Paired Intervention Analysis in Ecology

Paul A. MURTAUGH

The paired watershed experiments of Likens and coworkers in the Hubbard Brook Experimental Forest are examples of a classical design in ecology, in which a response in a manipulated unit is compared both to the response in the same unit before manipulation and to the response in an adjacent reference unit that remains undisturbed. Early proponents of this design did not attempt statistical analysis of their results but, more recently, before-after-control-impact analysis and randomized intervention analysis have been used by ecologists to draw statistical inferences from such data. These methods are simply two-sample comparisons (before vs. after) of between-unit differences, with significant results often interpreted as evidence for an effect of the intervention. This approach ignores variation caused by differences between units in the trajectories of the response through time, and it does not take into account possible serial correlation of errors. Consequently, the null hypothesis may be rejected much too often. I develop a new, two-stage analysis method that addresses these shortcomings by correcting for serial correlation and using half-series means to assess temporal variation. Unlike paired intervention analysis, the resulting test has close to the nominal level when the time course of the response is allowed to vary between units, but its power is extremely limited due to the lack of true replication in the design.

Key Words: Before-after-control-impact design; Environmental impact assessment; Environmental monitoring; Randomized intervention analysis; Serial correlation; Two-stage intervention analysis.

1. BACKGROUND

Large-scale ecological experiments often involve little or no replication, making statistical analysis of the results difficult or impossible. Classical examples include the whole-watershed manipulations of Likens and coworkers in the Hubbard Brook Experimental Forest (Likens 1985; Likens, Bormann, Johnson, Fisher, and Pierce 1970). Following a paradigm established by the Wisconsin limnologist A. D. Hasler, Likens and colleagues applied a treatment to an individual watershed and compared ecological responses in that watershed both to preintervention responses in the same watershed and to responses in an adjacent reference watershed that remained undisturbed. Statistical analyses of the these data were not attempted.

Paul A. Murtaugh is Associate Professor, Department of Statistics, Oregon State University, Corvallis, Oregon 97331.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 5, Number 3, Pages 280–292

Recently, an approach based on comparison of time series of observations in a single pair of control and manipulated units has been increasingly used by ecologists as a way of deciding whether a statistically significant change has occurred in the manipulated unit relative to the control. Variants of this approach have been labeled before-after-control-impact (BACI) analysis (Stewart-Oaten, Murdoch, and Parker 1986) and randomized intervention analysis (RIA; Carpenter, Frost, Heisey, and Kratz 1989).

The basic idea of these methods, which I will refer to jointly as paired intervention analysis, is as follows. A time series of some response is obtained for each of two ecological units, one of which receives an intervention after some time has passed and the other of which is an undisturbed control. The intervention may be a treatment imposed by an experimenter or a fortuitous event in an observational study. For each observation time, the difference in the response between the two units is calculated. The postintervention differences are then compared statistically to the preintervention differences with a two-sample test such as a t-test (BACI) or a randomization test (RIA). A significant p-value is interpreted as evidence of a difference in the time course of the response between the two units, which might be attributable to the intervention. Figure 1 shows a schematic diagram of the design.

There are at least two limitations of this approach:

(1) Neither BACI nor RIA accounts for possible serial correlation of observations in the times series of differences. If there is positive correlation, p-values will be too small. This problem has been recognized and discussed by proponents of paired intervention analysis (cf., Stewart-Oaten et al. 1986; Carpenter et al. 1989; Stewart-Oaten 1996).

(2) Even if serial correlation is not present in a particular data set, a disparity between the preintervention and postintervention differences could be due to an effect of the intervention *or* to a difference in the underlying response trends that has nothing to do with the intervention. A difference in the natural trajectories of the response between the units could be wrongly interpreted as an effect of the intervention, and a real effect of the intervention could be masked by fortuitous changes in the two time series (Underwood 1992).

There have been various criticisms of the precepts and implementation of paired intervention analysis (Hurlbert 1984; Underwood 1992, 1994; Smith, Orvos, and Cairns 1993), some of which were anticipated by the original proponents of this approach (Stewart-Oaten et al. 1986; Carpenter et al. 1989). Nevertheless, the methodology seems to be firmly entrenched in the ecological literature (cf., Faith, Humphrey, and Dostine 1991; Roberts 1993; Schroeter, Dixon, Kastendiek, Smith, and Bence 1993; Reitzel, Elwany, and Callahan 1994; Vose and Bell 1994; Stout and Rondinelli 1995; Uddameri, Norton, Kahl, and Scofield 1995; Hogg and Williams 1996; Lydersen, Fjeld, and Gjessing 1996; Schmitt and Osenberg 1996; Wallace, Eggert, Meyer, and Webster 1997). Using a combination of statistical modeling, data analysis, and computer simulation, I argue here that paired intervention analysis rejects the null hypothesis much too often, and I present and evaluate a new method of analyzing data from these sorts of studies.

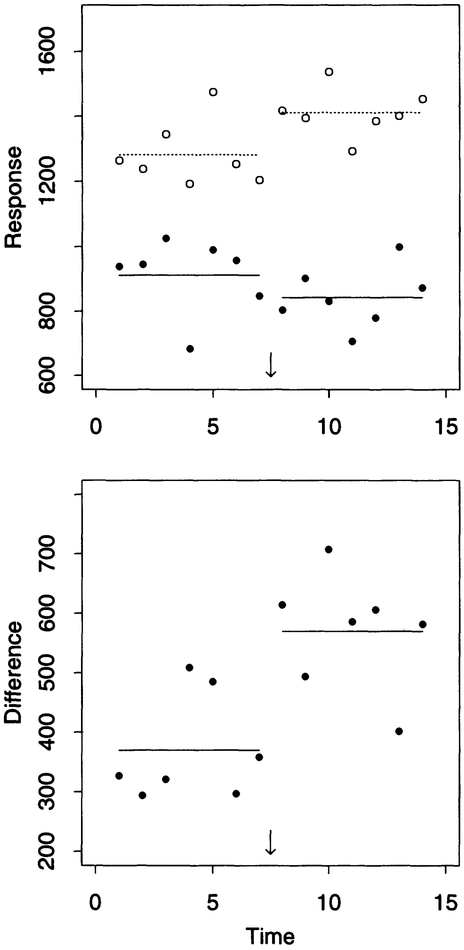


Figure 1. Plots of Hypothetical Data, With Arrow Indicating the Time of the Intervention. Top: The original time series from unit 1 (solid points and solid lines showing the half-series means) and unit 2 (open points, dashed lines). Bottom: The time series of differences, with half-series means indicated by horizontal lines. The distance between the lines is the estimate of the intervention effect.

2. STATISTICAL MODELS

2.1 Modeling the Two Time Series Separately

Let $y_1(t_i)$ and $y_2(t_i)$ be the measured responses in the control and manipulated units, respectively, at time t_i and let $\mu_j(t_i)$ be the true mean response in unit j at t_i. If an intervention applied at time t^* raises the true response in the manipulated unit by an amount δ, we can write

$$\begin{aligned} y_1(t_i) &= \mu_1(t_i) + \epsilon_{1i} \\ y_2(t_i) &= \mu_2(t_i) + \delta \cdot I_{[t^*,\infty)}(t_i) + \epsilon_{2i}, \end{aligned} \tag{2.1}$$

where ϵ_{1i} and ϵ_{2i} are measurement errors with variance σ^2, say. If the intervention is expected to have a multiplicative effect on the response, one can think of Equation (2.1) as applying to log-transformed data.

As a special case of the above model, consider a constant half-series means model,

$$\mu_1(t_i) = \begin{cases} \mu_{1\mathrm{B}} & \text{if } t_i \le t^* \\ \mu_{1\mathrm{A}} & \text{if } t_i > t^* \end{cases} \qquad \mu_2(t_i) = \begin{cases} \mu_{2\mathrm{B}} & \text{if } t_i \le t^* \\ \mu_{2\mathrm{A}} & \text{if } t_i > t^*, \end{cases} \tag{2.2}$$

where A and B denote periods after and before the intervention, respectively. Assume further that the half-series means are random,

$$\begin{aligned} \mu_{1\mathrm{B}} &= \mu_1 + \gamma_1 \qquad & \mu_{2\mathrm{B}} &= \mu_2 + \gamma_3 \\ \mu_{1\mathrm{A}} &= \mu_1 + \gamma_2 \qquad & \mu_{2\mathrm{A}} &= \mu_2 + \gamma_4, \end{aligned} \tag{2.3}$$

where μ_1 and μ_2 are the overall means in units 1 and 2, respectively, assumed independent $\mathrm{N}(\mu, \sigma_\mu^2)$, μ is a constant, and $\gamma_1, \ldots, \gamma_4$ are the effects of time intervals within units, assumed independent $\mathrm{N}(0, \sigma_\gamma^2)$.

The measurement errors in Equation (2.1) could have a complicated correlation structure since there is the potential both for serial correlation of errors within each time series and for correlation of concurrent errors across time series, as might be caused, e.g., by common environmental influences affecting the two units in similar ways. The time series of differences between the responses in the two units has a simpler error structure.

2.2 Modeling the Time Series of Differences

Suppose we have n_1 observation times before the intervention and n_2 observation times after the intervention ($n = n_1 + n_2$). The difference between measured responses in the two units at time t_i is

$$d_i \equiv y_2(t_i) - y_1(t_i) = (\mu_{2*} - \mu_{1*}) + \delta \cdot I_{[t^*,\infty)}(t_i) + \epsilon_i, \tag{2.4}$$

where $\epsilon_i = \epsilon_{2i} - \epsilon_{1i}$ is the error in the difference measurement ($i = 1, \ldots, n$) and the * subscript is A or B, depending on the value of t_i.

Substituting the expressions from Equation (2.3) into Equation (2.4), we get

$$\begin{aligned} d_i(\text{before}) &= (\mu_2 - \mu_1) + (\gamma_3 - \gamma_1) + \epsilon_i \\ d_i(\text{after}) &= \delta + (\mu_2 - \mu_1) + (\gamma_4 - \gamma_2) + \epsilon_i, \end{aligned} \tag{2.5}$$

and therefore

$$\begin{aligned} \bar{d}_B &= \frac{1}{n_1} \sum_{i=1}^{n_1} d_i = (\mu_2 - \mu_1) + (\gamma_3 - \gamma_1) + \frac{1}{n_1} \sum_{i=1}^{n_1} \epsilon_i \\ \bar{d}_A &= \frac{1}{n_2} \sum_{i=n_1+1}^{n} d_i = (\mu_2 - \mu_1) + (\gamma_4 - \gamma_2) + \delta + \frac{1}{n_2} \sum_{i=n_1+1}^{n} \epsilon_i. \end{aligned} \tag{2.6}$$

Assume the measurement errors are multivariate normal with zero means and covariance matrix $\mathbf{\Sigma}$, reflecting possible first-order serial correlation,

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \frac{\sigma_d^2}{1-\rho_d^2} \begin{pmatrix} 1 & \rho_d & \rho_d^2 & \cdots & \rho_d^{n-1} \\ \rho_d & 1 & \rho_d & \cdots & \rho_d^{n-2} \\ \vdots & & & \cdots & \vdots \\ \rho_d^{n-1} & \rho_d^{n-2} & \cdots & \rho_d & 1 \end{pmatrix} \right), \tag{2.7}$$

where ρ_d is the autoregression coefficient and $\sigma_d^2/(1-\rho_d^2)$ is the variance of each ϵ_i.

The estimate used in paired intervention analysis is $\bar{d}_A - \bar{d}_B$. With the assumptions of this and the preceding section, we have

$$\begin{aligned} \mathrm{E}(\bar{d}_A - \bar{d}_B) &= \delta \\ \mathrm{var}(\bar{d}_A - \bar{d}_B) &= 4\sigma_\gamma^2 + \mathbf{a}'\mathbf{\Sigma}\mathbf{a}, \end{aligned} \tag{2.8}$$

where $\mathbf{a} = (\,(-1/n_1) \;\; \ldots \;\; (-1/n_1) \;\; (1/n_2) \;\; \ldots \;\; (1/n_2)\,)'$.

This formulation differs from the the two-sample test usually done in paired intervention analysis by virtue of its inclusion of a component of variation (σ_γ^2) due to changes in the mean response through time and its allowance for possible serial correlation of errors.

3. A TWO-STAGE ANALYSIS STRATEGY

This approach, which I will be called two-stage intervention analysis, uses the two original time series to estimate σ_γ^2 and the time series of differences to estimate σ_d^2. These estimates are then combined to allow statistical testing.

(1) The estimation of σ_γ^2 relies on the following observation. If

$$\bar{y}_{1\mathrm{B}} = \left(\sum_{i=1}^{n_1} y_1(t_i)\right)/n_1 \quad \text{and} \quad \bar{y}_{1\mathrm{A}} = \left(\sum_{i=n_1+1}^{n} y_1(t_i)\right)/n_2$$

and we assume that the errors in Equation (2.1) are mutually independent, then $\mathrm{var}(\bar{y}_{1*} \mid \mu_1) = \sigma_\gamma^2 + \sigma^2/n_1$, if $n_1 = n_2$.

(a) First, estimate σ^2 as the pooled variance of the residuals from the constant half-series means model,

$$\hat{\sigma}^2 = \frac{1}{2n_1 + 2n_2 - 4}\left[\sum_{i=1}^{n_1}\left\{(y_1(t_i) - \bar{y}_{1\mathrm{B}})^2 + (y_2(t_i) - \bar{y}_{2\mathrm{B}})^2\right\} + \sum_{i=n_1+1}^{n}\left\{(y_1(t_i) - \bar{y}_{1\mathrm{A}})^2 + (y_2(t_i) - \bar{y}_{2\mathrm{A}})^2\right\}\right]. \tag{3.1}$$

(b) Next, combine this estimated variance with the method-of-moments estimator of the variance of the two half-series means from unit 1 and the before mean from unit 2 to get

$$\hat{\sigma}_\gamma^2 = \frac{1}{3}\left[(\bar{y}_{1\mathrm{B}} - \bar{\bar{y}})^2 + (\bar{y}_{1\mathrm{A}} - \bar{\bar{y}})^2 + (\bar{y}_{2\mathrm{B}} - \bar{\bar{y}})^2\right] - \frac{\hat{\sigma}^2}{n_1}, \tag{3.2}$$

where $\bar{\bar{y}}$ is the mean of the three half-series means. If $\hat{\sigma}_\gamma^2$ is negative, set it to zero. (The inclusion of $\bar{y}_{2B}$ in the sample variance, which is strictly justifiable only when $\mu_1 \approx \mu_2$, and the use of the method-of-moments form of the estimator lead to improved operating characteristics of the test described below.)

(2) Estimate $\mathbf{\Sigma}$ from the time series of differences using an iterative approach. If $\mathbf{X}$ is an $n \times 2$ design matrix with a column of 1's and a column of the values of $I_{[t^*,\infty)}(t_i)$, $\mathbf{d} = (d_1, \ldots, d_n)'$, and $\boldsymbol{\beta} = (\beta_0 \;\; \beta_1)'$, then we can write

$$\begin{aligned} \mathrm{E}(\mathbf{d} \mid \mu_{1\mathrm{B}}, \mu_{1\mathrm{A}}, \mu_{2\mathrm{B}}, \mu_{2\mathrm{A}}) &= \mathbf{X}\boldsymbol{\beta} \\ \mathrm{var}(\mathbf{d} \mid \mu_{1\mathrm{B}}, \mu_{1\mathrm{A}}, \mu_{2\mathrm{B}}, \mu_{2\mathrm{A}}) &= \mathbf{\Sigma}. \end{aligned} \tag{3.3}$$

Comparing this to Equation (2.4), we see that $\beta_0 = \mu_{2\mathrm{B}} - \mu_{1\mathrm{B}}$ and $\beta_0 + \beta_1 = \mu_{2\mathrm{A}} - \mu_{1\mathrm{A}} + \delta$.

(a) In the first iteration, use $\hat{\beta}_0 = \bar{d}_B$ and $\hat{\beta}_1 = \bar{d}_A - \bar{d}_B$, where $\bar{d}_A$ and $\bar{d}_B$ are the mean between-unit differences after and before the intervention, respectively. In subsequent iterations, use the generalized least-squares estimates from (d) below.

(b) From the time series of residuals, $\mathbf{e} = \mathbf{d} - \mathbf{X}\hat{\boldsymbol{\beta}}$, estimate ρ_d, the first-order autoregression coefficient, in the usual way (cf., Fuller 1996). If a test of $\rho_d = 0$ is nonsignificant, set $\hat{\rho}_d$ equal to zero. Calculate $\hat{\sigma}_d^2 = \widehat{\mathrm{var}}(e_i) \cdot (1 - \hat{\rho}_d^2)$.

(c) Estimate $\mathbf{\Sigma}$ by substituting $\hat{\rho}_d$ and $\hat{\sigma}_d^2$ into the expression in Equation (2.7).

(d) Use generalized least squares to calculate

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{\Sigma}}^{-1}\mathbf{d}, \tag{3.4}$$

and repeat the above steps if $\hat{\boldsymbol{\beta}}$ is not sufficiently close to its value from the preceding iteration.

(3) Combine the results from steps 1 and 2 as follows. If $\hat{\delta} \equiv \bar{d}_\mathrm{A} - \bar{d}_\mathrm{B}$ is the estimate used in paired intervention analysis, Equation (2.8) gives us

$$\widehat{\mathrm{var}}(\hat{\delta}) = 4\hat{\sigma}_\gamma^2 + \mathbf{a}'\hat{\mathbf{\Sigma}}\mathbf{a}, \tag{3.5}$$

where $\hat{\sigma}_\gamma^2$ is from step 1 and $\hat{\mathbf{\Sigma}}$ is from step 2. We can then test the hypothesis that $\delta = 0$ by using the test statistic $\hat{\delta}/(\widehat{\mathrm{var}}(\hat{\delta}))^{1/2}$. An ad hoc approximation of the distribution of this statistic comes from reasoning that, if $\hat{\sigma}_\gamma^2 \gg \hat{\sigma}_d^2$, then the variance estimate is based largely on the three means in Equation (3.2) (2 d.f.), whereas, if $\hat{\sigma}_\gamma^2 \ll \hat{\sigma}_d^2$, the variance depends largely on $\hat{\sigma}_d^2$ ($n - 2$ d.f.). So calculate

$$\mathrm{d.f.} = 2p + (n-2)(1-p), \tag{3.6}$$

where $p = 4\hat{\sigma}_\gamma^2/(4\hat{\sigma}_\gamma^2 + \mathbf{a}'\hat{\mathbf{\Sigma}}\mathbf{a})$, and compute a p-value by comparing the observed test statistic to a $t_{\mathrm{d.f.}}$ distribution.

4. APPLICATION TO PAIRS OF CONTROL UNITS

Table 1 lists six sources of data representing paired control time series from studies of forested watersheds, lakes, and aquatic microcosms. It should be emphasized that the authors of these studies did not do paired intervention analysis. I will use their data (i) to illustrate the performance of paired intervention analysis and two-stage intervention analysis when there is no intervention effect (since none was imposed in these control series) and (ii) to get an idea of realistic parameter values for use in the simulations described in the next section.

Table 2 summarizes some analyses of these data sets. Even though there are no interventions here, the differences between early and late mean differences ($\hat{\delta}$) range from 5 to 92% (mean, 33%) of the overall mean response in each pair of units. For the Hubbard Brook data (source 1), the early versus late difference is even judged statistically significant by RIA.

The p-values from two-stage intervention analysis are generally higher— sometimes much higher—than those from RIA. In general, positive serial correlation of measurement errors and temporal variability of half-series means will inflate p-values compared to those obtained with paired intervention analysis. For source 5, significant negative serial correlation of the residual differences causes two-stage intervention analysis to give a slightly lower p-value than does RIA.

5. SIMULATIONS

I used the constant half-series means model of Section 2.1 to simulate time series having properties similar to those of the data sets listed in Table 1. Key parameters were expressed

Table 1. Sources of Data From Paired Ecological Units That Were Not Manipulated During the Observation Period

Source number	*Data*
1	Concentrations of Ca^{++} (in mg/L) in streamwater from two watersheds in the Hubbard Brook Experimental Forest from June 1965 to January 1966 before W2 was clearcut ($n = 29$, from Fig. 9 of Likens et al. (1970))
2	Concentrations of NO_3^- (in mg/L) in streamwater from two watersheds in southwestern Oregon for 5 years following patch cutting in one of the watersheds ($n = 88$, from Display 15.3 of Ramsey and Schafer (1997), reporting data of Harr, Fredriksen, and Rothacher)
3	Chlorophyll concentrations (in μg/L) in two unmanipulated lakes in the Experimental Lakes area, Lake 240 and Lake 302N, in 1971 ($n = 14$, from Figs. 2 and 3 of Schindler and Fee (1974))
4	Densities of copepods (individuals/L) in two control mesocosms in a study of effects of the insecticide Guthion ($n = 14$, from Table 6 of Giddings, Biever, Helm, Howick, and deNoyelles (1994))
5	Densities of cladocerans, as above ($n = 14$, from Table 7 of Giddings et al. (1994))
6	Densities of rotifers, as above ($n = 14$, from Table 8 of Giddings et al. (1994))

Table 2. Summary of Data From the Sources Listed in Table 1, Assuming a Hypothetical Intervention (With No Effect) at the Midpoint of Each Time Series. $\hat{\rho}$ is an estimate of the serial correlation of the residuals from the constant half-series means model [Eqs. (2.1) and (2.2)], pooled over the two units, with asterisks denoting values significantly different from zero; coefficients of variation are the ratios of $\hat{\sigma}_\mu$, $\hat{\sigma}_\gamma$, and $\hat{\sigma}$ to the mean response; and $\hat{\delta}$ is the estimate of the intervention effect, defined in Equation (3.5). The p-value from RIA is based on 5,000 random permutations.

			Coefficient of variation				*p-values*	
Source	*Mean response*	$\hat{\rho}$	*Unit means*	*Half series*	*Within series*	$\hat{\delta}$	*RIA*	*Two-stage IA*
1	1.615	0.068	0.183	0.188	0.117	0.116	0.026	0.858
2	0.129	0.668*	0.148	0.220	1.210	−0.040	0.490	0.779
3	3.230	0.365	0.076	0	0.207	−0.279	0.700	0.774
4	22.46	0.272	0	0.343	0.749	12.71	0.121	0.684
5	6.393	−0.300*	0	0.277	1.179	−5.857	0.599	0.547
6	53.0	−0.004	0	0	1.315	−2.845	0.802	0.724

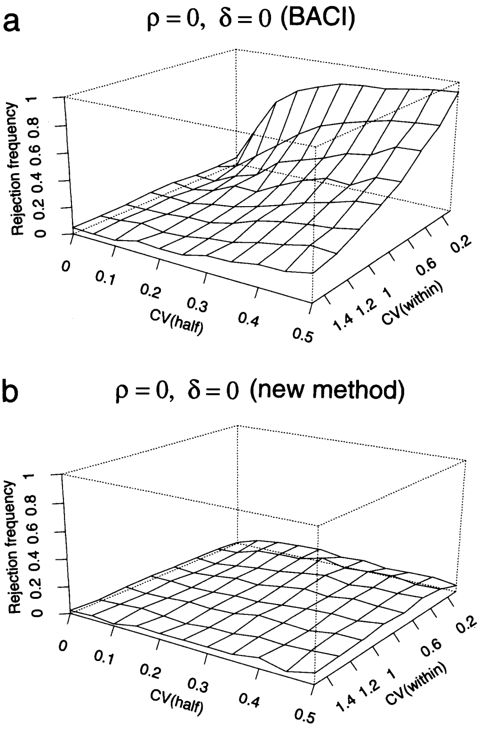
as coefficients of variation: $\text{CV}_{\text{mean}} = \sigma_\mu/\mu$, $\text{CV}_{\text{half}} = \sigma_\gamma/\mu$, and $\text{CV}_{\text{within}} = \sigma/\mu$, where μ is the overall mean, arbitrarily set to 1,000. Realistic ranges of values for these CVs were based on the estimates in Table 2.

In each run of the simulations for a particular set of parameter values, two time series of 30 observations each—15 before and 15 after a hypothetical intervention time—were generated according to the model in Section 2.1. The errors in each time series [Eq. (2.1)] had first-order autoregression coefficient ρ. Before and after differences were then compared with a two-sample t-test (the BACI protocol) and with two-stage intervention analysis. Five hundred such runs were done for each set of parameter values, and the results for each test were summarized as the proportion of runs for which the two-sided p-value was less than 0.05.

Figure 2 shows the results of the simulations for which $\delta = 0$ (no intervention effect) and $\rho = 0$ (no serial correlation of observations within each time series). The proportion of significant p-values from BACI analyses is close to the nominal 0.05 level only for small values of CV_{half} and large values of $\text{CV}_{\text{within}}$. As CV_{half} increases and $\text{CV}_{\text{within}}$ decreases, the rejection frequency rises dramatically above 0.05 to values as high as 0.92. Two-stage intervention analysis, on the other hand, has close to the nominal level (average rejection frequency over the grid is 0.065). The new method is somewhat conservative for small values of CV_{half} and anticonservative for large values.

The tendency for paired intervention analysis to reject too often is exacerbated when there is strong serial correlation ($\rho = 0.6$; see Fig. 3). The smallest rejection frequency on the grid is 0.29. Two-stage intervention analysis again has close to the correct level (average rejection frequency is 0.048).

When the intervention effects a 50% change in the mean response ($\delta = 500$), the rejection frequencies for paired intervention analysis are of course even larger than those in Figures 2 and 3. Of greater interest are the rejection frequencies for two-stage intervention analysis, shown in Figure 4. Power is generally poor; for about two-thirds of the grid points,


a ρ = 0, δ = 0 (BACI)

b ρ = 0, δ = 0 (new method)

Figure 2. Results of Simulations of Two Time Series of 30 Observations Each With No Intervention Effect and No Serial Correlation (See Text for Details). Graphs show the proportion of p-values that are less than 0.05 from (a) paired intervention analysis and (b) two-stage intervention analysis applied to the simulated data sets. Each point on the grids summarizes 500 data sets.

the rejection frequencies are less than 0.2. Only for small values of $\mathrm{CV}_{\mathrm{half}}$ and small values of $\mathrm{CV}_{\mathrm{within}}$ does the new method have anything approaching reasonable power to detect this large intervention effect (the maximum power is 0.80). As the serial correlation of observations increases, the general shape of the surface in Figure 4 is preserved, but the rejection frequencies decrease somewhat (e.g., the maximum power is 0.72 when $\rho = 0.6$).

The low power of two-stage intervention analysis is mostly a consequence of the method's accounting for variability of the half-series means through time. It might seem paradoxical that the power can be so low in some areas of the grid where the true value of σ_γ^2 is close to zero. In these situations, Equation (3.2) overestimates σ_γ^2, apparently because of the truncation of many values of $\hat{\sigma}_\gamma^2$ to zero. This overestimation, which happens especially when σ^2 is large, makes it harder for the test to reject the null hypothesis.

6. DISCUSSION

It is perhaps not surprising that some of the p-values from RIAs on pairs of control time series are fairly small (see Table 2). The null hypothesis being tested here is that the

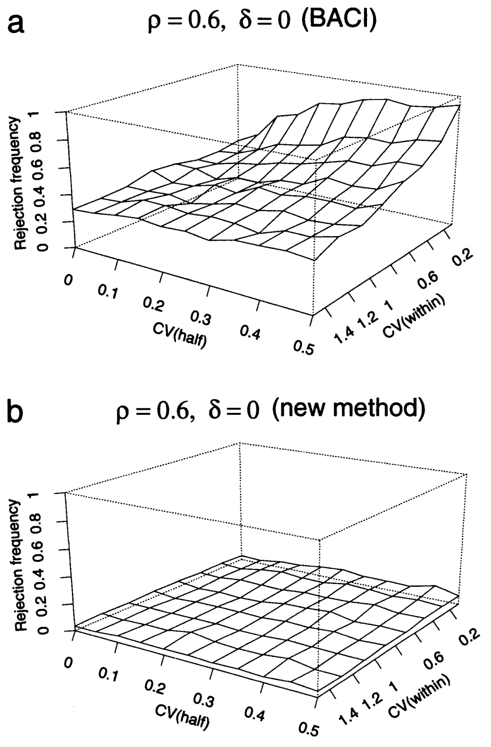


Figure 3. Results of Simulations With No Intervention Effect and Strong Positive Serial Correlation ($\rho = 0.6$). Details as in Figure 2 legend.

mean postintervention difference in response between manipulated and control units is identical to the mean preintervention difference. In nature, any two units are bound to differ in the trajectories of their responses through time, which will lead to a statistically significant comparison of early versus late differences if we take enough measurements and those measurements are assumed independent. This could explain, e.g., why Uddameri et al. (1995) found greater statistical significance with RIA on weekly than on monthly data.

There is evidence from other studies that RIA may frequently reject the null hypothesis for pairs of unmanipulated systems. Carpenter et al. (1989) applied RIA to a variety of responses measured in 12 lakes, 3 of which were manipulated and 9 of which were reference systems. Overall, 12.9% (14 of 108) pairwise comparisons of responses between reference lakes led to significant p-values—substantially larger than the 5% expected if the null hypotheses were actually true.

Stewart-Oaten et al. (1986) recommended transformation of time-series data to achieve preintervention trajectories that are as parallel as possible. This may stabilize the differences between the preintervention trajectories of the two units, but it does nothing to resolve the question of how much of the postintervention difference between units might be attributable to the intervention.

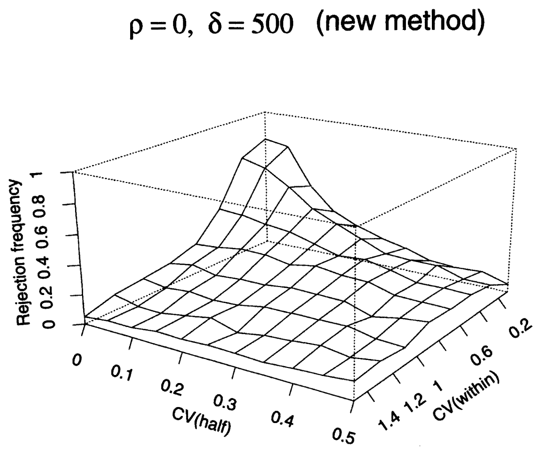


Figure 4. Results of Two-Stage Intervention Analysis Applied to Data Simulated With a Strong Intervention Effect ($\delta = 500$) and No Serial Correlation.

Two-stage intervention analysis attempts to improve on paired intervention analysis by adjusting for temporal variability of half-series means and serial correlation of measurement errors. The resulting test operates at close to the correct level (Figs. 2 and 3), but its power to detect even a large treatment effect is dismal (Fig. 4). This is a direct consequence of the extreme variability of a sample variance based on three numbers and the difficulty of using that variance to construct an unbiased estimate of σ_γ^2 [Eq. (3.2)].

A key assumption of the new method is that the time-interval effects [$\gamma_1, \ldots, \gamma_4$ in Eq. (2.3)] are independent. Proponents of paired intervention analysis, who are willing to assume that the time courses of the response in the two units would be parallel in the absence of intervention, might argue that $\text{corr}(\gamma_1, \gamma_3)$ and $\text{corr}(\gamma_2, \gamma_4)$ are close to unity. This would effectively eliminate the $4\sigma_\gamma^2$ term from Equation (2.8) and make the two-sample test used in paired intervention analysis more defensible, although the problem of serial correlation of errors remains.

The correlations of the γ's are not estimable from a single data set, but the large magnitudes of the $\hat{\delta}$'s for the pairs of control series in Table 2, which should be close to zero if parallelism holds, suggest substantial variability in the time courses of these responses. The best way to assess this variability, of course, is to include more than two ecological units in the study. Since it is often not possible to replicate the disturbance or intervention of interest, Underwood (1994) has advocated including multiple control units in the design, and others have suggested monitoring multiple sites near and far from the location of the intervention (Thomas, Mahaffey, Gore, and Watson 1978; Skalski and McKenzie 1982; Skalski and Robson 1992). This allows the putative intervention effect to be evaluated in the context of the natural variability of response trajectories in a set of similar units or locations.

7. CONCLUSIONS

Clearly, for data generated as in Section 2.1, a simple two-sample comparison of before versus after differences may reject the null hypothesis far too often, especially when there is temporal variation of half-series means and positive serial correlation of measurement errors. Two-stage intervention analysis attempts to improve on this situation using a crude estimate of temporal variation that can contribute enormous variability to the test statistic of interest and severely limit power. This seems an inevitable consequence of basing statistical inference on a design in which there is one control and one manipulated unit.

Large-scale experiments with little or no replication have been extremely influential in ecology (cf., Schindler 1973; Schindler and Fee 1974; Likens 1985; Carpenter, Chisholm, Krebs, Schindler, and Wright 1995), and some of the most famous examples have made their impact wholly in the absence of p-values (Likens et al. 1970; Schindler and Fee 1974). Few people looking at Figure 9 of Likens et al. (1970), e.g., would doubt that the elevation of streamwater nutrient levels in the deforested watershed, relative to the reference watershed, is due to the deforestation. That these results were presented without a p-value does not detract fom their impact.

Some ecological responses, like the export of nutrients from a watershed, can be measured only at a scale that may make true replication impossible. Analyses of the spatial and temporal dynamics of such responses can reveal much about the functioning of the system and can generate hypotheses that might be testable with a greater number of replicates at a smaller scale. If a p-value must be obtained for a study based on a single pair of units, the approach outlined here gives a more honest assessment, in my view, than do existing methods of paired intervention analysis.

ACKNOWLEDGMENTS

I thank M. Lucash and several anonymous reviewers for constructive comments on earlier versions of the manuscript and D. Derryberry, C. Pereira, and F. Ramsey for helpful discussions of this work.

[Received October 1998. Accepted December 1999.]

REFERENCES

Carpenter, S. R., Chisholm, S. W., Krebs, C. J., Schindler, D. W., and Wright, R. F. (1995), "Ecosystem Experiments," *Science,* 269, 324–327.

Carpenter, S. R., Frost, T. F., Heisey, D., and Kratz, T. K. (1989), "Randomized Intervention Analysis and the Interpretation of Whole-Ecosystem Experiments," *Ecology,* 70, 1142–1152.

Faith, D. P., Humphrey, C. L., and 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.

Fuller, W. A. (1996), *Introduction to Statistical Time Series* (2nd ed.), New York: Wiley.

Giddings, J. M., Biever, R. C., Helm, R. L., Howick, G. L., and deNoyelles, F. J., Jr. (1994), "The Fate and Effects of Guthion (Azinphos Methyl) in Mesocosms," in *Aquatic Mesocosms in Ecological Risk Assessment,* eds. R. L. Graney, J. H. Kennedy, and J. H. Rodgers, Boca Raton, FL: Lewis, pp. 469–495.

Hogg, I. D., and Williams, D. D. (1996), "Response of Stream Invertebrates to a Global-Warming Thermal Regime: An Ecosystem-Level Manipulation," *Ecology,* 77, 395–407.

Hurlbert, S. H. (1984), "Pseudoreplication and the Design of Ecological Field Experiments," *Ecological Monographs,* 54, 187–211.

Likens, G. E. (1985), "An Experimental Approach for the Study of Ecosystems," *Journal of Ecology,* 73, 381–396.

Likens, G. E., Bormann, F. H., Johnson, N. M., Fisher, D. W., and Pierce, R. S. (1970), "Effects of Forest Cutting and Herbicide Treatment on Nutrient Budgets in the Hubbard Brook Watershed-Ecosystem," *Ecological Monographs,* 40, 23–47.

Lydersen, E., Fjeld, E., and Gjessing, E. T. (1996), "The Humic Lake Acidification Experiment (HUMEX): Main Physico-Chemical Results After Five Years of Artificial Acidification," *Environment International,* 22, 591–604.

Ramsey, F. L., and Schafer, D. W. (1997), *The Statistical Sleuth: A Course in Methods of Data Analysis,* Belmont, CA: Duxbury Press.

Reitzel, J., Elwany, M. H. S., and Callahan, J. D. (1994), "Statistical Analyses of the Effects of a Coastal Power Plant Cooling System on Underwater Irradiance," *Applied Ocean Research,* 16, 373–379.

Roberts, E. A. (1993), "Seasonal Cycles, Environmental Change and BACI Designs," *Environmetrics,* 4, 209–232.

Schindler, D. W. (1973), "Experimental Approaches to Limnology—An Overview," *Journal of the Fisheries Research Board of Canada,* 30, 1409–1413.

Schindler, D. W., and Fee, E. J. (1974), "Experimental Lakes Area: Whole-Lake Experiments in Eutrophication," *Journal of the Fisheries Research Board of Canada,* 31, 937–953.

Schmitt, R. J., and Osenberg, C. W. (eds.) (1996), *Detecting Ecological Impacts: Concepts and Applications in Coastal Habitats,* San Diego: Academic Press.

Schroeter, S. C., Dixon, J. D., Kastendiek, J., Smith, R. O., and Bence, J. R. (1993), "Detecting the Ecological Effects of Environmental Impacts: A Case Study of Kelp Forest Invertebrates," *Ecological Applications,* 3, 331–350.

Skalski, J. R., and McKenzie, D. H. (1982), "A Design for Aquatic Monitoring Programs," *Journal of Environmental Management,* 14, 237–251.

Skalski, J. R., and Robson, D. S. (1992), *Techniques for Wildlife Investigations: Design and Analysis of Capture Data,* San Diego: Academic Press.

Smith, E. P., Orvos, D. R., and Cairns, J., Jr. (1993), "Impact Assessment Using the Before-After-Control-Impact (BACI) Model: Concerns and Comments," *Canadian Journal of Fisheries and Aquatic Sciences,* 50, 627–637.

Stewart-Oaten, A. (1996), "Problems in the Analysis of Environmental Monitoring Data," in *Detecting Ecological Impacts: Concepts and Applications in Coastal Habitats,* eds. R. J. Schmitt and C. W. Osenberg, San Diego: Academic Press, pp. 109–131.

Stewart-Oaten, A., Murdoch, W. W., and Parker, K. R. (1986), "Environmental Impact Assessment: 'Pseudoreplication' in Time?" *Ecology,* 67, 929–940.

Stout, R. J., and Rondinelli, M. P. (1995), "Stream-Dwelling Insects and Extremely Low Frequency Electromagnetic Fields: A Ten-Year Study," *Hydrobiologia,* 302, 197–213.

Thomas, J. M., Mahaffey, J. A., Gore, K. L., and Watson, D. G. (1978), "Statistical Methods Used to Assess Biological Impact at Nuclear Power Plants," *Journal of Environmental Management,* 7, 269–290.

Uddameri, V., Norton, S. A., Kahl, J. S., and Scofield, J. P. (1995), "Randomized Intervention Analysis of the Response of the West Bear Brook Watershed, Maine, to Chemical Manipulation," *Water, Air, and Soil Pollution,* 79, 131–146.

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. (1994), "On Beyond BACI: Sampling Designs That Might Reliably Detect Environmental Disturbances," *Ecological Applications,* 4, 3–15.

Vose, F. E., and Bell, S. S. (1994), "Resident Fishes and Macrobenthos in Mangrove-Rimmmed Habitats: Evaluation of Habitat Restoration by Hydrologic Modification," *Estuaries,* 17, 585–596.

Wallace, J. B., Eggert, S. L., Meyer, J. L., and Webster, J. R. (1997), "Multiple Trophic Levels of a Forest Stream Linked to Terrestrial Litter Inputs," *Science,* 277, 102–104.