Ecology, 83(6), 2002, pp. 1752–1761

ON REJECTION RATES OF PAIRED INTERVENTION ANALYSIS

PAUL A. MURTAUGH[1]

Department of Statistics, Oregon State University, Corvallis, Oregon 97331 USA

Abstract. Before–After-Control–Impact (BACI) analysis and randomized intervention analysis (RIA) are commonly applied to time series of response measurements obtained from two ecological units, one of which is subjected to an intervention at some intermediate time. Positive results from the analyses are interpreted as evidence of a potentially meaningful association between the intervention and the response. Applied to 154 pairs of actual ecological time series, RIA done at the 5% level rejected the hypothesis of no association 20% of the time when both units were in fact undisturbed, and 30% of the time when one of the two units had received an intervention. Correction for first-order serial autocorrelation in the time series of between-unit differences reduced these rejection frequencies to 15% and 28%, respectively. A two-stage analysis method that attempts to adjust for temporal variability of early and late response means failed to find an association in any of the pairs of ''control'' units, and found evidence of an association in only 14–15% of the pairs in which one unit was disturbed.

These results suggest that RIA (and BACI analysis) greatly overstate the evidence for associations of interventions with ecological responses, and that attempts to modify these methods to account for temporal variability of response trajectories result in tests with very limited power. It may be that the best strategy for interpreting data from BACI designs is to rely on graphical presentation, expert judgment, and common sense, rather than *P* values derived from hypothesis tests of questionable validity.

Key words: before–after-control–impact (BACI) design; environmental impact assessment; intervention effects; randomized intervention analysis (RIA); serial correlation; time series; two-stage intervention analysis; Type I error rate.

INTRODUCTION

In the Before–After-Control–Impact (BACI) design, a pair of ecological units is monitored over time, with regular measurements of some response of interest. At some point during the monitoring, an intervention is applied, or a disturbance is noted, in one of the two units. The post-intervention responses in the disturbed unit are compared to the pre-intervention responses in that unit, and to the responses measured in the undisturbed ''control'' unit. Large differences provide possible evidence of an effect of the intervention. This design has been used, without statistical analysis, in some famous and influential ecological experiments (Likens et al. 1970, Schindler and Fee 1974).

While statistical inference based on a single pair of units seems impracticable, some authors have suggested calculating the differences in the response between units for each observation time, and then comparing the pre-intervention to the post-intervention differences with either a two-sample *t* test (the original BACI analysis; Stewart-Oaten et al. 1986) or a randomization test (randomized intervention analysis, or RIA; Carpenter et al. 1989). Although this approach has been criticized on several grounds (Hurlbert 1984, Underwood 1992, 1994, Smith et al. 1993, Conquest 2000, Murtaugh 2000), it seems to have achieved currency in the ecological literature (Faith et al. 1991, Roberts 1993, Schroeter et al. 1993, Reitzel et al. 1994, Vose and Bell 1994, Stout and Rondinelli 1995, Uddameri et al. 1995, Hogg and Williams 1996, Lydersen et al. 1996, Schmitt and Osenberg 1996, Wallace et al. 1997, Johnston et al. 1999, del Rosario and Resh 2000, Keough and Quinn 2000, Solazzi et al. 2000, Wardell-Johnson and Williams 2000).

Recently I argued that BACI analysis and RIA are flawed because they ignore possible serial correlation of the between-unit differences, and they assume that the trajectories of the response would have been exactly parallel in the absence of the intervention (Murtaugh 2000). Stated another way, it is assumed that the difference between early and late responses would be the same in all unmanipulated units. Statistical simulations suggested that rejection frequencies based on this approach are probably much too high. I proposed an alternative, two-stage method that attempts to address these problems but appears to have very limited power.

Since it is never clear how effectively simulations capture the complexities of real data, in this paper I apply these analysis methods to actual data collected in ways consistent with the BACI design. I compare the performance of RIA to that of newer methods that adjust for serial correlation of between-unit differences and temporal variability of the response trajectories in the two units.

Manuscript received 16 November 2000; revised 8 May 2001; accepted 7 July 2001; final version received 10 September 2001.
[1] E-mail: murtaugh@stat.orst.edu

Methods and Materials

Statistical methods

BACI (Before–After-Control–Impact) analysis and RIA (randomized intervention analysis) have been described previously (Stewart-Oaten et al. 1986, Carpenter et al. 1989). In my implementation of RIA, two-sided P values were based on 20 000 permutations of the original time series of between-unit differences.

The newer methods are from Murtaugh (2000). *Two-stage intervention analysis* can be summarized as follows.

1) For the two original time series of observed responses:

a) Calculate the four half-series means for the two time series (i.e., control before, control after, treated before, treated after).

b) Use three of these means—control before, control after, and treated before—to estimate a component of variation reflecting the effects of time periods within units.

2) For the time series of between-unit differences:

a) Use generalized least squares to fit a model of constant half-series means plus errors having first-order serial autocorrelation. That is, assume the pre-intervention differences have one mean, and the post-intervention differences have another, with serially correlated errors added to the means to yield the observed differences.

b) Test the hypothesis that the first-order autoregression coefficient is zero. If the hypothesis is not rejected, revert to a model that assumes independent errors.

3) Using variance estimates from (1) and (2), test the hypothesis that the mean post-intervention difference is equal to the mean pre-intervention difference. A more detailed description of this methodology can be found in Appendix A.

I analyzed real data sets using (a) RIA; (b) adjustment for serial correlation of between-unit differences; (c) adjustment for temporal variability of half-series response means; and (d) adjustment for both serial correlation and temporal variability.

The data

I worked with data collected from 13 sources, which yielded 154 sensible comparisons of two time series (see Appendix B). These data were identified from literature searches and discussions with ecologists. Included are (a) data sets to which BACI analysis or RIA was applied by the original authors; (b) data generated by BACI-like designs, but not analyzed with the usual two-sample test; and (c) data from unmanipulated ''reference'' units that were identified by the authors as ecologically similar. While the 13 sources of data are not a random sample from any clearly defined population, it is hoped that they are representative of the kinds of studies to which BACI analysis and RIA are usually applied.

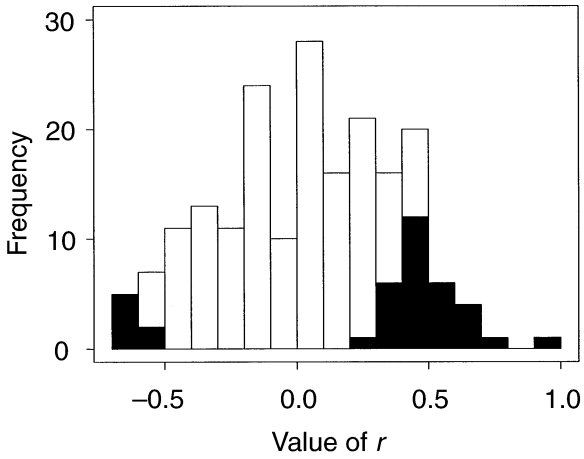

FIG. 1. Frequency distribution of estimates of the first-order autoregression coefficient for the 154 time series of between-unit differences. Solid areas represent estimates that are significantly different from zero ($P < 0.05$).

Ninety-three of the paired comparisons are from systems in which one of the two units was manipulated, and the other 61 comparisons involve pairs of undisturbed ''control'' units. The intent was to see how often the statistical methods reject true null hypotheses, as well as how powerful they are in the detection of possible intervention effects.

For pairs of unmanipulated units, a hypothetical intervention was assumed to have occurred at the midpoint of the time series: observation times up to and including the median time were considered to be ''pre-intervention.'' The pre-intervention differences were then compared to the post-intervention differences with the methods described in the preceding section. (Not all BACI studies have equal numbers of pre- and post-intervention observation times, but the mean, study-specific proportion of pre-intervention observation times in Table 1 is 0.49.) See Appendix C for additional details on the implementation of these analyses.

Results and Discussion

Serial correlation

Fig. 1 is a histogram of the estimated values of the first-order autoregression coefficient for the time series of between-unit differences. Of the 154 paired comparisons, 26 comparisons yielded positive estimates that were statistically different from zero ($P < 0.05$), and 7 comparisons yielded statistically significant negative values.

Most of the positive correlations seem to arise from a strong seasonal cycle of the response in one unit, unmatched by a cycle of comparable magnitude in the other unit (e.g., see Fig. 2). The between-unit differences retain a seasonal pattern, which, in the context

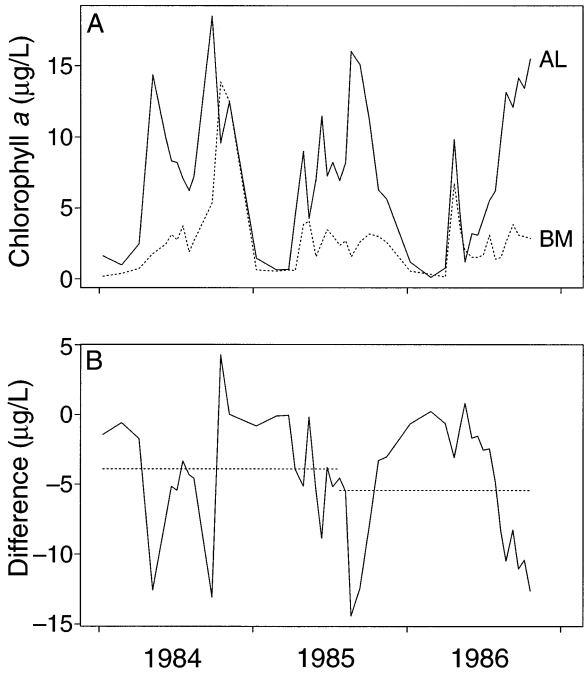

FIG. 2. Example of paired time series showing positive serial correlation of between-unit differences. (A) Time series of chlorophyll *a* measurements from two lakes (Allequash [AL] and Big Muskellunge [BM]) in Wisconsin, USA, over a three-year period (North Temperate Lakes Long Term Ecological Research Program; see Appendix B: data set A). (B) Time series of between-lake differences, with means of the early and late observations shown by dashed lines. The estimated first-order autoregression coefficient for the residuals from the constant half-series means model is 0.47 (P = 0.001).

of the constant half-series means model, is interpreted as positive serial correlation.

The negative correlations come from Giddings et al. (1994), the studies of zooplankton densities in mesocosms. These apparently result from asynchronous seasonal peaks of zooplankton densities in pairs of mesocosms. The between-unit differences are large and alternating in sign near these spikes in zooplankton numbers, leading to negative estimates of the serial autoregression coefficient. Fig. 3 shows an example of such a comparison. Analysis of the densities on the original scale yields a time series of differences with strong negative serial correlation. When the data are log-transformed, effectively ''expanding'' the early and late-season differences and reducing the influence of the mid-season differences, the evidence for serial correlation disappears (Fig. 3C).

The above observations point to the inadequacy of the constant half-series means model for the differences, which interprets all departures from the constant means as random error. Instead of being modeled as part of the ''signal,'' as is usually done in time-series analyses, seasonal variations in the between-unit differences are modeled as serially correlated errors added to the half-series means. It is obvious that many of the time series of differences are nonstationary (e.g., see Fig. 3B), so that application of the first-order autoregression model is of questionable value. More complicated models, including a seasonal component, would be needed for proper statistical treatment of these time series.

Rejection frequencies

Table 1 shows the results of applying RIA and two-stage intervention analysis to the 154 pairs of time series. Because of the questionable value of applying the first-order autoregression model to the time series of differences, Table 1 focuses on two-stage intervention analysis *without* adjustment for serial correlation of the between-unit differences.

Inspection of the results for pairs of ''control'' units in Table 1 shows that the P values from RIA tend to be much smaller than the P values from two-stage intervention analysis. The rejection frequencies for these

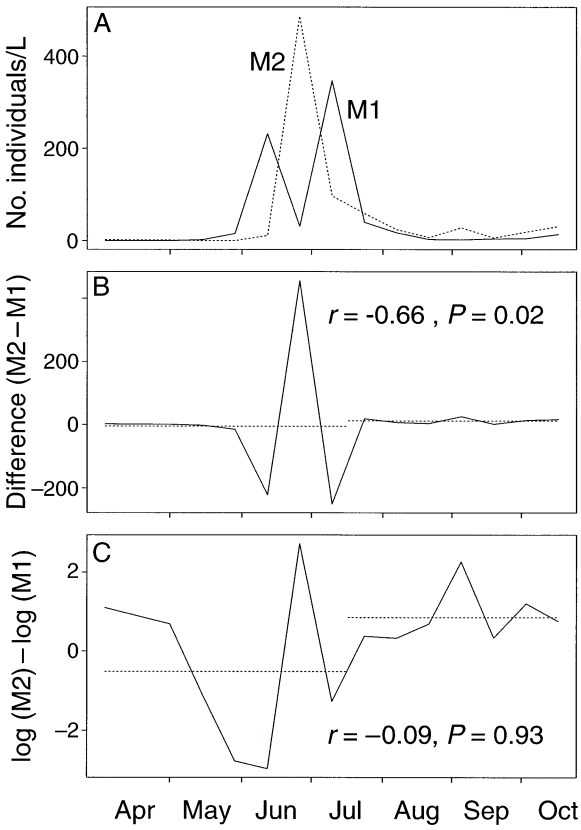

FIG. 3. Example of paired time series showing apparently negative serial correlation of between-unit differences. (A) Rotifer densities in two control mesocosms, M1 and M2 (Giddings et al. 1994) (see Appendix B: data set E). (B) Differences in densities between mesocosms. (C) Differences in log-transformed densities between microcosms (with 1 added to the densities before taking the logarithm). In (B) and (C) the dashed lines indicate half-series means.

TABLE 1. Results of applying randomized intervention analysis (RIA) and two-stage analysis *without* adjustment for serial correlation to the 154 pairs of time series, organized by data set.

Data set, source(s) and subject	Comparison†	Intervention?	*n*‡	*r*§	*P* value	
					RIA	Two-stage
A) Chlorophyll *a* in seven lakes in Wisconsin, USA (North Temperate Lakes LTER Program)	AL–BM	0	47	**0.47**	0.232	0.709
	AL–CB	0	45	0.05	0.069	0.376
	AL–CR	0	49	**0.45**	0.348	0.823
	AL–SP	0	49	**0.61**	0.466	0.810
	AL–TB	0	49	0.09	**0.007**	0.427
	AL–TR	0	48	**0.49**	0.638	0.875
	BM–CB	0	46	0.10	0.104	0.539
	BM–CR	0	48	**0.43**	0.655	0.775
	BM–SP	0	48	**0.54**	0.602	0.595
	BM–TB	0	47	0.09	**0.021**	0.576
	BM–TR	0	48	**0.39**	0.272	0.253
	CB–CR	0	47	0.07	0.081	0.561
	CB–SP	0	46	0.03	0.072	0.532
	CB–TB	0	46	0.19	0.877	0.886
	CB–TR	0	46	0.06	0.060	0.503
	CR–SP	0	51	**0.69**	0.487	0.730
	CR–TB	0	49	0.05	**0.016**	0.585
	CR–TR	0	51	**0.28**	0.198	0.752
	SP–TB	0	49	0.11	**0.016**	0.562
	SP–TR	0	50	**0.40**	0.354	0.353
	TB–TR	0	49	0.08	**0.013**	0.546
B) Surface pH in seven lakes in Wisconsin, USA (North Temperate Lakes LTER Program)	AL–BM	0	37	0.09	0.738	0.880
	AL–CB	0	35	**0.48**	0.915	0.996
	AL–CR	0	38	0.18	0.891	0.993
	AL–SP	0	38	−0.02	0.541	0.542
	AL–TB	0	38	**0.48**	0.221	0.961
	AL–TR	0	37	−0.13	0.714	0.721
	BM–CB	0	36	**0.53**	0.977	0.999
	BM–CR	0	37	**0.38**	0.94	0.996
	BM–SP	0	38	0.12	0.351	0.796
	BM–TB	0	38	**0.49**	0.075	0.958
	BM–TR	0	37	0.10	0.868	0.977
	CB–CR	0	36	**0.36**	0.324	0.946
	CB–SP	0	36	0.24	0.341	0.977
	CB–TB	0	37	0.19	**0**	0.593
	CB–TR	0	35	**0.52**	0.696	0.989
	CR–SP	0	37	−0.07	0.383	0.967
	CR–TB	0	38	**0.36**	**0.037**	0.905
	CR–TR	0	37	0.25	0.835	0.994
	SP–TB	0	37	0.22	**0.042**	0.962
	SP–TR	0	36	0.28	0.074	0.074
	TB–TR	0	36	**0.58**	0.052	0.957
C) Carpenter et al. (1987): effects of fish transplants on plankton in three lakes in Wisconsin, USA	zoo, PA–PT	1	14, 15	0.36	**0.036**	0.452
	zoo, PA–TU	1	14, 15	0.13	**0.002**	0.398
	zoo, PA–PT, pre	0	14	0.08	0.165	0.167
	zoo, PA–TU, pre	0	14	−0.04	0.986	0.995
	pct, PA–PT	1	14, 15	**0.43**	**0.002**	0.158
	pct, PA–TU	1	14, 15	0.42	**0.002**	0.623
	pct, PA–PT, pre	0	14	−0.17	**0.045**	0.210
	pct, PA–TU, pre	0	14	−0.16	0.051	0.708
	phyto, PA–PT	1	13, 13	0.41	0.302	0.259
	phyto, PA–TU	1	13, 13	0.3	**0.001**	0.593
	phyto, PA–PT, pre	0	13	0.25	0.717	0.686
	phyto, PA–TU, pre	0	13	0.24	0.167	0.534
	pp, PA–PT	1	42, 44	**0.64**	**0**	0.680
	pp, PA–TU	1	42, 44	**0.56**	0.168	0.912
	pp, PA–PT, pre	0	42	**0.35**	0.866	0.954
	pp, PA–TU, pre	0	42	**0.45**	**0.019**	0.256
D) Gerard et al. (1992): spider numbers, with and without deltamethrin	13–24	0	10, 7	−0.12	0.884	0.93
	13–32	0	10, 7	−0.11	0.627	0.864
	24–32	0	10, 7	−0.18	0.787	0.779
	13–11	1	10, 7	−0.23	0.137	0.311
	13–14	1	10, 7	−0.12	0.167	0.173
	13–16	1	10, 7	0.38	0.207	0.54
	13–12	1	10, 7	−0.09	**0**	**0.001**
	13–15	1	10, 7	0.2	**0.001**	**0**
	24–22	1	10, 7	−0.4	**0.005**	**0.004**

TABLE 1. Continued.

Data set, source(s) and subject	Comparison†	Intervention?	*n*‡	*r*§	*P* value	
					RIA	Two-stage
	24–25	1	10, 7	−0.09	**0.025**	**0.023**
	24–23	1	10, 7	−0.28	**0**	**0**
	24–26	1	10, 7	−0.14	**0**	**0.001**
	24–21	1	10, 7	−0.25	**0**	**0.02**
	32–33	1	10, 7	0.1	**0.007**	**0.009**
	32–31	1	10, 7	−0.34	**0**	0.054
	32–34	1	10, 7	−0.32	**0**	**0**
	32–35	1	10, 7	−0.02	**0**	**0**
E) Giddings et al. (1994): insecticide effects on densities of zooplankton	cop, 0–0	0	7, 7	0.2	0.122	0.612
	cop, 0–1	1	7, 7	−0.25	0.801	0.938
	cop, 0–1	1	7, 7	0	0.258	0.625
	cop, 0–2	1	7, 7	0.02	0.949	0.965
	cop, 0–2	1	7, 7	**−0.64**	0.878	0.96
	cop, 0–3	1	7, 7	0.34	0.090	0.539
	cop, 0–3	1	7, 7	−0.44	0.594	0.871
	cop, 0–4	1	7, 7	−0.3	0.445	0.645
	cop, 0–4	1	7, 7	0.06	**0.015**	0.492
	cop, 0–5	1	7, 7	−0.18	**0.021**	0.212
	cop, 0–5	1	7, 7	0.01	**0.037**	0.359
	cop, 0–1	1	7, 7	−0.26	0.349	0.766
	cop, 0–1	1	7, 7	−0.1	0.67	0.867
	cop, 0–2	1	7, 7	−0.01	**0**	0.681
	cop, 0–2	1	7, 7	−0.46	0.287	0.662
	cop, 0–3	1	7, 7	0.35	0.273	0.709
	cop, 0–3	1	7, 7	−0.49	0.627	0.817
	cop, 0–4	1	7, 7	−0.45	0.820	0.871
	cop, 0–4	1	7, 7	0.24	0.172	0.655
	cop, 0–5	1	7, 7	−0.30	**0.013**	0.163
	cop, 0–5	1	7, 7	0.1	0.117	0.408
	cld, 0–0	0	7, 7	−0.35	0.613	0.474
	cld, 0–1	1	7, 7	−0.07	0.412	0.761
	cld, 0–1	1	7, 7	−0.3	0.968	1
	cld, 0–2	1	7, 7	0	0.119	0.66
	cld, 0–2	1	7, 7	−0.19	0.592	0.842
	cld, 0–3	1	7, 7	0.08	0.604	0.794
	cld, 0–3	1	7, 7	−0.16	0.052	0.494
	cld, 0–4	1	7, 7	0.15	0.897	0.951
	cld, 0–4	1	7, 7	−0.16	0.833	0.833
	cld, 0–5	1	7, 7	−0.01	0.108	0.633
	cld, 0–5	1	7, 7	−0.19	0.733	0.439
	cld, 0–1	1	7, 7	−0.36	0.484	0.419
	cld, 0–1	1	7, 7	−0.33	0.316	0.257
	cld, 0–2	1	7, 7	−0.47	0.907	0.913
	cld, 0–2	1	7, 7	−0.36	0.526	0.398
	cld, 0–3	1	7, 7	−0.54	0.318	0.311
	cld, 0–3	1	7, 7	0.07	0.154	0.6
	cld, 0–4	1	7, 7	−0.13	0.930	0.591
	cld, 0–4	1	7, 7	−0.44	0.339	0.307
	cld, 0–5	1	7, 7	−0.5	0.903	0.936
	cld, 0–5	1	7, 7	−0.12	0.657	0.471
	rot, 0–0	0	7, 7	**−0.66**	0.813	0.850
	rot, 0–1	1	7, 7	**−0.63**	0.558	0.535
	rot, 0–1	1	7, 7	−0.3	0.653	0.495
	rot, 0–2	1	7, 7	**−0.60**	0.912	0.891
	rot, 0–2	1	7, 7	**−0.6**	0.411	0.533
	rot, 0–3	1	7, 7	0.06	**0.009**	**0.046**
	rot, 0–3	1	7, 7	−0.34	**0.045**	0.142
	rot, 0–4	1	7, 7	−0.5	0.999	0.766
	rot, 0–4	1	7, 7	−0.55	0.595	0.663
	rot, 0–5	1	7, 7	**−0.62**	0.278	0.296
	rot, 0–5	1	7, 7	−0.52	0.368	0.308
	rot, 0–1	1	7, 7	−0.31	0.488	0.449
	rot, 0–1	1	7, 7	−0.52	0.935	0.912
	rot, 0–2	1	7, 7	−0.3	0.223	0.241
	rot, 0–2	1	7, 7	−0.4	0.994	0.789
	rot, 0–3	1	7, 7	**−0.58**	0.862	0.793
	rot, 0–3	1	7, 7	−0.19	0.154	0.235
	rot, 0–4	1	7, 7	−0.28	0.989	0.491

TABLE 1. Continued.

Data set, source(s) and subject	Comparison†	Intervention?	*n*‡	*r*§	*P* value RIA	*P* value Two-stage
	rot, 0–4	1	7, 7	−0.43	0.881	0.875
	rot, 0–5	1	7, 7	−0.32	0.766	0.444
	rot, 0–5	1	7, 7	−0.49	0.999	0.818
F) Moring and Lantz (1975): numbers of salmonids before and after patch cutting and clearcutting	ch, C–T1	1	7, 5	−0.1	0.39	0.699
	ch, C–T2	1	7, 7	−0.17	0.314	0.303
	ct, C–T1	1	4, 7	0.03	0.152	0.69
	ct, C–T2	1	4, 9	−0.11	**0.001**	**0.005**
G) Moring (1975): max. daily temperature in streams in (F) above	all	1	7, 8	0.32	**0.007**	**0.007**
H) Likens et al. (1970): Ca^{++} concentrations in streams in two New Hampshire, USA, watersheds	all	1	145	**0.95**	**0**	**0**
	pre-int	0	29	0.01	**0.025**	0.79
I) Smith et al. (1993): fish abundance upstream and downstream before and after power-plant operation	all	1	80	0.15	0.243	0.532
	pre-int	0	56	0.16	0.675	0.684
J) Hurd et al. (1996): stonefly densities before and after diflubenzuron canopy treatment	all	1	29	0.08	**0.039**	0.164
	pre-int	0	18	0.05	0.669	0.684
K) Davis (1989): fire effect on Na^{+} concentration in streams	all	1	72	−0.06	0.096	0.130
	pre-int	0	23	−0.11	0.958	0.473
L) Crome et al. (1996): logging effects on numbers of rats	all	1	16	−0.25	0.928	0.981
M) Arsenault and Payette (1997): fire effects on tree-ring chronology	all	1	25	**0.80**	0.192	0.323
	pre-int	0	9	0.10	**0.021**	0.182

Notes: Statistically significant values of *r* and *P* values less than 0.05 are in boldface. For some pairs of time series involving interventions, analyses were done both on the full data set and on the pre-intervention data (mimicking two control series).

† See Appendix B for explanation of abbreviations.

‡ The number of observation times (pre-intervention [pre] and post-intervention sample sizes are shown for pairs in which one unit received an intervention).

§ The estimated first-order autoregression coefficient for the time series of between-unit differences.

methods, along with two other variants of two-stage intervention analysis, are summarized in Table 2.

When there is no intervention, RIA rejects the null hypothesis for 20% of the comparisons, even though the nominal testing level is 5% (Table 2). Adjustment for serial correlation reduces the rejection frequency to 15%. When temporal variability of half-series means is adjusted for, the null hypothesis is rejected for *none* of the 61 pairs of ''control'' units.

For comparisons in which one of the two units was disturbed, RIA rejects the null hypothesis 30% of the time. Adjustment for serial correlation reduces the rejection frequency slightly, and adjustment for temporal variability of half-series means results in rejections for only 14–15% of the comparisons.

Fig. 4 summarizes these results graphically. Obviously, adjustment for temporal variability of half-series means can substantially inflate the *P* values from RIA. The largest disparities (upper left-hand corners of plots) occur when there are large differences among the three half-series means used to calculate the temporal component of variation (control before, control after, and treated before).

TABLE 2. Rejection frequencies for the four different analysis methods (proportion of times the null hypothesis is rejected is shown in parentheses).

Intervention?	No. of data sets	No. with $P < 0.05$ RIA†	Corr.‡	Means§	Corr. and means
No	61	12 (0.20)	9 (0.15)	0 (0)	0 (0)
Yes	93	28 (0.30)	26 (0.28)	14 (0.15)	13 (0.14)

† RIA = randomized intervention analysis.

‡ Corr. = adjustment for serial correlation of between-unit differences.

§ Means = adjustment for temporal variability of half-series means (see *Methods and materials: Statistical methods*).

CONCLUSIONS

When applied to time series of responses from pairs of undisturbed ecological units, randomized intervention analysis done at the 5% level detected an association of the response with a hypothetical intervention 20% of the time (Table 2). This indicates that RIA rejects the null hypothesis much too often when there is no intervention effect. This rejection rate is not that different from the 30% frequency of positive results for paired time series in which one of the units did receive an intervention.

There are many practical problems with the application of RIA and BACI (Before–After-Control–Impact) analysis—including the effects of serial corre-

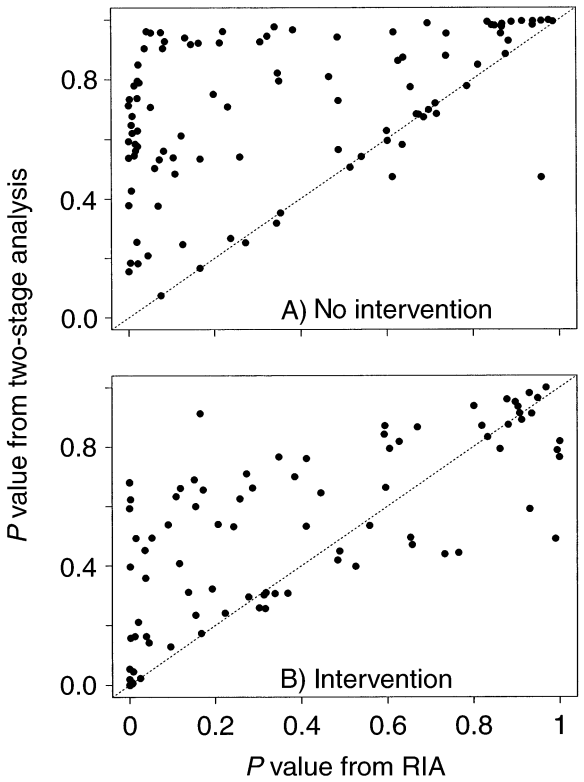

FIG. 4. *P* value from two-stage analysis with adjustment for temporal variability of half-series means (but not for serial correlation) vs. *P* value from randomized intervention analysis (RIA): (A) for pairs of ''control'' units and (B) for pairs in which one unit was manipulated. The dashed line indicates equality of the two *P* values.

lation, the choice of transformation (if parametric tests are used), and the possibility of delayed or transient effects of the intervention. Some of these problems could be alleviated by more sophisticated statistical modeling of the time series of between-unit differences (e.g., see Box and Tiao 1975). But, in my view, even the use of better statistical models would not eliminate the key problem in the interpretation of BACI-like analyses—namely, the unreasonableness of the assumption that the trajectories of the response in the two units would have been exactly parallel in the absence of the intervention.

The method presented by Murtaugh (2000) attempts to adjust for temporal variability of the response trajectories in the two units, but it does so by assuming that a sample variance based on the control-before, control-after and treated-before means is representative of the variability of half-series means that one could expect within a particular unit. Consequently, it will be very difficult to establish statistical significance for any pair of units in which the baseline mean responses differ substantially between units.

The challenge of obtaining efficient, unbiased estimates of the temporal variance component resulted in very low power of the two-stage approach in simulations (Murtaugh 2000). Ninety-three of the paired time series in the current study involved an intervention applied to one of the units, but, of course, there is no way of knowing how many of these interventions had nonzero effects. Unless many of the interventions were without effect, the 14–15% rejection frequency of the two-stage approach is consistent with the low power seen in simulations.

Proponents of BACI analysis and RIA might argue that concern with variability of response trajectories within or across ecological units is misguided: ''[I]mpact studies are usually concerned with effects at a particular place and time, not with generalizations: they are analagous to asking not whether smoking causes cancer but whether it caused a particular smoker's cancer'' (Stewart-Oaten 1996:126–127). I would suggest that many users of BACI analysis and RIA do not adhere to this strict interpretation of their results, and, in any case, the conclusion of an association between the intervention and the response—even at a particular time and place—depends on an untestable assumption about how the treated unit would have behaved in the absence of the intervention. Whatever view one has of the validity and appropriateness of BACI-like analyses, it is clear from the work described here that the method results in many false positives, i.e., it very frequently detects associations where none exist.

In my view, it is probably impossible to devise honest, powerful tests of the association of an intervention with the response in BACI designs, unless multiple ecological units can be observed, as suggested by Underwood (1992, 1994). Since this sort of replication is impracticable for units as large and unwieldy as those commonly studied with the BACI approach (Carpenter 1990), it may be best to resist the temptation and/or the editorial pressure to attempt statistical inference from such data. Instead, graphical displays, expert opinion, and common sense can be summoned to make plausible interpretations of the paired time series, and smaller-scale experiments in which replication is possible may be devised to explore the relationships and mechanisms suggested by the larger-scale observations.

ACKNOWLEDGMENTS

For assistance in identifying and gathering appropriate data sets, I thank the following members of Statistics 535: M. Baumgartner, M. Cavanaugh, R. Coshow, J. Dambacher, D. Jensen, R. Keim, S. McDowell, V. Monleon, G. Rosales, and J.-S. Taur. I am grateful to P. Guttorp and the National Research Center for Statistics and the Environment at the University of Washington for support during the writing of this manuscript, and I thank two anonymous reviewers for many constructive criticisms.

LITERATURE CITED

Arseneault, D., and S. Payette. 1997. Landscape change following deforestation at the arctic tree line in Quebec, Canada. Ecology **78**:693–706.

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

Carpenter, S. R. 1990. Special feature: statistical analysis of ecological response to large-scale perturbations. Ecology **71**:2037.

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

Carpenter, S. R., J. F. Kitchell, J. R. Hodgson, P. A. Cochran, J. J. Elser, M. M. Elser, D. M. Lodge, D. Kretchmer, and X. He. 1987. Regulation of lake primary productivity by food web structure. Ecology **68**:1863–1876.

Conquest, L. L. 2000. Analysis and interpretation of ecological field data using BACI designs: discussion. Journal of Agricultural, Biological, and Environmental Statistics **5**: 293–296.

Crome, F. H. J., M. R. Thomas, and L. A. Moore. 1996. A novel Bayesian approach to assessing impacts of rain forest logging. Ecological Applications **6**:1104–1123.

Davis, E. A. 1989. Prescribed fire in Arizona chaparral: effects on stream water quality. Forest Ecology and Management **26**:189–206.

del Rosario, R. B., and V. H. Resh. 2000. Invertebrates in intermittent and perennial streams: Is the hyporheic zone a refuge from drying? Journal of the North American Benthological Society **19**:680–696.

Faith, D. P., C. L. Humphrey, and P. L. Dostine. 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. Second edition. John Wiley and Sons, New York, New York, USA.

Gerard, A. J., M. Jagers op Akkerhuis, and H. van der Voet. 1992. A dose–effect relationship for the effect of deltamethrin on a linyphiid spider population in winter wheat. Archives of Environmental Contamination and Toxicology **22**:114–121.

Giddings, J. M., R. C. Biever, R. L. Helm, G. L. Howick, and F. J. deNoyelles, Jr. 1994. The fate and effects of Guthion (azinphos methyl) in mesocosms. Pages 469–495 *in* R. L. Graney, J. H. Kennedy, and J. H. Rodgers, editors. Aquatic mesocosm studies in ecological risk assessment. Lewis, Boca Raton, Florida, USA.

Hogg, I. D., and D. D. Williams. 1996. Response of stream invertebrates to a global-warming thermal regime: an ecosystem-level manipulation. Ecology **77**:395–407.

Hurd, M. K., S. A. Perry, and W. B. Perry. 1996. Nontarget effects of a test application of diflubenzeron to the forest canopy on stream macroinvertebrates. Environmental Toxicology and Chemistry **15**:1344–1351.

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

Johnston, N. T., M. D. Stamford, K. I. Ashley, and K. Tsumura. 1999. Responses of rainbow trout (*Oncorhynchus mykiss*) and their prey to inorganic fertilization of an oligotrophic montane lake. Canadian Journal of Fisheries and Aquatic Sciences **56**:1011–1025.

Keough, M. J., and G. P. Quinn. 2000. Legislative vs. practical protection of an intertidal shoreline in southeastern Australia. Ecological Applications **10**:871–881.

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.

Lydersen, E., E. Fjeld, and E. T. Gjessing. 1996. The humic lake acidification experiment (HUMEX): main physicochemical results after five years of artificial acidification. Environment International **22**:591–604.

Moring, J. R. 1975. The Alsea watershed study: effects of logging on the aquatic resources of three headwater streams of the Alsea River, Oregon. Part II. Changes in environmental conditions. Fishery Research Report 9. Oregon Department of Fish and Wildlife, Corvallis, Oregon, USA.

Moring, J. R., and R. L. Lantz. 1975. The Alsea watershed study: effects of logging on the aquatic resources of three headwater streams of the Alsea River, Oregon. Part 1. Biological studies. Fishery Research Report 9. Oregon Department of Fish and Wildlife, Corvallis, Oregon, USA.

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

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

Roberts, E. A. 1993. Seasonal cycles, environmental change and BACI designs. Environmetrics **4**:209–232.

Schindler, D. W., and E. J. Fee. 1974. Experimental Lakes Area: whole-lake experiments in eutrophication. Journal of the Fisheries Research Board of Canada **31**:937–953.

Schmitt, R. J., and C. W. Osenberg, editors. 1996. Detecting ecological impacts: concepts and applications in coastal habitats. Academic Press, San Diego, California, USA.

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

Smith, E. P., D. R. Orvos, and J. Cairns, 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.

Solazzi, M. F., T. E. Nickelson, S. L. Johnson, and J. D. Rodgers. 2000. Effects of increasing winter rearing habitat on abundance of salmonids in two coastal Oregon streams. Canadian Journal of Fisheries and Aquatic Sciences **57**: 906–914.

Stewart-Oaten, A. 1996. Problems in the analysis of environmental monitoring data. Pages 109–131 *in* R. J. Schmitt and C. W. Osenberg, editors. Detecting ecological impacts: concepts and applications in coastal habitats. Academic Press, San Diego, California, USA.

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

Stout, R. J., and M. P. Rondinelli. 1995. Stream-dwelling insects and extremely low frequency electromagnetic fields: a ten-year study. Hydrobiologia **302**:197–213.

Uddameri, V., S. A. Norton, J. S. Kahl, and J. P. Scofield. 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 S. S. Bell. 1994. Resident fishes and macrobenthos in mangrove-rimmed habitats: evaluation of habitat restoration by hydrologic modification. Estuaries **17**:585–596.

Wallace, J. B., S. L. Eggert, J. L. Meyer, and J. R. Webster.

1997. Multiple trophic levels of a forest stream linked to terrestrial litter inputs. Science **277**:102–104.
Wardell-Johnson, G., and M. Williams. 2000. Edges and gaps in mature karri forest, southwestern Australia: logging effects on bird species abundance and diversity. Forest Ecology and Management **131**:1–21.

APPENDIX A

Details of Two-Stage Intervention Analysis

The following approach is motivated and explained further in Murtaugh (2000).

Let $\bar{y}_{1B}$, $\bar{y}_{1A}$, $\bar{y}_{2B}$ and $\bar{y}_{2A}$ be the means of the observations in the control (1) and treated (2) units, before (B) and after (A) the intervention time. Let n_1 and n_2 be the numbers of pre- and post-intervention observation times, respectively, and let $n = n_1 + n_2$.

1) Estimate σ^2_γ, the temporal variability of the ''half-series means'' within units.

a) First, estimate σ^2, the variance of the observations around the four half-series means:

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

b) Next, calculate

$$\hat{\sigma}^2_\gamma = \frac{1}{3}[(\bar{y}_{1B} - \bar{\bar{y}})^2 + (\bar{y}_{1A} - \bar{\bar{y}})^2 + (\bar{y}_{2B} - \bar{\bar{y}})^2] - \frac{\hat{\sigma}^2}{n/2} \quad \text{(A.2)}$$

where $\bar{\bar{y}}$ is the mean of the three half-series means. If $\hat{\sigma}^2_\gamma$ is negative, set it to zero.

2) Estimate $\boldsymbol{\Sigma}$, the variance–covariance matrix of the between-unit *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 $I_{[t^*,\infty)}(t_i)$; $\mathbf{d} = (d_1, \ldots, d_n)'$, with $d_i = y_2(t_i) - y_1(t_i)$; and $\beta = (\beta_0\ \beta_1)'$, then we can write the following:

$$\mathrm{E}(\mathbf{d} \mid \mu_{1B}, \mu_{1A}, \mu_{2B}, \mu_{2A}) = \mathbf{X}\beta$$
$$\mathrm{Var}(\mathbf{d} \mid \mu_{1B}, \mu_{1A}, \mu_{2B}, \mu_{2A}) = \boldsymbol{\Sigma}. \quad \text{(A.3)}$$

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, $\mathbf{e} = \mathbf{d} - \mathbf{X}\hat{\beta}$, estimate ρ_d, the first-order autoregression coefficient, in the usual way (e.g., see Fuller 1996). If a test of $\rho_d = 0$ is nonsignificant, set $\hat{\rho}_d$ equal to zero. Calculate $\hat{\sigma}^2_d = \widehat{\mathrm{Var}}(e_i) \times (1 - \hat{\rho}^2_d)$.

c) Calculate

$$\hat{\boldsymbol{\Sigma}} = \frac{\hat{\sigma}^2_d}{1 - \hat{\rho}^2_d} \begin{pmatrix} 1 & \hat{\rho}_d & \hat{\rho}^2_d & \cdots & \hat{\rho}^{n-1}_d \\ \hat{\rho}_d & 1 & \hat{\rho}_d & \cdots & \hat{\rho}^{n-2}_d \\ \vdots & & & \cdots & \vdots \\ \hat{\rho}^{n-1}_d & \hat{\rho}^{n-2}_d & \cdots & \hat{\rho}_d & 1 \end{pmatrix}. \quad \text{(A.4)}$$

d) Calculate

$$\hat{\beta} = (\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{d} \quad \text{(A.5)}$$

and repeat steps (a) through (c) if $\hat{\beta}$ is not sufficiently close to its value from the preceding iteration.

3) If $\hat{\delta} \equiv \bar{d}_A - \bar{d}_B$ is the estimate used in paired intervention analysis, calculate

$$\widehat{\mathrm{Var}}(\hat{\delta}) = 4\hat{\sigma}^2_\gamma + \mathbf{a}'\hat{\boldsymbol{\Sigma}}\mathbf{a} \quad \text{(A.6)}$$

where $\hat{\sigma}^2_\gamma$ is from step 1, $\hat{\boldsymbol{\Sigma}}$ is from step 2, and $\mathbf{a} = (-1/n_1 \cdots -1/n_1\ 1/n_2 \cdots 1/n_2)'$. We can then test the hypothesis that $\delta = 0$ by comparing the test statistic $\hat{\delta}/\sqrt{\widehat{\mathrm{Var}}(\hat{\delta})}$ to a t_{df} distribution, where

$$\mathrm{df} = 2p + (n - 2)(1 - p) \quad \text{(A.7)}$$

with $p = 4\hat{\sigma}^2_\gamma/(4\hat{\sigma}^2_\gamma + \mathbf{a}'\hat{\boldsymbol{\Sigma}}\mathbf{a})$.

APPENDIX B

Descriptions of the Data Sets

Following are descriptions of the data sets used in the analyses, including explanations of the abbreviations used in Table 1.

A) Average concentrations of chlorophyll *a* (in micrograms per liter) in the top 4 m of seven unmanipulated study lakes in Wisconsin, USA (AL = Allequash, BM = Big Muskellunge, CB = Crystal Bog, CR = Crystal Lake, SP = Sparkling, TB = Trout Bog, TR = Trout Lake). Following Carpenter et al. (1989), I did all pairwise comparisons of these ''reference'' lakes, using 1984–1986 data. Source: Chlorophyll *a* in North Temperate Lakes Primary Study Lakes—Trout Lake Area, North Temperate Lakes Long Term Ecological Research program, National Science Foundation, Timothy K. Kratz, Center for Limnology, University of Wisconsin–Madison (available online).[2]

B) Non-equilibrated surface pH in the seven Wisconsin lakes listed in (A), 1984–1986. Source: Nutrient Data of North Temperate Lakes Primary Study Lakes, North Temperate Lakes Long Term Ecological Research program, National Science Foundation, Emily H. Stanley, Center for Limnology, University of Wisconsin–Madison (available online).[3]

C) Data from three lakes in Wisconsin, USA: Paul Lake (PA; reference), and Peter (PT) and Tuesday (TU) Lakes, which received reciprocal transplants of fish in the spring of 1985 (Carpenter et al. 1987). The responses are dry biomass of zooplankton (zoo; in grams per square meter); percentage of zooplankton biomass consisting of animals >1 mm in length (pct); phytoplankton biovolume density (phyto; in units of 10^6 μ^3 per mL; and primary production (pp) in units of milligrams of C per cubic meter per day). Sampling was done in the summers of 1984 and 1985—weekly for zooplankton biomass and phytoplankton biovolume, daily for primary production.

[2] URL: ⟨http://www.limnology.wisc.edu⟩

[3] URL: ⟨http://www.limnology.wisc.edu⟩

D) Numbers of spiders caught per day per plot, from a study of the effect of deltamethrin on a linyphiid spider population (Gerard et al. 1992). A randomized block design was used: in each of three blocks, one control plot (labeled 13, 24, or 32) was compared to four or five other plots that received different doses of deltamethrin on day 10 of the 17-d experiment. Also included are all pairwise comparisons of the three control plots.

E) Densities (no. of individuals per liter) of copepods (cop), cladocerans (cld) and rotifers (rot) in mesocosms in a study of effects of the insecticide Guthion, from Tables 6–8 of Giddings et al. (1994). For each zooplankton type, there were two control mesocosms (labeled 0) and five pairs of treated mesocosms (labeled 1 through 5) that received various doses of Guthion halfway through the experiment.

F) Numbers (per square meter) of coho salmon (ch) and cutthroat trout (ct) in streams draining three basins in the Alsea Watershed (Oregon, USA): a control basin (C), a patch-cut basin (T1), and a clearcut basin (T2). These are annual estimates from before (1959 to 1965) and after (1966 to 1974) the treatments. The original data were from Moring and Lantz (1975), with recent corrections supplied by J. Dambacher (*personal communication*).

G) From the above study, mean maximum daily temperature in July (°C), for streams from the control basin and the clearcut basin, 1959 to 1973 (Moring 1975).

H) Concentrations of Ca^{++} (in milligrams per liter) in streamwater from two watersheds in the Hubbard Brook Experimental Forest (West Thornton, New Hampshire, USA) one of which was clearcut at the end of 1965, from Fig. 9 of Likens et al. (1970). Analyses are done for the pre-intervention data only, and for the entire set of data (June 1965 to May 1968).

I) Fish abundance at sites upstream and downstream of a power plant for seven years before and seven years after the plant went into operation (Smith et al. 1993). These data are from the east side of the river (Fig. 1a and b in Smith et al. 1993).

J) Mean densities of the stonefly *Isoperla* in control streams and in streams under diflubenzuron-treated canopies (Hurd et al. 1996). Monthly sampling started in October 1989; the diflubenzuron treatment was applied in May 1992.

K) Sodium concentrations (in parts per million) in streams flowing through two contiguous watersheds, one of which was subjected to a complete burn (Davis 1989). These data span eight months of pre-treatment monitoring and six months of post-treatment monitoring.

L) Numbers of white-tail rats caught in logged and unlogged sites in Queensland, Australia (Crome et al. 1996). There are 11 pre-logging observations and 5 post-logging observations, between September 1982 and December 1986.

M) Mean tree-ring chronologies from spruce stumps and trunks in two forest stands in northern Quebec: one that survived a fire in 1568, and one nearby control stand (Arseneault and Payette 1997). The rings in the data set correspond to the years from 1560 to 1584.

APPENDIX C

Technical Details

Below are some details about the analyses of the data sets.

1) When observation times were not identical between units, the times for the less-observed unit were taken as the baseline, and response values in the other unit at those times were obtained by linear interpolation.

2) In the application of two-stage intervention analysis to pairs of reference units, the results depend on which unit is labeled the control and which is considered to have received a hypothetical intervention at the median observation time. I always made the choice that minimized the variance component due to time periods, providing a conservative comparison to RIA (i.e., the expected rejection rate is higher than it would be if the designation of manipulated vs. control units were random).

3) If the number of pre-intervention observation times (n_1) was different from the number of post-intervention times (n_2), I replaced n_1 in Eq. 3.2 of Murtaugh (2000) by the average of n_1 and n_2 (see Appendix A).

4) In the analysis of the data from data set C in Appendix B, where two years of observations were separated by a long inter-annual gap, r was calculated as the weighted average of estimates done separately for each year, and the covariance matrix used in two-stage intervention analysis (see Murtaugh 2000 and Appendix A) was modified to reflect independence of the two clusters of observations.