PLUS's 'gls' function performs the ML iterations, but may approximate the full ARMA(p, q) likelihood by conditioning on the first p observations [70, p. 415]. Both likelihoods are approximate since normality is unlikely, and the difference would be negligible in a longer series, but I used gls, then used the ITSM program [8] on the residuals to re-estimate the ARMA parameters, used these estimates in gls (without iteration) for new estimates of μ, Δ, β_A, etc. used iterative time series modeling on the new residuals, and so on, stopping when the effect estimate, $\hat{\Delta}$ or $(\hat{\alpha}_A, \hat{\beta}_A)$, was stable – about 5–8 (but up to 18) extra iterations. The regression parameters did not change much, though the ARMA parameters sometimes did.

The six estimates for each model are summarized in Table 1. The 'best' ARMA model was the one with the smallest value of the Akaike Information Criterion (AIC). The autocorrelation and partial autocorrelation functions of the residuals from all final ARMA models were within (usually well within) the standard 0.05-level test limits ($\pm 1.96/\sqrt{n}$; $n = 15$) for all lags. The uncorrelated model had seven borderline values and one excess value each for (12) and (13); all its (16) values were well within the limits. For $n = 15$, the power of the test and the reliability of the asymptotic distributions may be doubted.

The percent change is 100 × $\hat{\Delta}$/(Impact Before average) for (12) and (14), 100 × exp($\hat{\Delta}$) for (13) and (15), and $100 \times (\hat{\alpha}_A + \hat{\beta}_A C_\bullet)/(\hat{\alpha} + {}^{\wedge}C_\bullet)$ for (16), where $C\bullet$ is the average of the Control values over all times. Approximate P values compare Estimate/$\widehat{\text{SE}}$ with the t-table on 13 (12)–(15) or 11 (16) degrees of freedom (df).

One lesson is that long-term correlation can cause misleading patterns, which can be modified by using Controls. This benefit will not always be apparent from the data. Compared with Impact site values, BACI differences have more sampling error and local variability, which can be estimated, but less long-term variation, which cannot. Here, the $\widehat{\text{SE}}$s of the Before–After estimates (12) and (13) are about the same as those of the BACI estimates (14) and (15), but they could have been smaller. Another lesson may be that model form matters more than correlation structure. For a given form, the different ARMA error structures usually gave answers within 1 $\widehat{\text{SE}}$ of each other. Differences between forms were greater. Higher order ARMA models might have given different results, but the series is too short for credible use of these. Only (15) gave the 'right' answer, no evidence for a change, but I chose the example for illustration and convenience (e.g. no zeros or seasons), so it may not be typical. More disturbing is the difference between (15) and (16): if $\ln(I_i) - \ln(C_i) = \mu$, then $I_i = e^{\mu}C_i$, so added random fluctuations and sampling error make (16) plausible. It may be that the flattening of slope estimates due to sampling error at Control (the 'errors in variables' problem) is greater in the After period because of reduced regional variation.

Published Examples

The following list is a selection only and therefore incomplete. Most add purposes, impacts or variables to the range of applications, and use statistical inferences not as decisions or revealed truth but as guides to biological conclusions in concert with other information. Many add details of data transformation and selection that real studies rarely escape, and none seem misled by implausible models. Some inferences

Table 1

Model equation	Range of estimates	Range of $\widehat{\text{SE}}$s	Best	Estimate ($\widehat{\text{SE}}$)	% change
(12)	$\hat{\Delta}$: −126 to −102	38 to 53	MA(2)	−122(38)	−60
(13)	$\hat{\Delta}$: −0.8 to −1.05	0.27 to 0.35	AR(1)	−0.85(0.35)	−57
(14)	$\hat{\Delta}$: −65 to −70	24 to 29	MA(1)	−68(28).	−23
(15)	$\hat{\Delta}$: −0.53 to +0.22	0.26 to 0.57	AR(2)	+0.22(0.26)	+24
(16)	$\hat{\alpha}_A$: −5.6 to +28	38 to 40	MA(1)	−5.64(39.5)	
	$\hat{\beta}_A$: −1.39 to −1.11	0.58 to 0.64		−1.12(0.58)	
	$\hat{\alpha}_A + \hat{\beta}_A C_\bullet$: −83 to −67	24 to 30		−82(30)	−46

take insufficient account of model uncertainty and autocorrelation. Estimates based on short study periods can be misleading, and natural variation such as El Niño may cause alteration effects to vary [36]. Warnken and Buckley [73] note that developers are reluctant to pay for monitoring before approval or to delay construction after it: of 44 parameters monitored at 13 Australian tourist developments, only one had more than one year of 'Before' data, and none had 10 Before times.

Accounts of BACI analyses include [41], using (16) above, [2, 55, 63], and a Bayesian approach [15]. The most complete accounts may be in reports of particular impact studies, where the design problems, ranges of species and impact mechanisms, and need for a coherent, quantitative biological 'story' to compare with economic and other concerns, can be best appreciated. Two good examples deal with Southern California power plants: [44] and [64]. Biological examples of IA, where no Control existed, are given in [46]. Simulation studies of power and validity are given in [38], using 15, and [35], using the Ricker stock-recruitment model (their equation (1) has a typo).

Biological intuition can be guided by assessment experiments. Areas may be deliberately altered [12, 16, 30, 32, 48, 53], previously altered sites selectively restored [72], an ongoing disturbance source manipulated [25], or captive organisms introduced into Impact and Control areas. These studies can clarify mechanisms and test statistical models, but raise new questions. Unavoidably, most use few sites, some only one per treatment. Potential interactions between sites and treatment mechanisms can make the definition and comparison of Impact and Control 'populations' tricky. In some source manipulations, 'treatment' periods or sites may have crossover effects on 'controls'.

Most assessments involve many species, and thus a multiple testing problem: the chance that all 95% confidence intervals are correct is far less than 95%. Methods that control 'studywise' error rates are controversial in general [60] and rarely useful in assessment: with dozens of species, they use highly unreliable extreme quantiles (e.g. 99.9%) and increase interval widths which are already large. In providing an overall description of effects, and an account of its uncertainty, the individual confidence intervals, with their limitations clearly explained, can be interpreted and weighted by their coherence with each other and other information – measurements and models of the alteration's physical and chemical effects, the physiology and interactions of the species involved, and so on. A drawback of this informality is its reliance on biological intuition. A more formal approach is to seek multivariate measures that can yield fewer, more coherent results and greater power (e.g. [26] and [27]). This has difficulties too: interpretation of principal components and ordinations, dubious linearity assumptions and uncertainty measures that ignore the data-based selection of analysis procedures. Verdonschot and Ter Braak [71] and Kedwards et al. [34] used permutation tests and ordination to assess experiments with several treatment levels: even with 12 'sites', power was low unless samples (rather than sites) could be permuted.

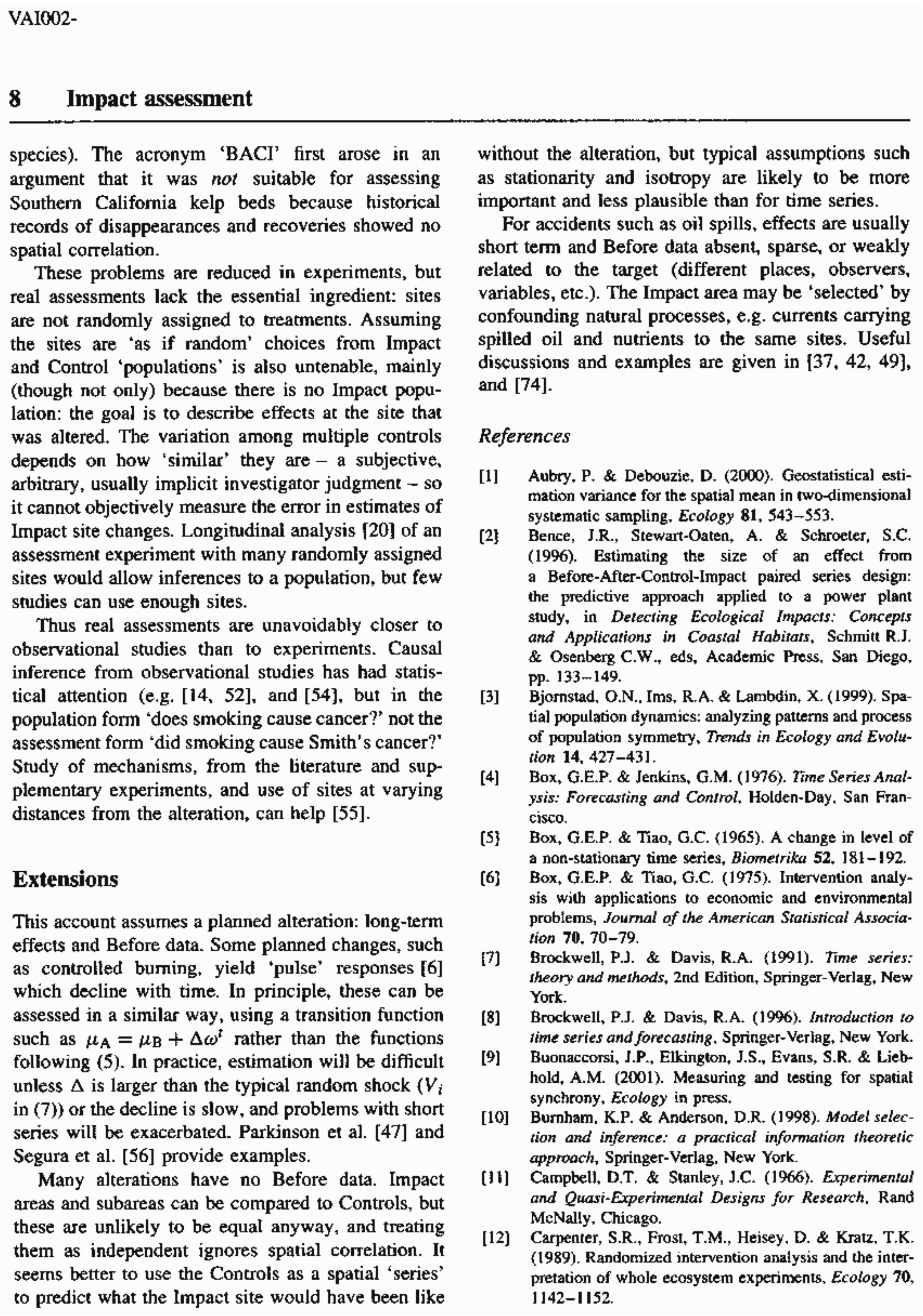
Experiments, multivariate methods, overall design and problems of early detection, are discussed and illustrated in a series of papers [23, 24, 31] on effects of mining runoff on streams (where Before–After, Impact–Control and upstream–downstream comparisons may all be relevant).

Assessments as Experiments?

Underwood [66–69], and others have argued that BACI and IA give insufficient evidence that the alteration *caused* a change, that different sites will always have different 'abundance trajectories' and that assessment should be treated as an experiment, with multiple randomly chosen controls used for error estimation. Stewart-Oaten and Bence [61] compare BACI with this radically different approach in detail.

IA and BACI do not use experimental controls. IA does not need controls. BACI uses carefully (not randomly) chosen controls as covariates, not to estimate error but to cope with long-term temporal variation. This does not require identical trajectories but enough 'synchrony' [3, 9] for the errors in models such as (15) or (16) to be of manageable size and adequately described by ARMA models of low order compared to T_B and T_A (2).

Choosing controls and models requires judgment and introduces uncertainty that is hard to measure. BACI controls may be unnecessary if long-term temporal variation at Impact is low, or even harmful if they track it poorly. Negligible long-term variation is rarely credible, but good Control sites may not exist or be hard to choose (especially for multiple



species). The acronym 'BACI' first arose in an argument that it was *not* suitable for assessing Southern California kelp beds because historical records of disappearances and recoveries showed no spatial correlation.

These problems are reduced in experiments, but real assessments lack the essential ingredient: sites are not randomly assigned to treatments. Assuming the sites are 'as if random' choices from Impact and Control 'populations' is also untenable, mainly (though not only) because there is no Impact population: the goal is to describe effects at the site that was altered. The variation among multiple controls depends on how 'similar' they are – a subjective, arbitrary, usually implicit investigator judgment – so it cannot objectively measure the error in estimates of Impact site changes. Longitudinal analysis [20] of an assessment experiment with many randomly assigned sites would allow inferences to a population, but few studies can use enough sites.

Thus real assessments are unavoidably closer to observational studies than to experiments. Causal inference from observational studies has had statistical attention (e.g. [14, 52], and [54], but in the population form 'does smoking cause cancer?' not the assessment form 'did smoking cause Smith's cancer?' Study of mechanisms, from the literature and supplementary experiments, and use of sites at varying distances from the alteration, can help [55].

Extensions

This account assumes a planned alteration: long-term effects and Before data. Some planned changes, such as controlled burning, yield 'pulse' responses [6] which decline with time. In principle, these can be assessed in a similar way, using a transition function such as $\mu_A = \mu_B + \Delta\omega^t$ rather than the functions following (5). In practice, estimation will be difficult unless Δ is larger than the typical random shock (V_i in (7)) or the decline is slow, and problems with short series will be exacerbated. Parkinson et al. [47] and Segura et al. [56] provide examples.

Many alterations have no Before data. Impact areas and subareas can be compared to Controls, but these are unlikely to be equal anyway, and treating them as independent ignores spatial correlation. It seems better to use the Controls as a spatial 'series' to predict what the Impact site would have been like without the alteration, but typical assumptions such as stationarity and isotropy are likely to be more important and less plausible than for time series.

For accidents such as oil spills, effects are usually short term and Before data absent, sparse, or weakly related to the target (different places, observers, variables, etc.). The Impact area may be 'selected' by confounding natural processes, e.g. currents carrying spilled oil and nutrients to the same sites. Useful discussions and examples are given in [37, 42, 49], and [74].

References

[1] Aubry, P. & Debouzie, D. (2000). Geostatistical estimation variance for the spatial mean in two-dimensional systematic sampling, *Ecology* **81**, 543–553.

[2] Bence, J.R., Stewart-Oaten, A. & Schroeter, S.C. (1996). Estimating the size of an effect from a Before-After-Control-Impact paired series design: the predictive approach applied to a power plant study, in *Detecting Ecological Impacts: Concepts and Applications in Coastal Habitats*, Schmitt R.J. & Osenberg C.W., eds, Academic Press, San Diego, pp. 133–149.

[3] Bjornstad, O.N., Ims, R.A. & Lambdin, X. (1999). Spatial population dynamics: analyzing patterns and process of population symmetry, *Trends in Ecology and Evolution* **14**, 427–431.

[4] Box, G.E.P. & Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco.

[5] Box, G.E.P. & Tiao, G.C. (1965). A change in level of a non-stationary time series, *Biometrika* **52**, 181–192.

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

[7] Brockwell, P.J. & Davis, R.A. (1991). *Time series: theory and methods*, 2nd Edition, Springer-Verlag, New York.

[8] Brockwell, P.J. & Davis, R.A. (1996). *Introduction to time series and forecasting*, Springer-Verlag, New York.

[9] Buonaccorsi, J.P., Elkington, J.S., Evans, S.R. & Liebhold, A.M. (2001). Measuring and testing for spatial synchrony, *Ecology* in press.

[10] Burnham, K.P. & Anderson, D.R. (1998). *Model selection and inference: a practical information theoretic approach*, Springer-Verlag, New York.

[11] Campbell, D.T. & Stanley, J.C. (1966). *Experimental and Quasi-Experimental Designs for Research*, Rand McNally, Chicago.

[12] Carpenter, S.R., Frost, T.M., Heisey, D. & Kratz, T.K. (1989). Randomized intervention analysis and the interpretation of whole ecosystem experiments, *Ecology* **70**, 1142–1152.

[13] Chatfield, C. (1984). *The analysis of time series: an introduction*, 3rd Edition, Chapman & Hall, London.

[14] Cox, D.R. (1992). Causality: some statistical aspects, *Journal of the Royal Statistical Society, Series A* **155**, 291–301.

[15] Crome, F.H.J., Thomas, M.R. & Moore, L.A. (1996). A novel Bayesian approach to assessing impacts of rain forest logging, *Ecological Applications* **6**, 1104–1123.

[16] Currie, D.R. & Parry, G.D. (1996). Effects of scallop dredging on a soft sediment community: a large-scale experimental study, *Marine Ecology Progress Series* **134**, 131–150.

[17] Davison, A.C. & Hinkley, D.V. (1997). *Bootstrap methods and their application*, Cambridge University Press, Cambridge.

[18] de Valpine, P. & Hastings, A. (2001). Fitting population models with process noise and observation error, *Ecology* in press.

[19] Diggle, P.J. (1990). *Time series: a biostatistical introduction*, Clarendon Press, Oxford.

[20] Diggle, P.J., Liang, K-Y. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*, Clarendon Press, Oxford.

[21] Eberhardt, L.L. (1976). Quantitative ecology and impact assessment, *Journal of Environmental Management* **4**, 27–70.

[22] Efron, B. & Tibshirani, R. (1993). *An introduction to the bootstrap*, Chapman & Hall, London.

[23] Faith, D.P., Dostine, P.L. & Humphrey, C.L. (1995). Detection of mining impacts on aquatic macroinvertebrate communities: Results of a disturbance experiment and the design of a multivariate BACIP monitoring programme at Coronation Hill, Northern Territory, *Australian Journal of Ecology* **20**, 167–180.

[24] Faith, D.P., Humphrey, C.L. & Dostine, P.L. (1991). Statistical power and BACI designs in biological monitoring: comparative evaluation of measures of community dissimilarity based on benthic macroinvertebrate communities in Rockhole Mine Creek, Northern Territory, Australia, *Australian Journal of Marine and Freshwater Research* **42**, 589–602.

[25] Granelli, E., Wallstrom, K., Larsson, U., Granelli, W. & Elmgren, R. (1990). Nutrient limitation of primary production in the Baltic area, *Ambio* **19**, 142–151.

[26] Green, R.H. (1979). *Sampling design and statistical methods for environmental biologists*, Wiley, New York.

[27] Green, R.H. (1989). Power analysis and practical strategies for environmental monitoring, *Environmental Research* **50**, 195–205.

[28] Harvey, A.C. (1989). *Forecasting, structural time series models and the Kalman filter*, Cambridge University Press.

[29] Hoeting, J.A., Madigan, D., Raftery, A.E. & Volinsky, C.T. (1999). Bayesian Model Averaging: A Tutorial, *Statistical Science* **14**, 382–417.

[30] Hogg, I.D. & Williams, D.D. (1996). Response of stream invertebrates to a global warming thermal regime: an ecosystem-level manipulation, *Ecology* **77**, 395–407.

[31] Humphrey, C.L., Faith, D.P. & Dostine, P.L. (1995). Baseline requirements for assessment of mining impact using biological monitoring, *Australian Journal of Ecology* **20**, 150–166.

[32] Hurd, M.K., Perry, S.A. & Perry, W.B. (1996). Nontarget effects of a test application of diflubenzuron to the forest canopy on stream macroinvertebrates, *Environmental Toxicology and Chemistry* **15**, 1344–1351.

[33] Jones, R.H. (1993). *Longitudinal data with serial correlation: a state space approach*, Chapman & Hall, London.

[34] Kedwards, T.J., Maund, S.J. & Chapman, P.F. (1999). Community level analysis of ecotoxological field studies: II. Replicated-design studies, *Environmental Toxicology and Chemistry* **18**, 158–166.

[35] Korman, J. & Higgins, P.S. (1997). Utility of escapement time series data for monitoring the response of salmon populations to habitat alteration, *Canadian Journal of Fisheries and Aquatic Science* **54**, 2058–2067.

[36] Lee, R.S. & Pritchard, T.R. (1996). How do long-term patterns affect time-limited environmental monitoring programmes? *Marine Pollution Bulletin* **33**, 260–268.

[37] Lopes, C.F., Milanelli, J.C.C., Prosperi, V.A., Zanardi, E. & Truzzi, C. (1997). Coastal monitoring program of São Sebastião Channel: assessing the effects of "Tebar V" oil spill on rocky shore populations, *Marine Pollution Bulletin* **11**, 923–927.

[38] Mandenjian, C.P., Jude, D.J. & Tesar, F.J. (1986). Intervention analysis of power plant impact on fish populations, *Canadian Journal of Fisheries and Aquatic Sciences* **43**, 819–829.

[39] Manly, B.J.F. (1997). *Randomization and Monte Carlo Methods in Biology*, 2nd Edition, Chapman & Hall, London.

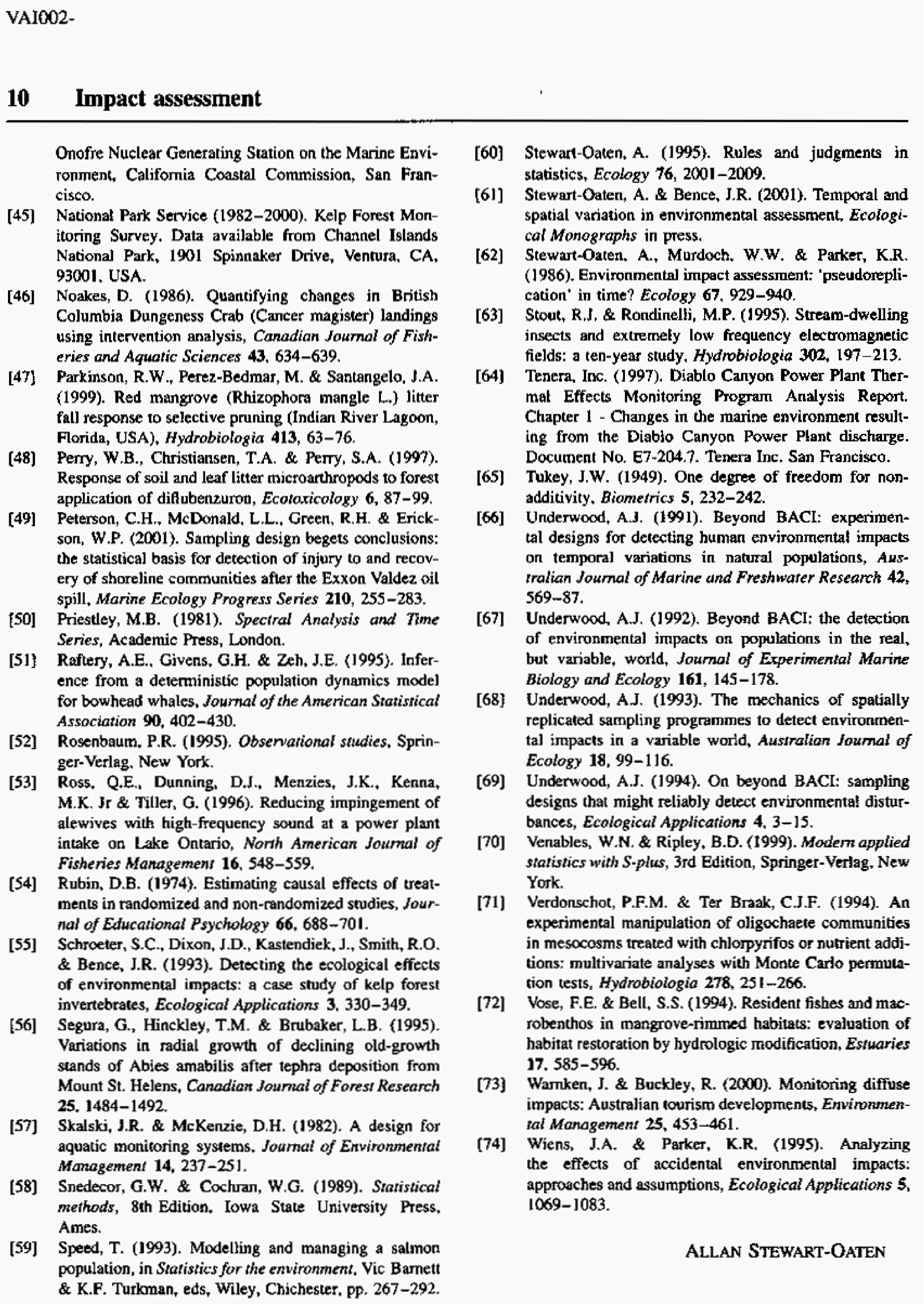
[40] Manly, B.F.J., Olsen, A.R. & Smith, E.P. (eds) (1999). Special issue on sampling over time, *Journal of Agricultural, Biological and Environmental Statistics* **4**, 327–507.

[41] Mathur, D., Robbins, T.W. & Purdy, E.J. (1980). Assessment of thermal discharges on zooplankton in Conowingo Pond, Pennsylvania, *Canadian Journal of Fisheries and Aquatic Science* **37**, 937–944.

[42] McDonald, L.L., Erickson, W.P. & Strickland, M.D. (1995). Survey design, statistical analysis, and basis for statistical inferences in coastal habitat injury assessment: Exxon Valdez oil spill, in *Exxon Valdez Oil Spill: Fate and Effects in Alaskan Waters, ASTM STP 1219*, P.G. Wells, J.N. Butler & J.S. Hughes, eds. American Society of Testing and Materials, Philadelphia, pp. 296–311.

[43] McDowall, S.P., McCleary, R., Meidinger, E.E. & Hay, R.A. (1980). *Interrupted time series analysis*, Sage, Beverly Hills.

[44] Murdoch, W.W., Mechalas, B. & Fay, R.C. (1989). Final report of the Marine Review Committee to the California Coastal Commission on the effects of the San



VAI002-

Onofre Nuclear Generating Station on the Marine Environment, California Coastal Commission, San Francisco.

[45] National Park Service (1982–2000). Kelp Forest Monitoring Survey. Data available from Channel Islands National Park, 1901 Spinnaker Drive, Ventura, CA, 93001, USA.

[46] Noakes, D. (1986). Quantifying changes in British Columbia Dungeness Crab (Cancer magister) landings using intervention analysis, *Canadian Journal of Fisheries and Aquatic Sciences* **43**, 634–639.

[47] Parkinson, R.W., Perez-Bedmar, M. & Santangelo, J.A. (1999). Red mangrove (Rhizophora mangle L.) litter fall response to selective pruning (Indian River Lagoon, Florida, USA), *Hydrobiologia* **413**, 63–76.

[48] Perry, W.B., Christiansen, T.A. & Perry, S.A. (1997). Response of soil and leaf litter microarthropods to forest application of diflubenzuron, *Ecotoxicology* **6**, 87–99.

[49] Peterson, C.H., McDonald, L.L., Green, R.H. & Erickson, W.P. (2001). Sampling design begets conclusions: the statistical basis for detection of injury to and recovery of shoreline communities after the Exxon Valdez oil spill, *Marine Ecology Progress Series* **210**, 255–283.

[50] Priestley, M.B. (1981). *Spectral Analysis and Time Series*, Academic Press, London.

[51] Raftery, A.E., Givens, G.H. & Zeh, J.E. (1995). Inference from a deterministic population dynamics model for bowhead whales, *Journal of the American Statistical Association* **90**, 402–430.

[52] Rosenbaum, P.R. (1995). *Observational studies*, Springer-Verlag, New York.

[53] Ross, Q.E., Dunning, D.J., Menzies, J.K., Kenna, M.K. Jr & Tiller, G. (1996). Reducing impingement of alewives with high-frequency sound at a power plant intake on Lake Ontario, *North American Journal of Fisheries Management* **16**, 548–559.

[54] Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies, *Journal of Educational Psychology* **66**, 688–701.

[55] Schroeter, S.C., Dixon, J.D., Kastendiek, J., Smith, R.O. & Bence, J.R. (1993). Detecting the ecological effects of environmental impacts: a case study of kelp forest invertebrates, *Ecological Applications* **3**, 330–349.

[56] Segura, G., Hinckley, T.M. & Brubaker, L.B. (1995). Variations in radial growth of declining old-growth stands of Abies amabilis after tephra deposition from Mount St. Helens, *Canadian Journal of Forest Research* **25**, 1484–1492.

[57] Skalski, J.R. & McKenzie, D.H. (1982). A design for aquatic monitoring systems, *Journal of Environmental Management* **14**, 237–251.

[58] Snedecor, G.W. & Cochran, W.G. (1989). *Statistical methods*, 8th Edition, Iowa State University Press, Ames.

[59] Speed, T. (1993). Modelling and managing a salmon population, in *Statistics for the environment*, Vic Barnett & K.F. Turkman, eds, Wiley, Chichester, pp. 267–292.

[60] Stewart-Oaten, A. (1995). Rules and judgments in statistics, *Ecology* **76**, 2001–2009.

[61] Stewart-Oaten, A. & Bence, J.R. (2001). Temporal and spatial variation in environmental assessment, *Ecological Monographs* in press.

[62] Stewart-Oaten, A., Murdoch, W.W. & Parker, K.R. (1986). Environmental impact assessment: 'pseudoreplication' in time? *Ecology* **67**, 929–940.

[63] Stout, R.J. & Rondinelli, M.P. (1995). Stream-dwelling insects and extremely low frequency electromagnetic fields: a ten-year study, *Hydrobiologia* **302**, 197–213.

[64] Tenera, Inc. (1997). Diablo Canyon Power Plant Thermal Effects Monitoring Program Analysis Report. Chapter 1 - Changes in the marine environment resulting from the Diablo Canyon Power Plant discharge. Document No. E7-204.7. Tenera Inc. San Francisco.

[65] Tukey, J.W. (1949). One degree of freedom for nonadditivity, *Biometrics* **5**, 232–242.

[66] Underwood, A.J. (1991). Beyond BACI: experimental designs for detecting human environmental impacts on temporal variations in natural populations, *Australian Journal of Marine and Freshwater Research* **42**, 569–87.

[67] Underwood, A.J. (1992). Beyond BACI: the detection of environmental impacts on populations in the real, but variable, world, *Journal of Experimental Marine Biology and Ecology* **161**, 145–178.

[68] Underwood, A.J. (1993). The mechanics of spatially replicated sampling programmes to detect environmental impacts in a variable world, *Australian Journal of Ecology* **18**, 99–116.

[69] Underwood, A.J. (1994). On beyond BACI: sampling designs that might reliably detect environmental disturbances, *Ecological Applications* **4**, 3–15.

[70] Venables, W.N. & Ripley, B.D. (1999). *Modern applied statistics with S-plus*, 3rd Edition, Springer-Verlag, New York.

[71] Verdonschot, P.F.M. & Ter Braak, C.J.F. (1994). An experimental manipulation of oligochaete communities in mesocosms treated with chlorpyrifos or nutrient additions: multivariate analyses with Monte Carlo permutation tests, *Hydrobiologia* **278**, 251–266.

[72] Vose, F.E. & Bell, S.S. (1994). Resident fishes and macrobenthos in mangrove-rimmed habitats: evaluation of habitat restoration by hydrologic modification, *Estuaries* **17**, 585–596.

[73] Warnken, J. & Buckley, R. (2000). Monitoring diffuse impacts: Australian tourism developments, *Environmental Management* **25**, 453–461.

[74] Wiens, J.A. & Parker, K.R. (1995). Analyzing the effects of accidental environmental impacts: approaches and assumptions, *Ecological Applications* **5**, 1069–1083.

ALLAN STEWART-OATEN

VAP046-

Pseudo-replication

Hurlbert [5] defines pseudo-replication as 'testing for treatment effects with an error term inappropriate to the hypothesis being considered'. An alternative definition is 'estimating the size of chance variation without accounting for all its likely sources'. Examples include:

1. estimating the decomposition rate of leaves in a lake using values obtained from N bags of leaves, all placed in the same plot;
2. comparing the rates of oak and maple leaves, using samples of N from two plots, one oak and one maple;
3. equivalent to case 2 with each sample coming from k plots, each with $N/k > 1$ bags;
4. testing effects of fox predation on rodent sex ratio, using counts of male and female rodents from one field with foxes and one without, in a χ^2 test. This is similar to case 2, but the χ^2 from may make it harder to recognize;
5. other ecological field and laboratory studies where values of units (e.g. organisms) treated in batches (in aquaria, microcosms or fields) are analyzed as if independent.

In each case, treating the observations as independent ignores sources of variation. If there is variation among plots or fields, then the true sample size is not N but k in case 3 and one in the other cases. The true level of variation is underestimated, so the uncertainty of conclusions is greater than the nominal level.

True variation cannot be estimated in cases 1, 2 and 4 without an assumption about variation among plots. It can be estimated from the two sets of k plot averages in case 3, if plot positions were randomly VAR011- chosen (*see* **Randomization**). Some investigators test that there is no 'plot' component of variation before pooling bags into a common sample. This protection can be unreliable: if the difference between the pooled and unpooled inferences is critical, then the test probably has low power against levels of variation that are low but non-negligible. Other justification is needed, especially strong evidence based on previous experience. This is rare in ecology.

One response, the claim that variation among 'replicable' units can be ignored, muddles the underlying concepts. Replication is needed for validity: to assess variation among units and thus the size of error and uncertainty. It does not require or imply homogeneity (which is useful for efficiency: to reduce variation, error and uncertainty). 'Replicability' among units has no meaning. Zero variation among units requires not only that they be identical (not just similar) in all relevant respects at the start of the experiment, but that they experience identical conditions during the experiment, and usually that treatments can be applied to different units in an identical manner. Most claims of 'replicability' establish only rough initial similarity in a few respects.

Recognizing Pseudo-replication

Several questions seem helpful.

1. Are units assigned to treatments independently?
2. Are treatments applied to units in batches (especially in a single batch per treatment)?
3. Does the experiment design make units given the same treatment likely to experience greater similarity under other conditions than units given different treatments?
4. Are there likely sources of variation whose effects will be more similar, on average, for units getting the same rather than different treatments?

To avoid pseudo-replication, an analysis that treats the unit values as independent should answer 'Yes' to question and 1 'No' to the rest. No single question seems to cover all possibilities. It is easy to violate question 1 while satisfying the rest. We satisfy questions 1 and 2 but not 3 and 4 if we randomly assign rats to injections of different drugs, applied independently, but then cage them so that each cage contains several rats, all with the same drug. It is not necessary that there be only one cage per drug, nor that the set of cages for one drug differs in a predictable direction from other sets. Questions 3 and 4 both ask 'are there sources of variation that cause correlation to be greater for units given the same treatments?'.

A design in which units arise in blocks need not imply pseudo-replication. Even if random treatment assignment ignores the blocks and leads to some treatments being concentrated in some blocks, p values and other inferences based on randomization remain valid. Pseudo-replication arises if the design – i.e. the assignment method – makes some lop-sided arrangements especially likely.

VAP046-

2 Pseudo-replication

It is not always obvious what a 'unit' is. In this section, it is the entity contributing a single value (which may be a vector) to the analysis. It could be a single rat, but if rats are weighed several times then it could be a single rat if the repeated weighings are treated as a single multivariate observation, or a combination of a rat and a weighing if all weighings from all rats are treated as independent (*see* VAR035- **Repeated measures**). Pseudo-replication occurs in the latter case. One can think of pseudo-replication as the misidentification of units, but this suggests that the analyst should first identify the units and then carry out an analysis that treats them as independent. If the rat weighings are taken over a period of time, then one may want to use them separately (rather than just averaging them) to trace the progress of drug effects or fit parameters of a theoretical model. Also, 'units' may be hard to define. Hurlbert [5, p. 190] argues that if the bags of example 2 are all randomly placed in the same plot (e.g. at random points on the same isobath), then the 'units' are the physical locations. Others might feel that the bags are the units VAM012- and the random plot effects constitute **measurement error** (cf. [9, Chapter 9]). In such cases it can be easier and more reliable to determine the analysis by specifying and justifying a formal mathematical model (e.g. outcome = species mean + bag effect + position effect, the last two being independent), a process too often discouraged in ecological data analysis.

A General Criticism and Some Consequences

Hurlbert treats probability as if it has a standard definition and method of measurement, like a physical property. Instead, at least three sources of chance are common in environmental work: (i) random assignment of units to treatments, (ii) random sampling from the population of interest, and (iii) random choices by 'Nature'.

Source (i) is the most credible since randomizing mechanisms like coin tosses or random number generators can in principle be checked for any patterns to any degree of precision. It is available only for experiments where the investigator controls assignment. Each unit is assumed to have a set of predestined values, one for each treatment. Only one of these is observed. Under the 'unit-treatment additivity' assumption, each treatment's effect is the same on each unit: there is a set of unit values, U_i, and a set of treatment values, T_j, such that the value of unit i under treatment j equals $U_i + T_j$. Under any null hypothesis specifying the T_js (e.g. H_0: all $T_j = 0$), the U_is can be calculated from the observations. We can then calculate the test statistic that would have been obtained for any assignment of the T_js to the U_is. Listing these values for all possible assignments gives the null distribution, and hence tests and confidence intervals. Inferences about units not used in the experiment are based on the additivity assumption. From this viewpoint, failure to randomize does not lead to approximate error probabilities [5, p. 197] but to none at all: Hurlbert's 'layout-specific α' is meaningless until the observations are made, and then it is one if the null was rejected and zero if it was not [4, Chapter 2], [6] and [9] are good guides to this approach.

Source (ii) makes derivations easier and is needed in observational studies. In experiments, it is equivalent to random selection from the population followed by random assignment to treatments. It avoids source (i)'s additivity assumption by adding the random selection assumption. This is usually a bad trade. Additivity may be approximately true (possibly after transformation), can be checked, and may be needed for intelligible results (e.g. to decide on commercial application or to speculate sensibly about mechanisms). Random selection is rarely true, and can lead to worthless data if it has tempted the investigator to dispense with literal random assignment. Kempthorne [7, p. 322] calls it 'usually completely ludicrous' for experiments, and it seems dubious for most observational studies of nonhuman populations. Few laboratory organisms, aquaria, or research station plots are randomly chosen from a relevant population. In medical experiments, the population of concern is usually future patients, who cannot be randomly sampled. It is also dubious for most observational studies of nonhuman populations. Hurlbert's example 5 [5, p. 193] is like example 2 above, except that all 2*N* bags are randomly assigned within the same plot. He accepts that the single plot would be sufficient for 'comparison of the two species', but contradicts this by arguing that a comparison of oak and maple for the 1 m isobath of a class of lakes requires random selection of lakes and positions. But this argument also requires random selection of the leaves that, taken literally, seems impossible for a single tree, let alone a species. Even then, results would apply only to the

VAP046-

time of the experiment, unless random times could somehow be selected as well. Of course, using several lakes and plots is preferable to using a single plot: it makes additivity easier to check (especially with blocking) and the results more robust against moderate deviations from it. Using randomness in selection, along with other considerations, can also be useful, if only to reduce bias. But the belief that inferences must be justified by random selection from the population of interest is an illusion that can lead to substandard inferences and damaged experiments (e.g. when random selection leads to poor coverage or inaccessible plots). At some point, we must assume our results apply to populations we have not randomly sampled; evidence from inside the experiment (e.g. of additivity) can help with this, but external evidence is usually required.

Source (iii) uses models in which deterministic but unpredictable natural events are represented by chance variables. It can be combined with source (i) or (ii), e.g. to allow for measurement error or covariates, but is also the only choice in a wide array of observational studies, especially of complex processes that either cannot be repeated under 'identical' VAT010- VAS039- conditions (such as **time series** or **spatial analysis in ecology**) or whose mechanisms or stages are of interest (e.g. studies of learning, evolution, invasion or disease). Hurlbert's 'layout-specific α' and his argument for interspersion of treatments are based on an implicit chance model, roughly of random variation about a monotone trend in times or sites. Probability calculation, selection of designs and assessment of the gains and losses of interspersion require explicit models. Polynomial trends of low degree may best combine plausibility and tractability. Cox [2] gives a succinct but thorough account with explicit recommendations. Cox [3] compares interspersion, blocking, use of position as a covariate, and other methods in the case of linear trend.

'Validity' must be assessed for all three sources by identifying assumptions and examining their plausibility. Hurlbert's criticism [5, p. 204] of impact studies and 'temporal pseudo-replication' mixes several issues. To dismiss inferences from analyses of time series or spatial data on the grounds of 'pseudo-replication' is clearly mistaken. Analyses based on independence are more vulnerable, but correlations can be positive, negative or also sometimes zero (or VAI002- at least negligible). **Impact assessment** is not an experiment. The task is to assess impact at a particular site, not a population of sites, so assignment of treatments to sites is irrelevant: 'pseudo-replication' does not arise from one impact and one control station [5, p. 204], or even from no control at all. (For example, Box and Tiao [1] had no 'control' for Los Angeles in their intervention analysis of its air pollution.) An impact will cause the 'before' and 'after' time series to differ. Inference about this requires a model, to be assessed by its plausibility and its fit to the data. For instance, treating multiple values from the same time and site as independent assumes populations do not vary naturally over time (or that sites vary in perfect unison); this is highly implausible and, with a single before and single after time, cannot be checked against the data, so 'pseudo-replication' is apt. However, the assumption that site values, or differences between sites, from several before and after times have negligible correlation might be plausible in some systems if the time gaps are large enough and can be checked against the data, so 'pseudo-replication' is inappropriate. A model that allows for serial correlation is usually better, but there are many of these, and any usable one will have to limit the complexity or order of the correlation.

Thus the 'plausibility' question is not whether the model is true (no model is completely so, not even those based on randomization), but whether it adequately represents our knowledge and uncertainty about the system. Adequacy may require judgment about tradeoffs between realism and simplicity. High order correlations may be required for realism but, if they are small, a model that ignores them is likely to give better predictions and more accurate inferences. Some cases of pseudo-replication (e.g. in physics) might be justifiable if experience shows variation among batches (e.g. laboratories) is negligible compared with variation within them (e.g. repeated runs in the same laboratory). Hurlbert implicitly allows for such cases, by giving empirical evidence that the ignored variation is not negligible in the cases he criticizes.

Inferences cover causes only in experiments, where source (i) is either used directly or implied by source (ii). Without a chance model for the assignment of units to treatments, inferences can cover differences among groups receiving different treatments but cannot distinguish group effects from treatment effects (*see* **Confounding**). This does VAC047- not mean that nothing can be said about cause:

ruling out plausible alternatives is useful, and even quantitative inference can be possible under additional assumptions about the limits of bias in assignment by 'Nature' [8]. Inference about causes using a chance model for assignment is the key difference between experiments and observational studies, and between Cox's 'treatment' and 'classification' factors within experiments [3]. Hurlbert's distinction between 'manipulative experiments' and 'mensurative experiments' muddies both. His discussion [5, p. 190] misunderstands Cox, for whom species certainly could be a treatment factor when its assignment to units is under the investigator's control rather than intrinsic.

Conclusions

This article has devoted more space to disagreements than to agreements only because the latter are so well discussed by Hurlbert. His paper still improves the statistical practice of its numerous readers and their colleagues and students, and stimulates them to think more clearly about issues far less obvious than they once seemed. It identifies, explains and memorably names an important error, with detailed examples showing its frequency, many disguises and distinguished victims. By facilitating and inspiring similar reviews covering more recent periods or other disciplines, it appears to have raised consciousness not just for a moment but over the long run. It remains a 'must read'.

References

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

[2] Cox, D.R. (1951). Some systematic experimental designs, *Biometrika* **38**, 312–323.

[3] Cox, D.R. (1958). The use of a concomitant variable in selecting an experimental design, *Biometrika* **44**, 150–158.

[4] Cox, D.R. (1958). *The Planning of Experiments*, Wiley, New York.

[5] Hurlbert, S.H. (1984). Pseudoreplication and the design of ecological field experiments, *Ecological Monographs* **54**, 187–211.

[6] Kempthorne, O. (1955). The randomization theory of experimental inference, *Journal of the American Statistical Association* **50**, 946–967.

[7] Kempthorne, O. (1975). Inference from experiments and randomization, in *A Survey of Statistical Design and Linear Models*, J.N. Srivastava, ed., North-Holland, Amsterdam.

[8] Rosenbaum, P.R. (1995). *Observational Studies*, Springer-Verlag, New York.

[9] Scheffé, H. (1959). *The Analysis of Variance*, Wiley, New York.

(*See also* **Alpha-designs**; **Clustering**; **Ecologi-** VAA014- VAC028-
cal study design; **Nested Experimental Designs**; VAE007- VAN010-
Spatial design, optimal; **Surveys, environmental** VAO010-
design/data collection) drop VAS066-

ALLAN STEWART-OATEN

The Department of the Interior Mission

As the Nation's principal conservation agency, the Department of the Interior has responsibility for most of our nationally owned public lands and natural resources. This includes fostering sound use of our land and water resources; protecting our fish, wildlife, and biological diversity; preserving the environmental and cultural values of our national parks and historical places; and providing for the enjoyment of life through outdoor recreation. The Department assesses our energy and mineral resources and works to ensure that their development is in the best interests of all our people by encouraging stewardship and citizen participation in their care. The Department also has a major responsibility for American Indian reservation communities and for people who live in island territories under U.S. administration.

The Minerals Management Service Mission

As a bureau of the Department of the Interior, the Minerals Management Service's (MMS) primary responsibilities are to manage the mineral resources located on the Nation's Outer Continental Shelf (OCS), collect revenue from the Federal OCS and onshore Federal and Indian lands, and distribute those revenues.

Moreover, in working to meet its responsibilities, the **Offshore Minerals Management Program** administers the OCS competitive leasing program and oversees the safe and environmentally sound exploration and production of our Nation's offshore natural gas, oil and other mineral resources. The MMS **Royalty Management Program** meets its responsibilities by ensuring the efficient, timely and accurate collection and disbursement of revenue from mineral leasing and production due to Indian tribes and allottees, States and the U.S. Treasury.

The MMS strives to fulfill its responsibilities through the general guiding principles of: (1) being responsive to the public's concerns and interests by maintaining a dialogue with all potentially affected parties and (2) carrying out its programs with an emphasis on working to enhance the quality of life for all Americans by lending MMS assistance and expertise to economic development and environmental protection.