On Rejection Rates of Paired Intervention Analysis: Reply
Author(s): Paul A. Murtaugh
Source: *Ecology*, Vol. 84, No. 10 (Oct., 2003), pp. 2799-2802
Published by: Wiley on behalf of the Ecological Society of America
Stable URL: https://www.jstor.org/stable/3450122
Accessed: 20-02-2019 22:41 UTC

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

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

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

$\hat{\delta}$ = (average *post*-intervention difference in *observed* abundances between the two sites) − (average *pre*-intervention difference in *observed* abundances between the two sites). (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 $\Delta = 0$.

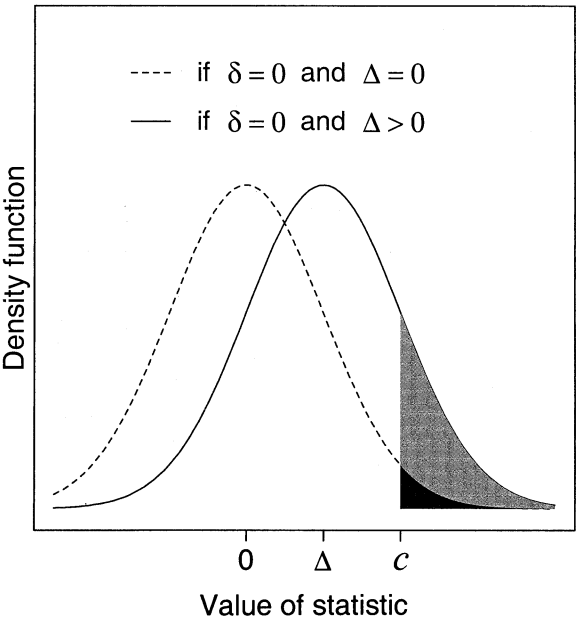

FIG. 1. Schematic diagram of the probability density function of the BACI test statistic, $\hat{\delta}$, when $\delta = 0$ and either $\Delta = 0$ (dashed line) or $\Delta > 0$ (solid line). The black area is the nominal 5% rejection region, to the right of the critical value, c; the grey region shows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