Comments

Ecology, 84(10), 2003, pp. 2795–2799

ON REJECTION RATES OF PAIRED INTERVENTION ANALYSIS: COMMENT

Allan Stewart-Oaten[1]

Murtaugh (2000, 2002) claims the Before–After, Control–Impact (BACI) approach to assessment of long-term local effects of a planned environmental alteration (such as a development) ignores serial correlation and assumes the Control and Impact values would have ''parallel trajectories'' in the absence of an impact.

In the BACI approach, observations are taken at the altered (''Impact'') site at a set of times before the alteration and another set after it, and matching observations are taken at one or more unaltered ''Control'' sites, sufficiently near and similar to share major natural events (e.g., weather, seasons) and respond in similar ways, but far enough away for the alteration effect on them to be negligible. A formal statistical analysis compares the Before and After time series at the Impact site, using the Control site values as concomitants, e.g., as covariates, or by adjusting each Impact site value by subtracting the average of the corresponding Control site values. Here, I focus mainly on the second approach, comparing the Before and After time series of (Impact − [average of Controls]).

Murtaugh (2000, 2002) criticizes this approach, claiming it ignores possible serial correlation of the between-site differences and assumes that Impact and Control trajectories would have been exactly parallel in the absence of an intervention. Both flaws would increase the rate of false ''detection'' of an impact, i.e., make the chance of rejecting a true null hypothesis of ''no alteration effect'' greater than the nominal chance. Murtaugh (2002) reports that RIA (randomized intervention analysis), a nonparametric version of BACI that compares the Before and After values of (Impact − [average of Controls]) by a permutation test, rejected the null hypothesis in 12 out of 61 data sets (20%) involving no known alteration. His own ''two-stage'' procedure had no rejections for these data. (I do not discuss his other results, for data sets with known alterations.)

I discuss these criticisms, arguing that BACI analyses can, and usually should, allow for serial correlation, and that ''parallel trajectories'' (previously called ''additivity'') is not required of the site values but only of the means of a stochastic model of them. It can be judged only by judging the whole model, especially the assumptions about the stochastic errors, and is then often plausible, at least as an approximation—which is all we demand of other statistical models. Some of the points have been made before, by Stewart-Oaten, Murdoch, and Parker (1986) and by Stewart-Oaten and Bence (2001), henceforth abbreviated to SOMP and SOB. I also argue that Murtaugh's (2002) test results have more plausible explanations than a tendency for BACI to produce false positives. However, concerns about additivity and correlation are not misplaced; I discuss two of the most difficult of these problems, and possible responses.

The Criticisms

Serial correlation.—The claim that this problem has been ignored seems strange: it takes up nearly 20% of SOMP (pp. 935–937). This mainly argues that serial correlation in the Impact-Control differences might be negligible compared to sampling error and uncorrelated temporal variation if the Impact and Control sites have similar-enough behavior, but an example is given where the correlation is not negligible, and there is brief reference to methods allowing for it. Such methods were used in several of the BACI-based analyses of the effects of the San Onofre Nuclear Generating Station (Murdoch et al. 1989). Serial correlation is also discussed by Mathur et al. (1980) and at length by SOB (e.g., pp. 313–318).

Positive serial correlation increases the error in effect estimates and, if ignored, causes this error to be underestimated. If there is enough synchrony between Control and Impact site fluctuations, much of the serial correlation at the Impact site could be eliminated in the Impact–Control differences, or in an analysis using Controls as covariates. This might reduce the error in effect estimates, though it is not guaranteed to do so because it also adds the unrelated variation and sampling error in the Control site observations. It may have more chance of reducing the underestimation of error, by removing long-term fluctuations that are especially hard to model and fit in the short series usually available for assessment.

Manuscript received 26 August 2002; revised 30 January 2003; accepted 31 January 2003. Corresponding Editor: A. M. Ellison.

[1] Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara, California 93106-9610 USA. E-mail: stewart@lifesci.ucsb.edu

In principle, BACI analyses can include serial correlation, of any structure, just as easily as any other time-series analyses. In practice, the series are usually short, so only simple models can be fitted. BACI's potential benefit is that, for suitable Controls, a simple and weak (low order and low values) correlation structure or even independence may provide an adequate (not perfect) representation for the differences. Whether it does so must be judged by biological and environmental arguments, plots, analytical criteria like the Akaike Information Criterion (AIC), tests like Durbin-Watson, and other methods. Almost always, models with at least first-order autocorrelation would need to be considered. A worked example appears in Stewart-Oaten (2002).

Parallel trajectories.—This assumption is called ''additivity'' in SOMP (pp. 931–932) and SOB. It often seems misunderstood. If the variable of interest is the abundance of some species, three values must be distinguished. The observed abundance is usually an estimate based on a sample of the site. The ''censused'' abundance is the value that would be obtained by an error-free census of the entire site. The third value is more abstract. The censused abundance is assumed to be the outcome of a process involving births, deaths, and migrations resulting from weather, predation, parasitism, disease, starvation, encounters with mates, behavioral choices by individuals and groups, etc. Being unable to measure all these, or predict their effects accurately, we model them collectively as ''chance'' variation. This represents the censused abundance at any time as a random variable. The mean of this variable is the third value.

This is not a mean over time, but over possible outcomes of the process, which is assumed to be repeatable. SOB (pp. 311–315) give a basis for this modeling approach, in terms of predicted future abundances with vs. without the alteration. We can write

$$S(t) = A_S(t) + \xi_S(t) = \mu_S(t) + \varepsilon_S(t) + \xi_S(t) \qquad (1)$$

where $S(t)$ is the observed abundance at time t and site S, $A_S(t)$ is the censused abundance, $\xi_S(t)$ is sampling error, $\mu_S(t)$ is the mean of $A_S(t)$ over possible outcomes of the abundance-producing process, and $\varepsilon_S(t)$ is the difference, $A_S(t) - \mu_S(t)$, due to chance variation in this process.

Using I = Impact and C = Control (or average of Controls), it is alteration-induced change in $\mu_I(t)$ that is the target of the assessment. From Eq. 1 we have

$$I(t) - C(t) = \mu_I(t) - \mu_C(t) + \varepsilon_I(t) - \varepsilon_C(t) + \xi(t) \qquad (2)$$

where $\xi(t)$ combines the sampling errors. If we can assume that $\mu_I(t) - \mu_C(t)$ is constant within a period (Before or After), the problem reduces to estimating the Before-After change in this constant. It is this assumption which Murtaugh (2002) calls ''unreasonable.''

By itself, it is not an assumption at all. In Eq. 1 the distinction between $\mu_S(t)$ and $\varepsilon_S(t)$ depends on what we choose to regard as ''random'' (SOB, p. 315). We can easily satisfy additivity by treating all temporal variation as stochastic, so that $\mu_S(t) = \mu_S$, a constant. The additivity ''assumption'' arises only when we describe the distribution of the $\varepsilon_S(t)$'s, so it can be judged only as part of the whole model: Is it biologically reasonable, an adequate fit to these data, and in accord with other relevant data?

The assumption $A_S(t) = \mu_S + \varepsilon_S(t)$, where the ε's are independent and identically distributed, would surely be unreasonable in many situations, though there might be exceptions, e.g., annual levels of a short-lived species, with no significant interactions with long-lived species, in an area with more weather variability within years than between them.

But BACI's additivity assumption can follow from weaker assumptions. The analysis need not be on raw abundances—logs, reciprocals or other transformations could be used. The mean, $\mu_S(t)$, does not need to be constant: it can have seasonal or other variation, provided this is the same at all sites. Models like $\log(S(t)) = \mu_S + \alpha\cos(2\pi t) + \beta\sin(2\pi t) + \varepsilon_S(t) + \xi_S(t)$ might reasonably represent multiplicative response to seasonal variation if t is measured in years (SOMP, pp. 933–934). BACI does need a tractable error structure, but this refers to the difference, $\varepsilon_I(t) - \varepsilon_C(t)$, so would be satisfied by $\varepsilon_S(t) = \phi_S(t) + \psi(t)$, where $\phi_S(t)$ is tractable while $\psi(t)$ is the same for all sites. The general requirement for the additivity assumption is that

$$f(S(t)) = \mu_S + \phi(t, \gamma_S) + \psi(t) + \xi_S(t) \qquad (3)$$

where f is a known transformation, ϕ is a time series whose correlation structure is known except for the unknown parameters, γ_S, ψ is a function that is common to all sites, and ξ is sampling error. Since ψ cancels in the differences, there is no need to separate it into deterministic and stochastic parts. For short time series we need the correlation in ϕ to be weak, and the number of unknown parameters in γ to be small.

Additivity is not a necessary assumption in BACI. It is not used if Controls are used as covariates. Even the ''difference'' analysis, Eq. 2, can allow variation in the mean function:

$$f(I(t)) - f(C(t)) = g(t, \beta) + \Phi(t, \lambda) + \xi(t) \qquad (4)$$

where $\Phi(t, \lambda) = \phi(t, \gamma_I) - \phi(t, \gamma_C)$ from Eq. 3, and $\xi(t) = \xi_I(t) - \xi_C(t)$. Here, g is a known function of the unknown parameters, β, and the ''effect'' is defined by Before-After change in these. For example, g could include linear or seasonal terms:

$$g(t, \beta) = \beta_0 + \beta_1 t + \beta_2 \sin(2\pi t) + \beta_3 \cos(2\pi t). \quad (5)$$

Of course, specific choices must be made for *f*, *g*, and the correlation structure of Φ in Eq. 4. These choices are unlikely to be exactly correct—just like assumptions of normal data, straight-line regressions, additive treatment effects and random samples in other models. Given the difficulty of defining frequentist probability, a ''correct'' stochastic model may be not just nonexistent but indefinable. It is a statistical adage that ''all models are false but some are useful'' (e.g., Chatfield 1995: 428): they make biological sense, fit the data, and address the question of interest. Burnham and Anderson (1998) give a good account of frequentist methods for assessing models, and of modeling philosophy in general. Since the additivity assumption applies only to the underlying process, is flexible with respect to transformations, additional functions of time, and covariates, and does not require perfection, it seems unreasonable to call it ''unreasonable.''

Sites as ''units''.—Murtaugh (2002:1752) hints at a criticism that has been taken seriously by others: ''statistical inference based on a single pair of units seems impractical . . . .'' SOB (pp. 322, 326, and ''The basis for inference'' [p. 327]) point out that assessment is the comparison of Before and After conditions at the Impact site. A major part of the evidence for these conditions is the Before and After time series at this site. Their effective ''units'' are times, though these must usually be treated as dependent. Sites are not ''units'' any more than are other potential covariates, like rainfall.

The Test Results

Murtaugh (2002) used the BACI model

$$I(t) - C(t) = \lambda + \eta(t) \quad (6)$$

on 61 data sets involving no known intervention. This is Eq. 2 with $\lambda = \mu_I(t) - \mu_C(t)$ and $\eta(t) = \varepsilon_I(t) - \varepsilon_C(t) + \xi(t)$. He found a ''significant'' Before-After change in λ in 12 cases when the η's were assumed to be independent and identically distributed, and in 9 cases where they were assumed to follow the autoregressive with one step (AR(1)) model

$$\eta(t) = \rho\eta(t - 1) + a(t) \quad (7)$$

where the *a*'s are assumed to be independent and identically distributed. These false ''detection'' rates, 20% and 15%, are well above the nominal 5%, but there are several explanations other than failure of the BACI approach.

If the 61 tests were independent, and each had a 5% chance of ''significance,'' then the number of significant cases should be about 3, with 1 SD $\approx$ 1.7. Thus either 12 or 9 is much larger than would be expected by chance. But the tests are not independent. They include 42 pairwise comparisons of seven Wisconsin lakes with respect to two response variables; eight of these comparisons are ''significant,'' but all eight involve a single lake, TB. Without this lake, there would be 4 significant tests out of 49—still greater than 5% but not ''significantly'' so if we regard these tests as independent (2.45 $\pm$ 1.53 [mean $\pm$ 1 SD]).

All the tests used untransformed data. Whether this was appropriate cannot be judged without more information, but the combination of additivity and simple correlation structure may require log or other transformations. Murtaugh's (2002) Figs. 2A and 3B and C seem to support this. If the appropriate transformation is very different from the one used (in this case, none), the ''$\psi(t)$'' of Eq. 3 varies among sites and may not approximately cancel in the difference, $I(t) - C(t)$. In long Before-and-After time series the main effect of this may be an increase in unexplained variation, and thus a decrease in power; but in short series, it may be an increase in bias and ''false positives.'' For example, there may be bias if seasonal variation is not allowed for and cancels poorly, and the fraction of winter observations is higher in the Before period than in the After period. If large natural fluctuations with long-lasting effects cancel poorly in the Impact–Control differences, then they can function as ''effects'': one or two of them can dominate one part of the series (Before or After), making it ''significantly'' different from the other part. Both problems are more likely if the series are short. Only 2 of Murtaugh's (2002) 61 ''no impact'' series (data sets I and M in his Table 1) use Before and After periods of two years or more.

Murtaugh's AR(1) structure gives some protection against these problems, but it may be weak. When the series is short, the standard estimate of ρ in Eq. 7 is biased low so the *t* test has too many false positives even when adjusted for correlation (Bence 1995). Murtaugh's (2002) test, using permutation of residuals with the same adjustment, seems likely to have the same problem. Of course, this is a problem for the BACI procedure also, just as it is for any method requiring adjustment for serial correlation in a short series.

The AR(1) structure may also be too simple for the autocorrelated case, whose simplest form may arise when $I(t)$ and $C(t)$ both have the form of Eq. 3, with $\phi_I(t)$ and $\phi_C(t)$ satisfying Eq. 7. Then $\phi_I(t) - \phi_C(t)$ is AR(1) if the ρ's of Eq. 7 are the same, and autoregressive moving average (ARMA) (2,1) otherwise; the BACI errors, $\phi_I(t) - \phi_C(t) + \xi(t)$, are then ARMA(1,1) or ARMA(2,2) respectively. In practice, such theoretical arguments should not get too much weight, but they do show that the BACI errors may be more complicated than AR(1).

Discussion

Murtaugh's (2000, 2002) objections to BACI are based either on an oversimplified version or on misunderstanding, and his test results have likely explanations other than BACI yielding too many false positives (or badly underestimating the standard errors of effect estimates), when properly used. This removes the basis for Murtaugh's conclusion that statistical inference should not be attempted with BACI data, and ''graphical displays, expert opinion and common sense'' be used instead (Murtaugh 2002:1758).

This conclusion also seems an abdication of responsibility. Graphs and experts have important roles in inference; both can help us choose among models and among possible causal explanations, and help clarify complicated conclusions. However, both can be quantitatively vague and usually downplay the uncertainty of conclusions. In many cases, both sides can find supporting graphs and experts (especially if there is a fee), in part because we lack clear principles for resolving disputes among them. ''Common sense'' is too often a rhetorical refusal to look beyond the superficial. Statistical inference, though far from perfect, has gone far towards overcoming these weaknesses—properly carried out and explained, it can be thought of as a systematized, objective form of common sense. Most of its disputes are about models, where differences can be clarified and narrowed, if not always resolved. If scientists reject the use of these tools, environmental decisions will still be made—but not as objectively or as well.

However, BACI is not problem-free (SOB, pp. 322 and 334). Murtaugh's concerns about additivity and correlation are not insolvable but must be addressed in any implementation. Two problems seem especially difficult; both are related to a main purpose of BACI, to reduce the role of large, long-lasting natural changes, which are hard to model, can mimic alteration effects, and add greatly to the uncertainty of effect estimates:

(1) The choice of transformation, or more generally of the model linking $\mu_I(t)$ and the $\mu_C(t)$'s and other covariates, will be based on biological arguments and the within-period time series, especially the Before series. Often, too little is known for biological arguments to narrow the choice very much. The data may not help much either, if the $\mu(t)$'s vary little within periods; it may be hard to choose between (say) ''$\mu_I(t) - \mu_C(t) = \text{constant}$'' and ''$\log(\mu_I(t)) - \log(\mu_C(t)) = \text{constant}$.'' But the wrong choice might give misleading conclusions if there is a large regional change between the Before and After periods, just as a straight line can fit a nonlinear function well over a small range but be misleading when extrapolated over a large one. Even here, an imperfect BACI model is likely to do better than an analysis using no Control sites at all, especially if several models or transformations are used and give similar answers, but BACI results need to be treated with special caution if the Controls show a greater difference between periods than within them.

(2) With the right model, BACI reduces contamination from large, long-lasting changes as long as they are broad enough to affect both Impact and Controls similarly. Local changes that affect Impact and Controls very differently remain a problem. They are part of the BACI model's chance error, so part of the error in effect estimates, whose variances will be underestimated if they are not properly accounted for. Short-lived local changes will affect results much as sampling error does: they make confidence intervals wider and tests less powerful, but should not make confidences or *P* values inaccurate. Long-lived local changes will also contribute to variance estimates, but may still cause underestimation if the correlation they induce is underestimated or is more complex than our model allows. Local changes that are large, long-lasting, and rare (e.g., other human disturbances) are a special problem. If one of these occurs during the Before or After period, it plays the role of an outlier, inflating the variance estimate. If none occur within a period, the variance estimate is too small: it does not allow for the possibility of an occurrence between the periods, which can mimic an alteration effect. Such changes could explain some of Murtaugh's (2002) test results; a change at lake TB could explain 8 out of 12 of them. Murtaugh's (2000) two-stage procedure is an attempted cure, but it uses the separate site averages, which BACI avoids because they are affected by fixed site differences and by large, region-wide temporal variation. There are other possibilities. Part of any assessment is construction of the biological ''story'': Do the size and spatial pattern of apparent alteration effects make sense physiologically and ecologically? If some types of local changes are known to occur but are not included in the model, there may be ''signals'' that indicate them, like changes in chemicals, other species, or spatial patterns of changes among the Controls. It may also be possible to use estimates of the temporal and spatial rates of particular types of potentially confounding changes in the region over recent years to modify (decrease) the confidence with which an apparent effect is attributed to the alteration.

These problems and others (many related to model uncertainty) show that inferences from BACI assessments need caution. BACI is not alone in this. Cox (2001) comments that, in model-based inference generally, ''significance tests, confidence intervals (etc.) . . . indicate the uncertainty that would apply under . . . idealized conditions and as such are often lower bounds to real uncertainty'' (p. 217) and that ''representations of under-

lying process have to be viewed with . . . caution, but this does not make them fruitless'' (p. 218). So too with BACI.

Acknowledgments

This 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 Initiative). Comments by the Associate Editor and two anonymous reviewers improved the organization of this paper, and especially clarified and deepened the Discussion.

Literature cited

Bence, J. R. 1995. Analysis of short time series: correcting for autocorrelation. Ecology **76**:628–639.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. Second edition. Springer-Verlag, New York, New York, USA.

Chatfield, C. 1995. Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society A **158**:419–466.

Cox, D. R. 2001. Comment on ''Statistical modeling: the two cultures'' (by L. Breiman). Statistical Science **16**:216–218.

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 Science **37**:937–944.

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.

Murtaugh, P. A. 2000. Paired intervention analysis in ecology. Journal of Agricultural, Biological and Environmental Statistics **5**:280–292.

Murtaugh, P. A. 2002. On rejection rates of paired intervention analysis. Ecology **83**:1752–1761.

Stewart-Oaten, A. 2002. Impact assessment. *In* A. H. El-Shaarawi and W. W. Piegorsch, editors. Encyclopedia of Environmetrics. Volume 2. John Wiley and Sons, Chichester, UK.

Stewart-Oaten, A., and J. R. Bence. 2001. Temporal and spatial variation in environmental impact assessment. Ecological Monographs **71**:305–339.

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

Ecology, 84(10), 2003, pp. 2799–2802

ON REJECTION RATES OF PAIRED INTERVENTION ANALYSIS: REPLY

Paul A. Murtaugh[1]

Stewart-Oaten (2003) criticizes my paper on paired intervention analysis (Murtaugh 2002) on several grounds. By ''paired intervention analysis'' I mean before–after, control–impact (BACI) analysis and randomized intervention analysis (RIA) applied to data from a single pair of ecological units.

The problem of serial correlation

Increasing numbers of authors are looking for, and attempting to adjust for, serial correlation (Hewitt et al. 2001, Levin and Tolimieri 2001, Rumbold et al. 2001, Zimmer et al. 2001), but many others continue to overlook the problem (Basset et al. 2001, Guidetti 2001, Guillemette and Larsen 2002, Roman et al. 2002, Rybczyk et al. 2002). They may have good reason: it is well known that modeling serial correlation requires large numbers of observations—more than are collected in the typical BACI study. For example, Ramsey and Schafer (2002:454) feel that, with $n < 50$, the usual tools for adjusting for serial correlation ''are unlikely to yield reliable results.''

The key result of my paper is that, even after adjustment for serial correlation, the rejection frequency for pairs of units receiving no intervention is still 3 times the supposed 0.05 level of the tests (my original Table 2). The small reduction in false-positive rate effected by the serial-correlation adjustment suggests there are more fundamental problems with the BACI approach.

Parallel trajectories

A. Stewart-Oaten defends the assumption of parallel trajectories of the response in the two units (also called ''additivity''). His models for the observed abundance at time t in site S (Stewart-Oaten et al. 1986, Stewart-Oaten and Bence 2001, Stewart-Oaten 2003) all take the general form

$$Y_S(t) = \mu_S(t) + \text{error} \quad (1)$$

where $\mu_S(t)$ is the expected value of abundance at time t, or, in Stewart-Oaten's (2003) parlance, the ''mean of [censused abundance] over possible outcomes of the abundance-producing process,'' with ''censused abundance'' meaning ''the value that would be obtained by an error-free census of the entire site''; and the error

Manuscript received and accepted 20 February 2003. Corresponding Editor: A. M. Ellison.

[1] Department of Statistics, Oregon State University, Corvallis, Oregon 97331 USA. E-mail: murtaugh@stat.orst.edu

is ''due to chance variation in this [abundance-producing] process'' plus sampling error (Stewart-Oaten 2003).

Much has been made of the nature and labeling of the components of variance of the error term (e.g., see Murtaugh 2000, Stewart-Oaten 2003), but this is perhaps academic, given that the components cannot be separately estimated from the single time series of between-site differences available from the simple BACI design. It is of course necessary to specify the covariance structure of the errors in order to attempt statistical inference, but debates over details of that structure divert attention from what I feel is a more important issue.

If an intervention with effect δ is applied to site I at time t^*, Eq. 1 implies that the differences in abundance between sites I and C can be written as

$$Y_I(t) - Y_C(t) = \mu_I(t) - \mu_C(t) + \delta \times I(t > t^*) + \text{error} \quad (2)$$

where $I(t > t^*)$ is 1 for times greater than t^* and zero otherwise. The error term here is a composite of the site- and time-specific errors in Eq. 1.

The BACI estimate of the effect of the intervention is

$$\hat{\delta} = (\text{average } \textit{post}\text{-intervention difference in } \textit{observed} \text{ abundances between the two sites}) - (\text{average } \textit{pre}\text{-intervention difference in } \textit{observed} \text{ abundances between the two sites}). \quad (3)$$

Given the model in Eq. 2, it's clear that the expected value of $\hat{\delta}$ is

$$E(\hat{\delta}) = \delta + (\text{average } \textit{post}\text{-intervention difference in } \textit{expected} \text{ abundances between the two sites}) - (\text{average } \textit{pre}\text{-intervention difference in } \textit{expected} \text{ abundances between the two sites}) = \delta + \Delta. \quad (4)$$

If desired, one can replace ''expected abundances'' by ''mean censused abundance, over possible outcomes of the abundance-producing process'' (Stewart-Oaten 2003). Note that $\hat{\delta}$ is an unbiased estimator of the intervention effect (i.e., $E(\hat{\delta}) = \delta$) only if the two averages in Eq. 4 are identical, i.e., if $\Delta = 0$.

There are many ways that the time series of $\mu_I(t)$ and $\mu_C(t)$ could, fortuitously, have the property that $\Delta = 0$, but the most natural is that the mean trajectories of abundance in the two sites are parallel (i.e., $\mu_I(t) - \mu_C(t)$ is constant for all t). This is the so-called ''additivity'' assumption.

One's view of the usefulness of BACI analysis therefore hinges on how likely one thinks it is that $\Delta = 0$.

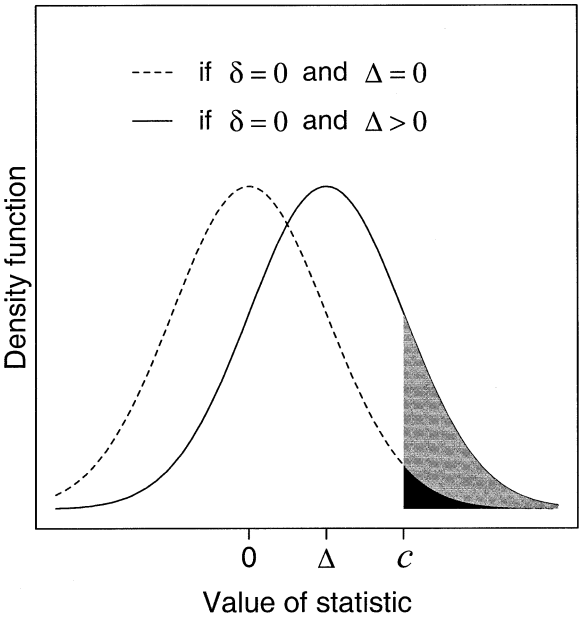

FIG. 1. Schematic diagram of the probability density function of the BACI test statistic, $\hat{\delta}$, when $\delta = 0$ and either $\Delta = 0$ (dashed line) or $\Delta > 0$ (solid line). The black area is the nominal 5% rejection region, to the right of the critical value, c; the grey region shows the inflated rejection rate when $\Delta > 0$.

Since there is no replication of the process that results in the time series of differences, $Y_I(t) - Y_C(t)$, there are no data on which to build a view of how likely it is that $\Delta = 0$. If Δ is non-zero, then we are ''shooting for the wrong target'' when we estimate the intervention effect δ using Eq. 3, and the rejection frequency of the BACI test will be too high (see Fig. 1). Note that this result obtains *whatever* error structures are assumed for the two time series of abundances and/or the time series of differences—all we are assuming is that the errors have mean zero.

Of course, one way to guarantee that $\Delta = 0$ is to *assume* that it is; nothing stops Stewart-Oaten (2003) from ''treating all temporal variation as stochastic, so that $\mu_S(t) = \mu_S$, a constant''. The idea that one can arbitrarily partition variability between signal and noise, as if by decree, runs counter to the precepts of frequentist statistics, which is essentially a tool to provide an objective basis for doing that partitioning. Of course, the usefulness of that tool hinges on the existence of replication at the pertinent level, which is lacking in the simple BACI design.

The data analyses

Hoping to take the debate beyond abstract musings like those in the preceding sections, I assembled data from pairs of unmanipulated units described in the eco-

logical literature, and found that, in 15–20% of cases, the results of randomized intervention analysis applied to paired ''reference'' units were statistically significant (Murtaugh 2002). Stewart-Oaten is unfazed by this result, for which he offers several possible explanations (italicized):

1) *Inadequate modeling of the error covariance structure causes P values to be underestimated.* My original Table 2 shows that incorporating first-order serial correlation in the BACI analyses—probably the most complicated error modeling these short time series can bear—caused only a small reduction in the false-positive rate.

2) *Inadequate transformation of responses to achieve additivity.* Even if one could find a transformation that stabilized the between-unit differences *before* the intervention, there is no basis for projecting that stability to the post-intervention period (i.e., for assuming $\Delta = 0$ in Eq. 4).

3) *If we remove the comparisons involving the Wisconsin lake labeled TB, the rejection frequency drops to a level indistinguishable from 0.05, using a statistical criterion based on the assumption of independence of the remaining comparisons.* This assumption is patently false, which is why I avoided such calculations in the first place. It is obvious that if one removes a subset of the data having a high false-positive rate the overall rate will drop. The problem is, investigators don't know a priori which units are going to end up being false positives.

It is worth noting here that, in an earlier version of the manuscript, I recorded a 39% false-positive rate in 101 comparisons of unmanipulated units. In response to a reviewer who questioned my classifications of ''reference'' units and choices of study periods, I eliminated data from seven sources and reduced the time scales of some of the other analyses.

Conclusions

Consider a pair of human subjects, A and B, whose blood pressures are monitored over time. Suppose that subject A is given an antihypertensive drug at time t^*, and that the difference between A's and B's blood pressures increases after t^*. A BACI analysis attributes that increase to an effect of the drug, by assuming that, in the absence of intervention, the expected difference in blood pressure between subjects would not vary with time ($\Delta = 0$)—an assumption I doubt many physicians would be willing to make. Taken alone, this result has no statistical value in testing for efficacy of the drug; only when it is combined with results from other pairs of subjects, having an array of Δ's centering on zero, can we construct a meaningful confidence interval for the drug's effect. If useful statistical inference could be based on a single pair of subjects, why do medical scientists work so hard to boost enrollment in clinical trials?

Ecologists have long recognized the importance of ecosystem-level studies, which often preclude replication. But, are we justified in relaxing our statistical standards because lakes and forests are harder to ''enroll'' and measure than are human subjects, or because we feel compelled to ''get something significant'' out of the enormous effort required to do ecological experiments on a large scale?

Proponents of BACI analysis have defended their approach by asserting, correctly, that inference cannot be extended beyond that specific pair of sites. I would argue that, taken alone, such inference gives a biased estimate of the intervention effect, and, in any case, in real studies authors are almost always interested in making general statements about the effect of an intervention on sites like those used in the study. Such inference *must* be based on designs having some replication of control and/or manipulated units (e.g., see DeLucia et al. 1999, Stanley et al. 2002).

Does that mean that unreplicated ecosystem-level manipulations are without merit? Of course not. Would the studies of Likens et al. (1970) on a single pair of watersheds be more compelling if they had been accompanied by BACI-derived *P* values? Would their results have been less compelling if *three* pairs of watersheds had been used, and an analysis correctly based on this level of replication yielded $P = 0.10$? In my opinion, this sort of slavish devotion to *P* values (and, yes, confidence intervals) gets in the way of good science.

Stewart-Oaten (2003) views my skepticism about BACI analyses as an ''abdication of responsibility,'' an abandonment of the objectivity that we must bring to scientific investigations. I would respond that *no P* values are better than incorrect ones. As a statistician, I could not agree more that ''properly carried out and explained, [statistical inference] can be thought of as a systematized, objective form of common sense.'' Improperly carried out, statistical inference can be misleading, distracting, and detrimental to the progress of science.

Acknowledgments

I thank Allan Stewart-Oaten for his comments on my paper, and the editors of *Ecology* for allowing me to respond to them. I am also grateful to the dozens of students, colleagues, and friends who have endured my ranting about this subject over the past several years!

Literature cited

Basset, Y., E. Charles, D. S. Hammond, and V. K. Brown. 2001. Short-term effects of canopy openness on insect herbivores in a rain forest in Guyana. Journal of Applied Ecology **38**:1045–1058.

DeLucia, E. H., J. G. Hamilton, S. L. Naidu, R. B. Thomas, J. A. Andrews, A. Finzi, M. Lavine, R. Matamala, J. E.

Mohan, G. R. Hendrey, and W. H. Schlesinger. 1999. Net primary production of a forest ecosystem with experimental CO^2 enrichment. Science **284**:1177–1179.

Guidetti, P. 2001. Detecting environmental impacts on the Mediterranean seagrass *Posidonia oceanica* (L.) Delile: the use of reconstructive methods in combination with ''beyond BACI'' designs. Journal of Experimental Marine Biology and Ecology **260**:27–39.

Guillemette, M., and J. K. Larsen. 2002. Postdevelopment experiments to detect anthropogenic disturbances: the case of sea ducks and wind parks. Ecological Applications **12**: 868–877.

Hewitt, J. E., S. E. Thrush, and V. J. Cummings. 2001. Assessing environmental impacts: effects of spatial and temporal variability at likely impact scales. Ecological Applications **11**:1502–1516.

Levin, P. S., and N. Tolimieri. 2001. Differences in the impacts of dams on the dynamics of salmon populations. Animal Conservation **4**:291–299.

Likens, G. E., F. H. Bormann, N. M. Johnson, D. W. Fisher, and R. S. Pierce. 1970. Effects of forest cutting and herbicide treatment on nutrient budgets in the Hubbard Brook watershed-ecosystem. Ecological Monographs **40**:23–47.

Murtaugh, P. A. 2000. Paired intervention analysis in ecology. Journal of Agricultural, Biological and Environmental Statistics **5**:280–292.

Murtaugh, P. A. 2002. On rejection rates of paired intervention analysis. Ecology **83**:1752–1761.

Ramsey, F. L., and D. W. Schafer. 2002. The statistical sleuth: a course in methods of data analysis. Second edition. Duxbury, Pacific Grove, California, USA.

Roman, C. T., K. B. Raposa, S. C. Adamowicz, M.-J. James-Pirri, and J. G. Catena. 2002. Quantifying vegetation and nekton response to tidal restoration of a New England salt marsh. Restoration Ecology **10**:450–460.

Rumbold, D. G., P. W. Davis, and C. Perretta. 2001. Estimating the effect of beach nourishment on *Caretta caretta* (loggerhead sea turtle) nesting. Restoration Ecology **9**:304–310.

Rybczyk, J. M., J. W. Day, and W. H. Conner. 2002. The impact of wastewater effluent on accretion and decomposition in a subsiding forested wetland. Wetlands **22**:18–32.

Stanley, T. R., and F. L. Knopf. 2002. Avian responses to late-season grazing in a shrub-willow floodplain. Conservation Biology **16**:225–231.

Stewart-Oaten, A. 2003. On rejection rates of paired intervention analysis: comment. Ecology **84**:2795–2799.

Stewart-Oaten, A., and J. R. Bence. 2001. Temporal and spatial variation in environmental impact assessment. Ecological Monographs **71**:305–339.

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

Zimmer, K. D., M. A. Hanson, and M. G. Butler. 2001. Effects of fathead minnow colonization and removal on a prairie wetland ecosystem. Ecosystems **4**:346–357.