Foundations of Restoration Ecology
D. A. Falk, M. A.. Palmer, and J. B.
Zedler, editors. Island Press, 2006

Chapter 13

Statistical Issues and Study Design in Ecological Restorations: Lessons Learned from Marine Reserves

Craig W. Osenberg, Benjamin M. Bolker, Jada-Simone S. White, Colette M. St. Mary, and Jeffrey S. Shima

Scientists and managers often seek to restore degraded systems to more desirable states. A system might be restored by eliminating a putatively deleterious factor(s) and allowing the system to recover naturally (e.g., by removing a sewage outfall or abolishing pesticide application) or by aggressively managing the system to reduce the time required for natural recovery. Regardless of the approach taken, we need to know if the restoration has fulfilled expectations. Thus, two fundamental questions underlie the scientific assessment of any restoration project: (1) What is the goal (e.g., to what state should the system be restored)? and (2) Did the restoration project achieve this goal (or, more generally, what were the effects of the restoration project)? Both aspects are central to the inferences we draw about restoration efforts and intimately linked to the statistical tools that we use to make these inferences.

Goals of restoration projects fall into two broad categories. The first, which we call *endpoint based*, aims to restore the system to a predefined state. We may define endpoints theoretically (e.g., that the density of an endangered bird species be restored to ≥50 breeding pairs based on a population viability analysis) or empirically, by comparison to a more "pristine" reference site (e.g., that species richness be ≥90% of that found at the reference site). Outcomes can be assessed by sampling the restored system and comparing it with the stated endpoint. To help formulate inferences, we might use a standard statistical null-hypothesis framework in which a single sample is compared with a theoretical expectation, or two samples are directly compared. Statistical power could also be considered in the assessment of restoration effects (low power will reduce our ability to detect the effects of restoration: Mapstone 1995). Although useful in many contexts, these endpoint-based approaches fail to provide an estimate of the *effect* of the restoration activity. In fact, the restoration effort may not have had any effects and yet the site may reach the desired state (e.g., due to natural variation independent of the restoration). This may be satisfactory in many contexts, but such a result would fail to inform future restoration projects.

Thus, we also define *effect-size-based* goals, in which we quantify the effects (and the associated uncertainty) of the restoration activity (e.g., determine the increase in the abundance of a threatened species caused by the restoration project), possibly by comparison with similarly degraded sites (as opposed to pristine sites), so that the response to the restoration can be quantified.

A combination of both approaches is likely ideal—we would like to know how much of an effect we have produced (effect-size-based outcomes) and if that change is "sufficient" (endpoint-based outcomes). In this chapter, however, we focus on effect-size-based goals and the study designs that facilitate this assessment, because endpoint-based approaches can be tackled with well-known statistical tools (e.g., ANOVA). In contrast, the apparently "simple" task of quantifying an effect size requires approaches that often are distinct from the standard quantitative tools we learn in basic statistics or experimental design courses, especially when dealing with large-scale, unreplicated assessments. These solutions are not, therefore, generally appreciated or applied. Fortunately, the complex challenges (and solutions) that are posed are very similar to those shared by assessments of unreplicated human interventions, such as the study of the effects of sewage outfalls, foresting practices, or nuclear power plants. As a result, we borrow heavily from the literature on impact assessment (e.g., Stewart-Oaten et al. 1986; Schmitt and Osenberg 1996). Effect sizes may be either univariate or multivariate, but for simplicity of discussion and presentation, we lay out the framework for univariate measures. Multivariate analogues exist for our univariate examples. For more general discussion of statistical issues in restoration studies, we refer the reader to the useful reviews by Michener (1997) and Schreuder et al. (2004).

To provide context to our discussion of assessment designs, we draw examples from the restoration of marine systems through the establishment of marine protected areas (MPAs) (Allison et al. 1998; Lubchenco et al. 2003; Norse et al. 2003). MPAs share many features with other restoration activities: (1) they are expected to have local effects within the boundaries of the restoration activity; (2) they also may have effects that extend beyond the MPA boundaries and therefore help restore degraded sites that are not actively managed, but may nonetheless benefit from distant restoration activities; and (3) there remains a considerable need for improved tools to document and estimate the local and regional effects of a given restoration effort.

Below, we discuss the central concepts drawn from experimental design and contrast these approaches with those needed in large-scale restorations (and impact assessments in general). We then discuss the major types of assessment designs, including their advantages and limitations, and highlight these issues with a critique of MPA studies. Last, we propose future directions, including more appropriate designs that will address current shortcomings and enhance the practice of restoration ecology.

Central Concepts

The basic question posed in any effect-size-based assessment study is simple to state and hard to solve: how does the state of the system after restoration compare with the state of the system that would have existed had the restoration activity not taken place (Stewart-Oaten et al. 1986; Stewart-Oaten 1996a)? Of course, the latter cannot be observed directly (because the restoration activity *did* take place) and must therefore be estimated. That is the crux of the problem: how do we estimate this unknown state and therefore (i.e., by comparison with the observed state) infer the effect of the restoration activity? The classic approach is experimental and employs null-hypothesis tests. Indeed, experiments are the primary tool of many restoration ecologists, so we begin with a discussion of issues germane to field experiments.

P-values Versus Estimation

Most ecologists use frequentist statistics, epitomized by P-values and tests of null hypotheses. If the observed data are not very unlikely under the null hypothesis (typically, $P > 0.05$), then we tentatively accept the null hypothesis, which is often erroneously interpreted as indicating "no effect" (Yoccoz 1991). Alternatively, if the data are sufficiently unlikely under the null (typically, $P < 0.05$), then we conclude that there was "an effect." The *P*-value itself (or the test statistic), however, gives little indication of the likely effect size or the associated uncertainty; we know only whether the confidence interval on this effect includes or excludes zero.

Consider two studies of the effects of two restoration approaches on the abundance of an endangered species. Approach A leads to an estimated increase in population density of 0.1% per year (± 0.11%), whereas approach B yields an effect of 100% (±101%). Although neither result is "significant," in approach A, we have high confidence that the effect is "small" because of the high precision in the estimate. In B, we do not even know the direction of the effect—the restoration might have very detrimental effects or extremely positive effects. A conclusion of "no effect" cannot be made with any confidence.

Instead of *P*-values we need to estimate the magnitude of effects and their uncertainty (Yoccoz 1991; Stewart-Oaten 1996a; Johnson 1999; Osenberg et al. 1999, 2002; Anderson et al. 2000). This is especially true in assessment studies in which policy makers, the public, and the scientific community should care less about whether there is a demonstrable (but possibly tiny) effect and more about the magnitude (and uncertainty) of the response (Stewart-Oaten 1996a). In this chapter, we emphasize estimation and refer the reader to these other sources for greater detail about the *P*-value culture.

An Experimental Approach: Why Do We Need an Alternative?

A restoration project might be conducted using a standard experimental approach, with multiple treatments (including appropriate controls), replication (multiple independent units that receive a given treatment), and random assignment of units to treatments (Underwood 1997; Scheiner and Gurevitch 2001). Imagine a site in which sea grass was previously present but was severely damaged by an anthropogenic activity (e.g., dredging or an oil spill). An investigator could choose multiple plots within this site and randomly assign them to two or more treatments (e.g., a suitable "control" plus different "restoration" treatments). After some appropriate amount of time the plots could be sampled and compared using standard statistical procedures. In principle, such an approach can be useful, especially to compare different possible restoration techniques. However the extension to a large-scale restoration project requires that (1) the spatial scale of the plots is appropriate to the overall goals of the large-scale restoration project; (2) the plots are independent of one another (e.g., restoration treatments do not affect adjacent control plots); and (3) the analysis focuses on effect sizes and their uncertainty.

To explore this issue, we reviewed all papers from the 2003 volume of *Restoration Ecology*. Of the 68 papers that reported results from studies that could be used to infer effects of a restoration activity (or activities), 41 were experimental, with replication, random assignment, and a control. Of these, the modal scale of manipulation was 10 m^2 (range: ~0.025–4 $\times 10^6$ m^2) with all but 6 occurring on scales <100 m^2.

However, propagules disperse, herbivores colonize, and predators typically forage over scales larger than 100 m^2. Indeed, scaling up small experiments to their larger-scale implications is a continuing challenge for ecologists (Englund and Cooper 2003; Melbourne and Chesson 2005; Schmitz 2005). As a result, small-scale experiments (e.g., conducted on the scale of 10 m^2), although useful for revealing mechanisms and evaluating likely restoration strategies, may be poor predictors of actual effects of a large-scale restoration project or the success of a restoration conducted at a larger scale (e.g., involving hectares or km^2). Of course, we could conduct experiments at larger spatial scales (e.g., using many different sea grass beds as replicates and having half randomly assigned to controls), but this is rarely feasible. For example, in our survey, only 2 of 68 studies were replicated and conducted at a scale >10,000 m^2 (also see Michener 1997 and Schindler 1998 for examples of associated constraints).

Even if a replicated large-scale experiment were possible, it would only reveal the *average* effect of a restoration activity on the population of potential sites and not the effect at any particular site. This would be useful to compare among possible restoration approaches; however, we are often most interested in understanding the effect of restoration at a *particular site* (e.g., for mitigation or regulation). Furthermore, if some sites were positively affected by restoration while others were negatively affected, one could conclude "no effect" overall. Instead, we would prefer to know which sites were positively and negatively affected (and, ideally, why). In an experiment, a "positive," site-specific effect cannot be inferred by the deviation of one site from the pool of replicates, because the restoration effect is confounded with other aspects of that site (e.g., initial conditions). This limitation is a generic feature of replicated experiments and standard statistical approaches.

Thus, we propose an approach that departs from our standard experimental training and that (1) can be applied to spatially unreplicated interventions; (2) is site-specific; and (3) yields defensible estimates of the effect of the restoration activity (rather than *P*-values or "yes/no" answers). That approach is the BACIPS (Before-After–Control-Impact Series) assessment design, which is currently used in impact assessments (Stewart-Oaten et al. 1986; Schmitt and Osenberg 1996). Interestingly, none of the studies we reviewed in *Restoration Ecology* used a BACIPS study or presented a cogent description of these assessment issues, suggesting that BACIPS could be a valuable addition to the restoration ecology tool kit.

Local Versus Regional Effects

Local effects, which are the focus of most restoration studies (and all of those we reviewed), arise within the boundaries of the specific restoration activity. However, effects will not be limited to the boundaries of the restoration project. Indeed, we expect that there will be regional effects that arise outside the restored site, for example, due to movement of plant propagules, animals, detritus, or nutrients. Selection of control sites (which need to be independent of the restoration effects) must therefore consider the life history and dispersal capabilities of the interacting species and the transport of materials. Local and regional effects also must be studied with different study designs, because one emphasizes effects that occur within the boundary of the project and the other focuses on effects outside of the boundary. In some cases (as we illustrate below), the regional effects are of equal (if not greater) importance than the local effects, yet they remain understudied because of problems inherent to their assessment.

Assessment Designs and Their Application to Restoration Ecology

For our discussion of effect-size-based approaches, we assume that the restoration effort is unreplicated, that reference and restoration site(s) are not necessarily assigned at random, and that all sites are initially degraded, although these conditions are not required. We refer to the reference site(s) as a "Control" and the site to be restored as the "Impact" site, as in the impact assessment literature (Schmitt and Osenberg 1996). Our goal is to estimate the change in some variable (say population density of a focal species) at the Impact site resulting from the restoration activity. Below we summarize common assessment designs to highlight the differences in their approach and the problems that may arise in drawing conclusions from the resulting data.

Control-Impact (CI) Designs

In this common design, multiple samples are typically taken from plots within an Impact site and at least one Control site. These two sets of samples are compared statistically to determine if the two sites differ. If they do, then we conclude that there was an effect of the restoration activity. Of course, because no two sites are identical (although Control and Impact sites may be *similar*), there will likely be statistically significant differences between the two sites. This will be true even before the restoration project begins. Thus, the Control-Impact design confounds the effect of the restoration project with other processes that produce spatial variation in parameters (e.g., Figure 13.1a).

Before-After (BA) Designs

The Before-After design avoids problems with spatial variation by sampling only the Impact site and comparing its state Before versus After restoration (e.g., see Figure 13.1b). We discuss two variants of this basic design.

BA-Single Time

The Impact site is sampled once Before and once After the restoration activity (with many plots within each site providing "replication"). However, all systems change through time, so any two sets of samples from the same site (but different times) will be different (assuming sufficient sampling). Thus, the BA-single-time design confounds the restoration effect with other processes that produce temporal variation.

BA-Time Series

Multiple sampling times within a period provides a form of replication that allows the investigator to incorporate, and potentially deal with, temporal variation. By using time-series methods that account for serial correlation, BA designs can be used to infer effects. Indeed, one of the most famous of all intervention studies was Box and Tiao's (1975) BA-time-series study of ozone in downtown Los Angeles and its response to two separate interventions: (1) the simultaneous reformulation of gasoline designed to reduce reactive hydrocarbons and

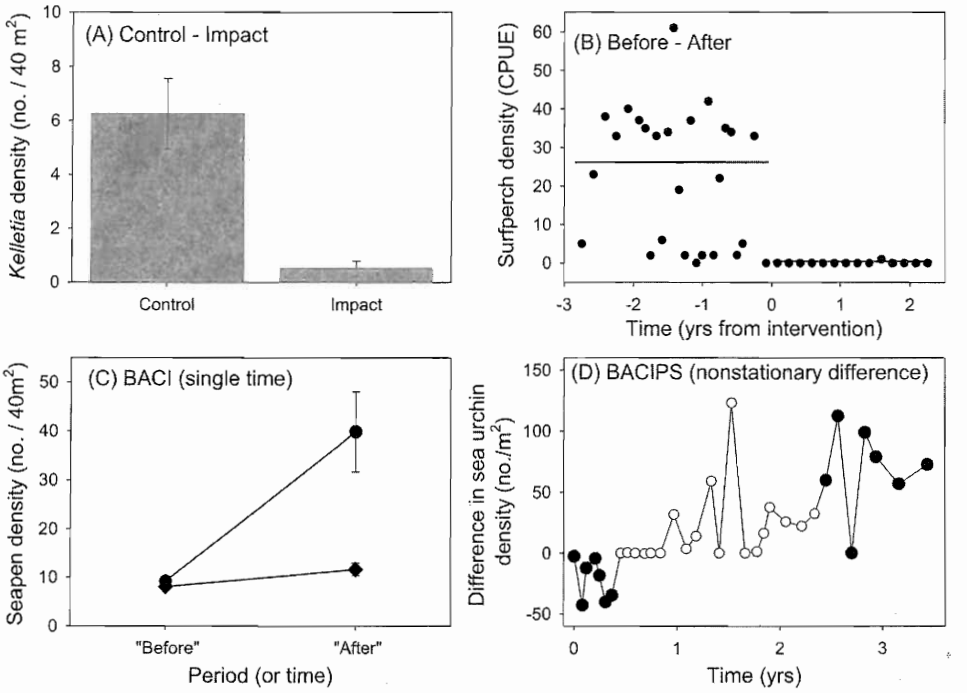


FIGURE 13.1 Empirical examples of assessment designs in which erroneous inferences would be drawn due to confounding of natural variability with effects of an intervention. (A) A Control-Impact design investigating effects of oil and gas production on a benthic mollusc (*Kelletia kelletii*) (see Osenberg et al. 1992, 1994; Osenberg and Schmitt 1996). The data were taken in a Before period and therefore represent preexisting spatial variation in density and not an effect of the oil production activity. (B) A Before-After design investigating effects of the cooling tower effluent of a nuclear power plant on the abundance (catch per unit effort = CPUE) of pink surfperch, *Zalembius rosaceus* (see Murdoch et al. 1989). Time = 0 indicates the date on which power was first generated following expansion of the power plant. However, these data came from a Control site and indicate natural temporal variability, not effects of the power plant. (C) A BACI design (without a time series) studying effects of oil production on the density of seapens (*Acanthoptilum* sp.). Production did not begin when expected, so this relative change in the Control and Impact sites represents a natural space-by-time interaction and not an effect of oil production. (D) A BACIPS study showing a time series of differences in sea urchin (*Lytechinus anamesis*) between an Impact and Control site, illustrating the possible confounding of an effect with long-term natural changes in density (e.g., if the two time periods indicated by filled circles happened to define the Before and After periods). The data come from a Before period and indicate a long-term trend in the differences independent of the intervention.

rerouting of traffic following the opening of the 405 freeway (these two were considered together due to their temporal confluence); and (2) redesign of the engines of new cars. The first intervention was predicted to produce a step-change reduction in ozone, and the second was expected to gradually reduce ozone as new cars replaced older versions. Box and Tiao framed a stochastic model of the interventions, defined an analytic approach based on that

model, ran diagnostics to determine model inadequacies, and, barring the latter, derived inferences about the response of ozone to the interventions. They concluded that both interventions had demonstrable effects (Figure 13.2).

Box and Tiao's success was, in part, due to (1) the long and dense time series (monthly averages of ozone from an 18-year period); (2) the well-behaved temporal dynamics of ozone; and (3) the simple expectations about plausible effects of the interventions on ozone. These advantages are unlikely to exist for most ecological studies (perhaps with the possible exception of some epidemiological studies: e.g., Earn et al. 2000). Figures 13.1b and 13.1d offer examples of ecologically "long" time series (five years) that were too short to capture relevant background temporal dynamics. We return to this issue in the next section.

Before-After–Control-Impact (BACI) Designs

BACI designs attempt to deal with both spatial and temporal variation by sampling at one or more Control site(s) and the Impact site both Before and After the intervention. A variety of permutations on the basic theme have been proposed.

BACI (Single Time)

Green (1979) proposed a BACI design in which a Control and Impact site were sampled once Before and once After an intervention. A site-by-time interaction indicates an effect of

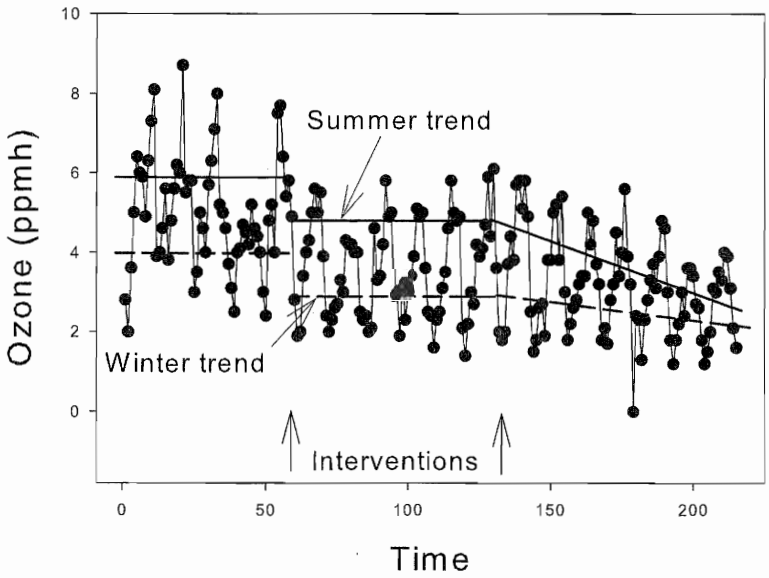


Figure 13.2 Summary of the results of Box and Tiao's (1975) Before-After study of ozone in the Los Angeles basin. Points give monthly ozone concentrations. Arrows indicate the timing of the two interventions: one hypothesized to result in a step change and one hypothesized to result in a gradual reduction in ozone. The solid line gives the estimated trend for summer conditions and the dashed line gives the estimated trend for winter conditions (other seasonal trends are excluded for clarity).

the intervention. However, no two sites show the same temporal dynamics. Thus, we expect site-by-time interactions when two sites are sampled intensively on two different dates (Figure 13.1c). This BACI design therefore confounds effects of the intervention with other factors that cause site-by-time interactions.

BACI-Paired Series (BACIPS)

In the basic BACIPS design, a Control (or set of Controls) and an Impact site are sampled simultaneously several times Before and After the perturbation (Stewart-Oaten et al. 1986). The parameter of interest is the *difference* in a chosen variable (e.g., density of a target species) between the Control and Impact sites estimated on each sampling date. Each difference from the Before period provides an estimate of the spatial variation between the two sites and thus is an estimate of the expected difference that should exist in the After period in the absence of an effect of the intervention. The difference between the average Before and After differences provides an estimate of the magnitude of the effect of the intervention. The simplest design assumes that there is no serial correlation (or temporal trend) in the differences between the Control and Impact sites (serial correlation will result if the sampling within a site is done at too short an interval). If there is serial correlation in the differences, then an autoregressive approach can be used to account for the correlation structure (see Stewart-Oaten et al. 1992; Stewart-Oaten and Bence 2001).

If sampling is done too close together for too short a time period, the serial correlation structure cannot be detected and may be confounded with the "effect" (Figure 13.1d); instead of indicating a true effect of the intervention, the change in the difference from Before to After may be the result of oversampling during a single, short-lived, local perturbation in each period, or sampling over a time interval in which the true difference was changing naturally and gradually through time. Indeed, the Before period is critical for developing diagnostic tests of the patterns of covariation between the Control and Impact sites (but see Murtaugh 2002, 2003). Especially important are the pattern of serial correlation and the additivity of site and time effects (Stewart-Oaten et al. 1986, 1992; Bence 1995; Stewart-Oaten 1996b; Bence et al. 1996). We return to this below.

Predictive BACIPS

The BACIPS design uses the Control site to predict the Impact's state (Bence et al. 1996): the Impact site's state in the After period (and assuming no effect of the intervention) can be predicted as the sum of the Control's state in the After period plus the mean difference between the Control and Impact site estimated during the Before period. Bence et al. (1996) have advocated a more flexible approach in which the relationship between the Control and Impact values is compared Before and After the intervention (Figure 13.3). This approach is intuitively appealing and has the advantage of allowing the effect size to vary (e.g., with the overall environmental conditions, as indexed by the Control value), but it has the disadvantage that the independent variable (the Control value) is measured with error and thus violates a standard assumption of Model I regression models. This problem has not been clearly resolved in the predictive-BACIPS approach.

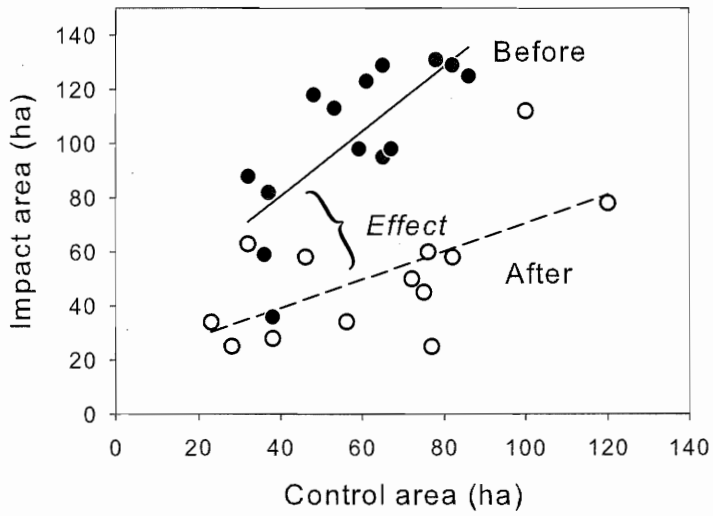


FIGURE 13.3 Illustration of the predictive-BACIPS design using the Bence et al. (1996) study of the effect of a nuclear power plant on the areal extent of kelp (*Macrocystis pyrifera*) in southern California. The difference between the relationships between the Impact and Control site from Before to After gives an estimate of the effect of the intervention (operation of the power plant). In this case, the effect ranges from a reduction in kelp cover of ~40–80 ha, with the largest effects expected when conditions are good (i.e., when there is more kelp at the Control site).

BEYOND BACI

Underwood (1991, 1992, 1994) promoted a different elaboration of BACI that uses an "asymmetrical design" in which there are multiple Control sites. The data are not paired in time (i.e., the samples at the Controls and Impact sites do not share a common time effect) and thus the differencing approach of BACIPS is not relevant. Stewart-Oaten and Bence (2001) have critiqued this approach in depth, so we concentrate on the BACIPS designs.

Why Have a Control? What Makes a Good One?

Recall that Box and Tiao (1975) successfully used a BA design to examine effects of interventions on ozone in downtown Los Angeles. Yet we dismissed BA designs above as confounding effects of the intervention with other sources of temporal variation. However, Box and Tiao had a very long-time series of data, from which they were able to construct (and evaluate) plausible models of ozone dynamics with and without the interventions. In essence, their model of ozone dynamics from the Before period could be extrapolated to the After period and contrasted with the observed behavior to infer effects of the intervention. In ecological assessments we usually lack long-time series and well-defined temporal dynamics. Thus, a predictive ecological model from the Before period is not likely to provide an accurate null expectation for the After Period (Figure 13.1b). This is where the Control site helps.

Imagine that the variable of interest at the Impact site varies considerably (and possibly erratically) through time. Developing a predictive model of these dynamics may be very difficult. However, if another site (the Control) exhibits similar temporal dynamics in the ab-

sence of the intervention, then the Control site can be used to develop a more accurate model of the Impact's dynamics. Indeed, this is the key feature of a good Control site: it is not necessarily a site that is most like the Impact site, but rather it is one that changes through time in a way comparable to the Impact site in the absence of an intervention (Figure 13.4) (Magnuson et al. 1990; Osenberg et al. 1994). If the Control and Impact sites track one another through time (show high coherence), then there will be low variation in the differences through time, and BACIPS and predictive-BACIPS will have high power and will give rise to more accurate estimates of the effect sizes (Figure 13.4).

To illustrate this more specifically, let the parameter of interest be the difference (hereafter referred to as "delta," Δ, or D for its estimate) in density or other suitable variable between the Control and Impact sites as estimated on each sampling date (e.g., $D_{P,i} = N_{I,P,i} - N_{C,P,i}$), where $N_{I,P,i}$ and $N_{C,P,i}$ are sampled densities (often log-transformed) at the Control and Impact sites on the i^{th} date of Period P (i.e., Before or After). Each difference Before provides an estimate of the spatial variation between the two sites (Δ_B), which is the expected difference that should exist in the After period in the absence of an effect of the intervention. The difference between the average Before and After differences ($\bar{D}_B - \bar{D}_A$) provides an estimate of the effect of the intervention. Confidence in this estimate is determined by the variance in differences pooled across periods (s^2), as well as the number of sampling dates (i.e., replicates) in each of the Before and After periods (n_B, n_A). In the absence of serial correlation in the time series of differences (see also Stewart-Oaten and Bence 2001):

$$\text{Effect Size: } E = \bar{D}_B - \bar{D}_A = \frac{\sum D_{B,i}}{n_B} - \frac{\sum D_{A,i}}{n_A} \tag{1}$$

$$\text{Variance: } s^2 = \frac{s_B^2}{n_B} + \frac{s_A^2}{n_A} \tag{2}$$

$$95\% \text{ Confidence Interval: } E \pm s \cdot t_{n_B+n_A-2,0.025} \tag{3}$$

where for period P,

$$s_P^2 = [\sum (D_{P,i} - \bar{D}_P)^2]/n_P - 1 \tag{4}$$

In a standard null-hypothesis testing context, low variability (s^2, Equation 2 or 4) will lead to a more powerful test of the intervention effect and more accurate estimates of the effect (i.e., smaller confidence limits, Equation 3): see Osenberg et al. (1994). By taking differences between the Control and Impact sites (E, Equation 1), BACIPS removes the effects of background sources of variation that are common to both sites (e.g., responses to climatic events). By emphasizing differences and using a time-series approach, the BACIPS design accounts for some sources of spatial and temporal variation ignored in the BA and CI and BACI designs (Stewart-Oaten et al. 1986; Stewart-Oaten 1996a, 1996b).

Notice that the two main sources of variation in a BACIPS design are quite different from those used in other designs. The estimate of the effect (E, Equation 1) is derived from the Period-by-Location term (in standard ANOVA terms), which indicates how much the response variable at the Impact site (*relative* to the Control site) changed from the Before to After periods (i.e., $\bar{D}_B - \bar{D}_A$). The error component (Equation 2 or 4) measures how much the difference between the response variable at the Control and Impact sites varies in the

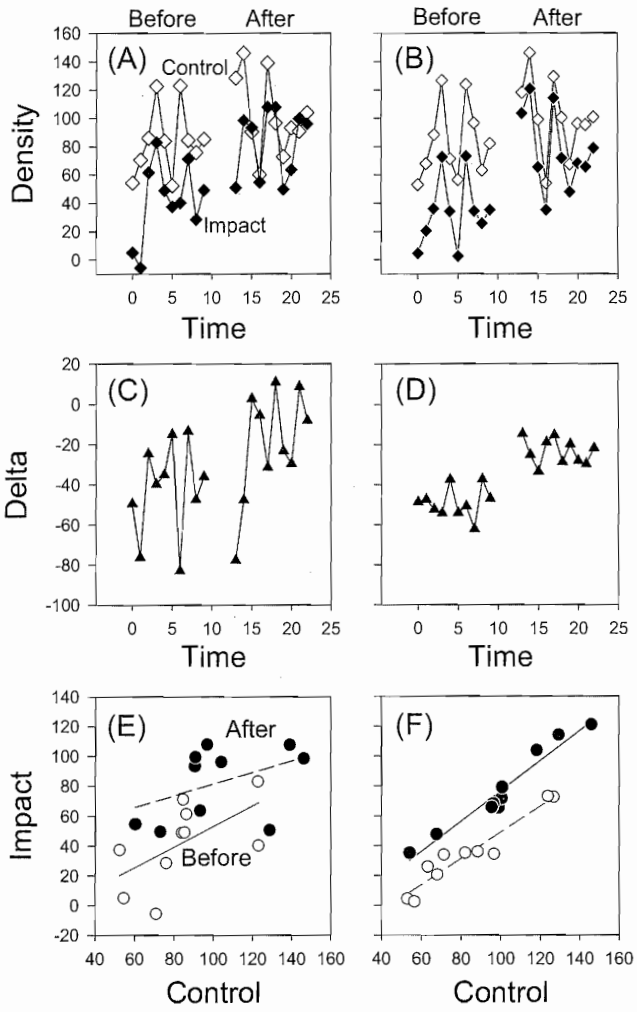


FIGURE 13.4 The effect of coherence between the Control and Impact site on the ability of the BACIPS and predictive-BACIPS designs to detect effects of an intervention. Coherence is the degree of strength of the correlation between the Control and Impact sites through time in the absence of a change in the status of the intervention (Magnuson et al. 1990; Osenberg et al. 1994). The panels on the left (A, C, and E) are for a system with relatively low coherence, whereas the panels on the right (B, D, and F) apply to a system with relatively high coherence. The data were simulated by constructing a time series from a random distribution and imposing a temporal trend in densities at both sites with a sine function. The variance in densities is the same under low and high coherence; the effect size is also identical (25). The only difference is the correlation between the two sites (high: $r = 0.99$; low: $r = 0.63$). The influence on inferences from BACIPS analyses is potentially dramatic. Effect sizes were estimated to be 25.6±6.6 (95% CI) for high coherence versus 21.9±24.9 for low coherence. Notice that under low coherence, the CI was very wide and included positive effects as well as deleterious effects; a *t*-test failed to reject the null hypothesis of no effect ($t_{18} = 1.91$, $P = 0.07$). For predictive-BACIPS, estimated effects and uncertainty were similarly affected (note difference in elevation and scatter in panels E and F, which give separate regression lines for the Before and After periods).

absence of a change in the intervention (i.e., the interaction between site and time *within* a period). Other designs use error terms based on the within-site sampling variation (Control-Impact and BACI designs) or temporal variation (Before-After design). Of course, inferences about cause and effect can be increased with ancillary studies of the mechanisms that might elicit change at the sites (e.g., Stewart-Oaten et al. 1992; Schroeter et al. 1993).

Case Studies: Marine Restoration Using Reserves

Marine reserves, or marine protected areas (MPAs), have been touted as a powerful tool to restore degraded marine systems, improve fisheries management, and conserve biodiversity. By limiting human activities, MPAs are thought to produce long-lasting increases in the density, size, diversity, and productivity of marine organisms within reserve boundaries due to decreased mortality and habitat destruction, as well as indirect ecosystem effects (e.g., Halpern 2003). Importantly, the effects of MPAs are hypothesized to extend beyond the boundaries of the MPA by "spillover": that is, via the density-dependent migration of juveniles or adults from inside to outside the MPA, or via increased production of planktonic larvae (spawned within the MPA), which are then exported outside of the MPA (e.g., Sanchez Lizaso et al. 2000). Thus, we expect both local and regional effects of MPAs. Indeed, it is the regional effect that is often used to motivate the designation of MPAs to the fishing community: reserves must enhance fisheries enough to compensate for the loss of fishing habitat (Palumbi 2000). Similar regional effects are expected in other conservation contexts, for example, by protecting the wintering grounds of a migratory bird or butterfly, effects should also arise in the breeding grounds.

Given the potential importance of MPAs as a restoration tool, many studies have examined effects of marine reserves on fishes and invertebrates and a recent meta-analysis by Halpern (2003) summarized those effects. We evaluated the designs of studies reviewed by Halpern and added additional studies by searching Web of Science for papers with the key words "marine protected area" or "marine reserve." We maintained the criteria for inclusion used by Halpern (2003): (1) data had to allow an inference about effects of the MPA; (2) measured variables had to include ecological responses (e.g., density or biomass); and (3) MPAs had to be "no-take" reserves. In total, we found 118 studies of MPA effects. Each study was conducted under various constraints (both political and scientific) and therefore the studies used different designs (e.g., CI versus BA versus BACI) and examined different scales of effects (i.e., local versus regional).

The majority of studies (70%) used a Control-Impact design to study local effects (Table 13.1). Fewer than 8% of the studies explored regional effects. No studies used a full BACIPS design with time series in both the Before and After periods (although some studies had time series in the After period and a single sample date in the Before period). Thus, not a single study used the most powerful assessment design (BACIPS) to study the regional effects that are of most interest to managers and often promoted by the scientific community.

Below, we look at several different approaches that have been taken, and highlight their limitations based on our previous generic discussions of assessment designs. We do this to emphasize the differences among the various study designs and their ability to look at appropriate scales of effects, and to inform future restoration studies, especially of marine reserves.

TABLE 13.1

Designs and scales of effects examined in studies of marine protected areas. Studies were obtained from Halpern's (2003) review and supplemented with further searches of the literature.

	Scale of study	
Design	Local	Regional
Control-Impact	82	3
Before-After	17	5
BACI	10	1
BACIPS	0	0

Control-Impact Studies

Because most of the studies that Halpern (2003) tabulated used a CI design to evaluate local effects (Table 13.1), we discuss Halpern's results in that context. Halpern achieved replication by combining the results from many unreplicated studies. Indeed, he observed strikingly consistent responses across the studies: for example, densities in reserves were 91% (95% CI: ~ 35–147%) greater than outside the reserves. He concluded that this consistent pattern was the result of a beneficial effect of MPAs on the densities of marine organisms. Increases also were observed for species richness (23%), organismal size (31%) and biomass (192%). Is there a reasonable alternative explanation to the appealing interpretation that the designation of MPAs has these beneficial effects?

In any single CI design the MPA effect is confounded with other factors whose effects vary spatially. Thus, we would expect the MPA to sometimes be placed in a "better" site and other times that the Control would go in the "better" site. On average, however, there should be no difference between the MPA and Control in the absence of an effect (assuming the MPA was assigned at random). Thus, the meta-analysis, which achieved replication by looking across studies, is comparable to a large-scale experiment (with MPA systems representing blocks, but lacking replication within blocks).

Of course, MPAs and Controls are not usually assigned randomly. Instead, MPAs are typically established following a laborious site selection process. Controls are rarely if ever discussed in the process; indeed, planning for a scientific assessment is rare. This is why CI designs are so common—the assessments are done after the fact, and the Control sites are often chosen by the investigator in a post hoc attempt to find sites that are otherwise "identical" to the MPA. Of course this is impossible. In most cases, MPAs (like most restoration sites) are put in specific sites—for example, the best remaining shallow coral reef habitat.

Thus, an alternative explanation for Halpern's result is that it reflects differences between the MPA and Control site that existed *prior* to the establishment of the MPA. Indeed, other meta-analyses indicate that the size of the reserve effect does not increase with time since the establishment of the MPA (Cote et al. 2001; Halpern and Warner 2002), suggesting a large role of initial conditions (but see Halpern and Warner 2002 for an alternative explanation). The problem, of course, is that the data cannot distinguish between the two alternatives. Hence, we are left either "believing" that MPAs are good and are in no better position than we were before the study was conducted or being skeptical and arguing that we need better data.

To further complicate inferences derived from such approaches, note that in the presence of regional effects, CI designs will underestimate true local effects because the Impact (MPA) site response will cause a concordant response at the Control site (i.e., they are not independent). Our hope is that by understanding the limits of even the best studies, such as Halpern's, we can ultimately obtain more defensible and less ambiguous interpretations. This requires Before data using designs conducted at appropriate scales.

Before-After Studies

Given the problems with site selection and possible non-independence between Control and Impact sites, why not simply avoid the use of Control sites all together and attempt to emulate the success of Box and Tiao (1975)? To explore this approach, we have extracted data from the studies of Russ and Alcala (1996, 2003) in the central Philippines. Although Russ and Alcala had Control sites, many of their inferences were based on patterns of change at two sites on Sumilon Island where fishing was "turned on and off" through time. As with most ecological studies, the data set is relatively sparse. We used these data (Equation 5a–b) to fit a model of fish dynamics that allowed us to estimate the effect of fishing:

$$N(t+1) = N(t) + (r + \varepsilon_r(t)) - (a + fF(t))N(t) \tag{5a}$$

$$N_{obs}(t) = N(t) + \varepsilon_{obs}(t) \tag{5b}$$

where $N(t)$ was the sampled density in year t; r was the average recruitment of new settlers into the local population and $\varepsilon_r(t)$ represents independent, normally distributed error with mean 0 and standard deviation σ_r; a is the background (nonfishing induced) mortality; f is the effect of fishing when it was allowed; $F(t)$ is the fraction of the year during which fishing was allowed (between 0 and 1); and $\varepsilon_{obs}(t)$ represents independent, normally distributed observation error with mean 0 and standard deviation σ_{obs}. We specified $N(0)$, the starting density of the population, as a parameter. When $\sigma_r = 0$, the other parameters can be estimated by simple least-squares fitting of the estimated population densities over time to the observed population densities, with σ_{obs} estimated from the residual sum of squares. To fit the model with process error ($\sigma_r > 0$), we ran many (up to 50,000) realizations of the population dynamics for a given set of parameters and used these realizations to compute the theoretical mean vector $\mathbf{m}$ of the observations as well as the variance-covariance matrix $\mathbf{V}$ among the observations. We then calculated the log-likelihood of the observed data given a multivariate normal distribution with mean $\mathbf{m}$ and variance-covariance matrix $\mathbf{V}$, and used a nonlinear fitting routine to maximize the log-likelihood. In practice, since estimates of standard errors were available for individual measurements, we determined σ_{obs} from the estimated sample standard error for a given census rather than trying to estimate this parameter from data. We used published data from 1983–2000 for the Sumilon Nonreserve (SNR) and from 1983–1994 for the Sumilon Reserve (SR). Limited fishing was permitted from 1995–2000 at SR, so we excluded this period of partial protection. Over these time periods SNR was opened for fishing except for 1987–1992 and therefore had a pattern of open-closed-open. SR was opened for fishing during two, approximately two-year, periods, and therefore had a closed-open-closed-open pattern of exploitation. These repeated "on-off" patterns potentially provide greater ability to detect interventions than the more standard single switch in Box

and Tiao's study. We used estimates of f (the fishing effect) to infer effects of the MPA on fish dynamics.

When we included process error, we obtained estimates of $f = -0.15 \pm 0.20\ yr^{-1}$ (95% CI) for SNR, which overlapped zero and failed to distinguish between beneficial and deleterious effects, and $f = 0.60 \pm 0.25\ yr^{-1}$ for SR, which provided good evidence for a demonstrable effect of the MPA (i.e., an increase of ~60% per year in the growth of the fish population released from fishing). Indeed, the confidence intervals of the fishing effect at the two reserves do not overlap, suggesting heterogeneity in the efficacy of the reserves. However, estimates of σ_r were large (e.g., 3.4 for SNR), suggesting a major role of environmental variability due to recruitment, r. Without process error, it was difficult to reconcile the data from SNR with a biologically plausible model, due in part to the large fluctuations in density that occurred when the site was continually fished (Figure 13.5). In contrast, the fit of the SR data was quite good, even in the absence of process error (Figure 13.5).

Our approach assumes that all fluctuations in the growth rate r are independent (and that there is no variation in the effect of fishing, f) and thus ignores serial correlation in the process error (in Box and Tiao's terms, we are fitting the autoregressive part of the model and ignoring the moving-average terms). Accounting for the serial correlation should be done but would only make our estimates even more uncertain. Despite estimating a significant effect of fishing for one of the sites, we are dangerously short on data. We are trying to fit a model

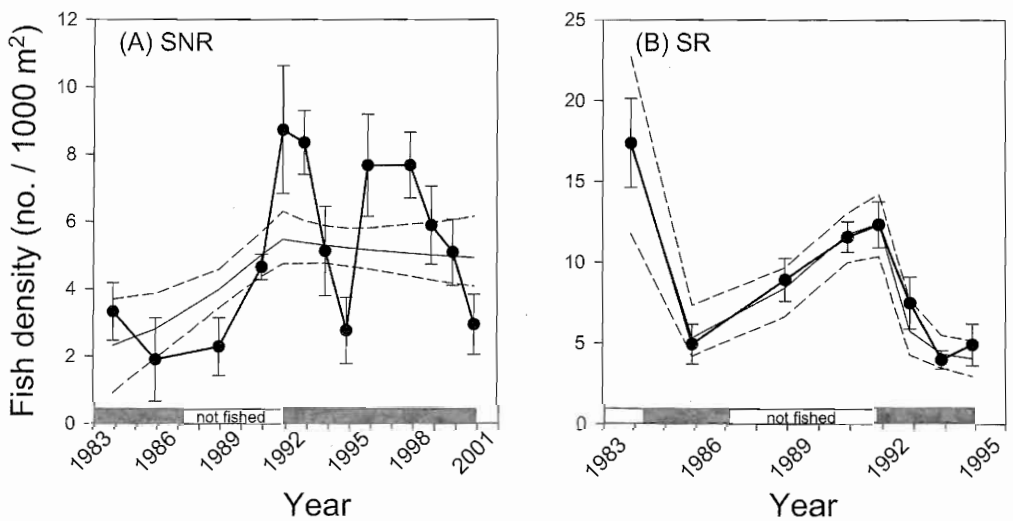


FIGURE 13.5 Data from Russ and Alcala's (2003) study of the response of large predatory fishes to the implementation of marine reserves in the Philippines. SNR and SR are two sites in which fishing was allowed or prohibited at different times between 1983 and 2000. Data points with error bars (±SE) give the observed fish densities. Solid lines without points give the predicted dynamics based on the mean of 1,000 simulations using parameter values drawn from the sampling distribution of the parameters (estimated from the curvature of the likelihood surface at the MLE): see equation 5a–b. The dashed lines bound 95% of all simulations. The simulation did not include process error (i.e., we set $\varepsilon_r = 0$) and thus the confidence bands reflect uncertainty in the parameter estimates and not temporal variation in the recruitment parameter.

with five parameters (two of them variances, which are notoriously hard to estimate) to 8 (or 13) data points in a time-series, which are not even independent of one another (and hence represent less than three, or eight, degrees of freedom). Indeed, most ecological data will not be sufficient in these regards. Furthermore, it will be difficult to develop detailed diagnostic checks and to compare alternate model formulations (e.g., functional forms, as well as error structure and serial correlation).

Unfortunately, Russ and Alcala's study is one of the best available with a fairly extensive time series by ecological standards. It helped that we were able to use a semimechanistic model, based on at least a caricature of a population growth model, that we had relatively detailed data on the interventions (=fishing intensity), and that the intervention fluctuated more than once (providing a stronger signal to pull out from the noise). Despite doing better than we initially expected (being able to pull out a signal at all), the estimates of fishing effects were uncertain. Can we do better?

Before-After–Control-Impact Studies and Spatial Scale

Although there are not any well-designed BACIPS regional studies, there are several that have elements of a BACIPS design. Here we discuss an important study by Roberts et al. (2001) on the St. Lucia reserve network in the Caribbean. Roberts et al. focused on regional effects on fisheries. Although they did not present a formal model, they did present data and results from a statistical analysis relevant to assessment designs. They used data collected once prior to the establishment of the MPA and four times during the subsequent five years. Data on fish biomass were taken inside and outside the reserve (Figure 13.6). They analyzed the data using a statistical model that included time, location (MPA versus outside of MPA),

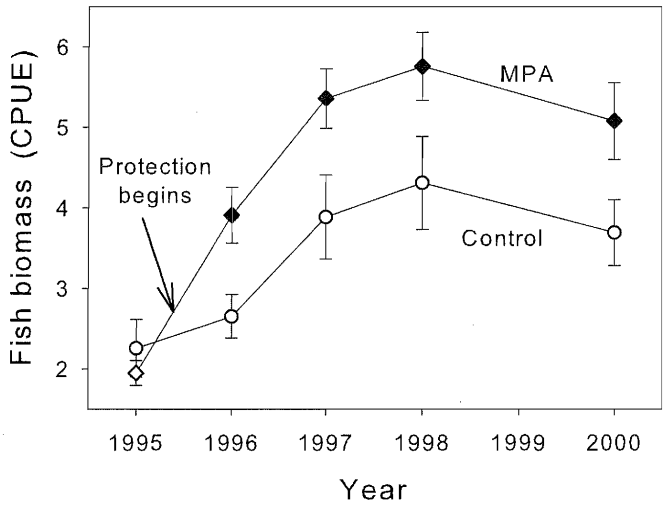


FIGURE 13.6 Relative abundance (CPUE, based on visual counts) of commercially important fishes from the Roberts et al. (2001) study of the MPA network on St. Lucia in the Caribbean. One survey date was available prior to, and four after, establishment of the MPA. Samples were taken inside (MPA) and outside (Control) the protected areas.

and a time-by-location interaction. The interaction was nonsignificant, suggesting little evidence for a differential change with time at the two sites. Taken alone, these data would argue against a local enhancement, but Roberts et al. were interested primarily in the regional effects, so their working hypothesis was that densities both inside and outside the reserve would increase, which they observed.

However, the fish responses observed by Roberts et al. apparently occurred within a year of establishment of the MPA network (Figure 13.6). This rapid response may be plausible for local effects where migration of fishes into MPAs may exaggerate local effects (e.g., note the relatively quick response suggested in Figure 13.5). It seems implausible, however, that *regional* effects would be manifest in a year's time, because they arise primarily through enhanced larval production from increased *adult* stocks or density-dependent movement, probably of older life stages (Russ and Alcala 2003; Russ et al. 2004).

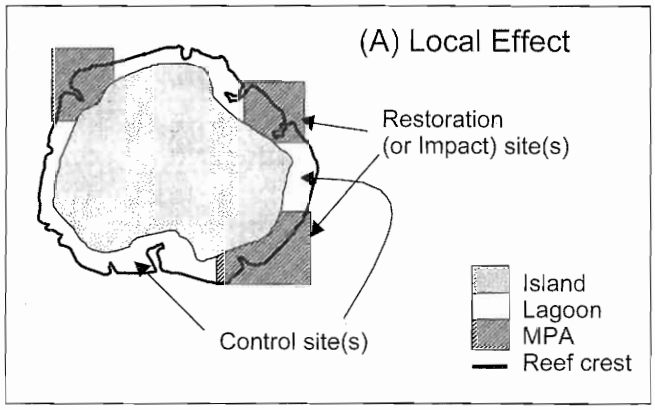
An alternative explanation for this result, as was noted by the authors, is that there was another factor that caused the regional increase in fish stocks. Because fish stocks fluctuate for many reasons, and each Control site was within ~1 km of the nearest reserve, a common response of the Control and Impact sites to another factor is plausible.

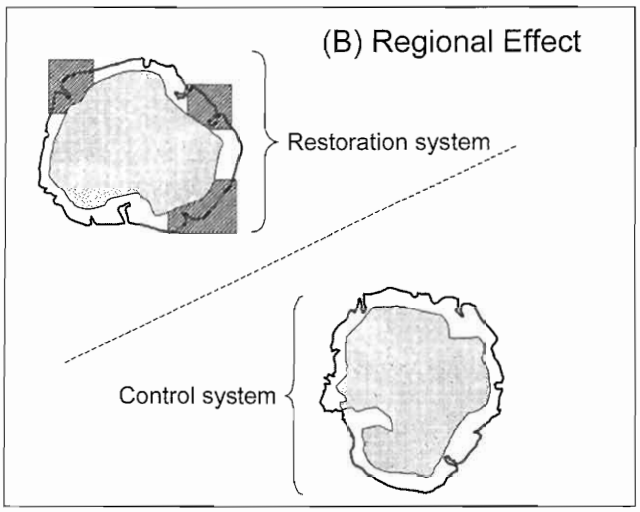
The two competing hypotheses (regional effects of the MPA versus other factors) cannot be distinguished with the available data. Interviews suggested that fishers thought the MPA had worked; however, the fishers might have reasonably, but perhaps falsely, inferred an effect based on the general increase in fish biomass (no matter the cause). The authors observed in a footnote that they had no evidence for similar increases in fish abundance on nearby islands, although quantitative data were not collected. These interviews illustrate a more suitable approach: what is needed is an *appropriate* Control instead of nearby sites that are expected to be influenced by the MPA (see below and Box 13.1).

Box 13.1

Case Study: The Application of BACIPS to Lagoonal Fisheries

Local and Regional BACIPS: BACIPS designs can be used to assess both local (Figure A) and regional (Figure B) effects of marine reserves (MPA) on lagoonal fisheries. Local and regional designs are distinguished by the location of the Control site(s). The regional assess-





ment design (B) has never been used in any system but is feasible for lagoonal fisheries because local retention of larvae is thought to be high (Planes et al. 1993, 1998; Bernardi et al. 2001) and fishing effort localized. Thus, two adjacent islands may have fairly independent lagoonal fisheries yet may be close enough to one another to be affected similarly by oceanographic and weather conditions. One island could serve as the Impact (receiving an MPA network within its lagoons) and one island could serve as a Control (lacking MPAs). Monitoring of population densities, size-structure, and fisheries yields in the two islands, Before and After implementation of the MPA, would provide a test of regional effects of MPAs on fisheries. The expectation is that the fishing yields (and stocks) outside of the MPA, but on the island with an MPA network, would increase relative to the Control island, despite removal of habitat from the fishers.

Similar opportunities likely exist in other restoration projects that may have regional-scale effects, such as prairie restorations designed to rescue other nearby habitats; habitat protection programs for migrating birds or butterflies that affect dynamics on other continents; or fire management regimes that promote local diversity and thus enhance the regional pool and therefore the richness of sites outside of the managed areas.

Coordinated Assessments: X-BACIPS. Consider ten islands in the Indo-Pacific, with five receiving MPA networks and five others remaining as Controls (as in Figure B). If MPAs were assigned at random and each site sampled after MPA enforcement, the resulting data could be analyzed using a standard experimental approach. Although this design could discern average effects, it could not estimate effects of MPAs on any particular island. Instead, if each island pair was sampled Before and After MPA enforcement, then inferences could be made about individual islands (as in assessment designs) and about the population of MPAs as a whole (as with a standard experimental design). Mean effects and variances could be estimated, for example, using mixed-model meta-analysis (Gurevitch and Hedges 1999; Osenberg et al. 1999) or maximum likelihood. Due to the combination of experimental and assessment approaches, we term this design "X-BACIPS."

Regional Assessments: BACIPS with an Appropriate Control

Studies of MPAs highlight the need to better match the spatial scale of interest to the assessment design. To study regional effects of MPAs (or any other restoration effort), we require not a comparison of inside versus outside the MPA, but instead comparison of a region *with* an MPA network with a region *lacking* an MPA (Box 13.1; see Russ 2002 for an alternate design). At least two major scientific problems are likely to arise in implementing such a study: (1) the spatial scale of movement of organisms can be large, suggesting that spillover effects will be fairly dispersed in time and will be hard to detect; and (2) large-scale movement will also require that the Control site be located a sufficient distance from the region with the MPA, thus reducing coherence and the power of the resulting analysis. Fortunately, there are some systems in which such a design is feasible (Box 13.1), and it is imperative that we take advantage of these opportunities.

Summary of Lessons Learned from MPAs

Roberts et al. (2001) and Halpern (2003) are among the best of all available studies of MPA effects. They were constrained by the available data and absence of key design features (such as random assignment, suitable Controls, and Before data). However, if these studies of restoration effects lead, at best, to equivocal results, then it is clear that additional studies (with poorer designs) will lead to even greater equivocation. Thus, we need a better approach (not just more of the same). This is not unique to restoration of marine systems but constrains the assessment of most large-scale restoration projects.

Future Directions

Future assessments will be enhanced through the use of better designs, such as BACIPS, and improved statistical tools. Although statistical tools are important, we believe design issues and the increased use of better designs are even more critical. We conclude with a brief discussion of one analytic tool (Bayesian approaches combined with meta-analysis) that we believe is relevant and provide a final discussion about the application of BACIPS.

Bayesian analyses are increasingly common in the ecological literature, although many ecologists avoid them because of concerns about the subjectivity of the prior distribution of effect sizes. As better assessment studies accumulate (with unbiased estimates of effect sizes and variances), mixed model meta-analyses can be used to quantify the distributions of effect sizes (Gurevitch and Hedges 1999; Osenberg et al. 1999). This distribution, useful in its own right, also can be used to define the prior distribution in a later Bayesian analysis of a new restoration project. Of course, each restoration setting is unique, and the number of potential BACIPS studies will likely be small for many types of restoration activities. Thus, the prior distribution itself will be estimated poorly and should probably carry relatively little weight (obviating one advantage of the Bayesian method).

Crome et al (1996) took a different Bayesian approach and used interviews with different parties involved in forestry practices to assess how much their initial opinions (the priors) would change (as reflected in the posteriors) by a scientific study based on a BACIPS design. This was an innovative way to incorporate an assessment into a public policy arena, attempting to gauge the interaction between public opinion and scientific information. If opposing

sides of an environmental issue are sufficiently intransigent, then even a well-designed scientific study may do little to bring the groups to consensus. Of course, we already may be in such a position today, because past studies used poor designs and led to debate among scientists. In such instances, the public has had little choice but to ignore the science and base their opinions on other matters, like economics.

In far too many cases (as illustrated for marine reserves) assessments of restoration projects are post hoc and lack Before data and, therefore, are open to alternate interpretations. As a result, these scientific studies do little to inform the science or public policy. Moving beyond BA, CI, and BACI-single-sample designs and toward greater reliance on BACIPS designs will not be trivial. Successful application of BACIPS requires planning. For politically charged projects (like MPAs), the science often takes a back seat to social considerations. Sites may be relocated several times during the planning phases. This may prevent the collection of Before data from appropriate sites. However, in many cases, candidate sites are known. Sampling can be done at several sites during the planning phase. One of these will likely become the restoration site (e.g., MPA); the others could be used as Control(s). Some sites may not track the restoration site well and can be dropped later from the study (for discussion of some of these issues, see Stewart-Oaten 1996b). Of course, conducting such a "risky" study requires foresight on the part of funding agencies. This is sometimes possible (Piltz 1996 and Ambrose et al. 1996 give two nice examples). However, if the scientific community does not aspire to conduct BACIPS studies, then regulatory and funding agencies will never support them. By recognizing the limitations of existing studies, we hope to facilitate the execution of better-designed and more informative studies that will lead to the development of more effective, large-scale, restoration activities.

Summary

Rigorous statistical evaluation and sound inference of restoration efforts is difficult to achieve. As a result, quantitative assessments are often missing, incomplete, or misinterpreted. Appropriate analyses must be applied within the broader context of the study design and the limitations of these designs evaluated within the context of the restoration goals (including spatial scale). We presented the central statistical concepts relevant to restoration evaluation and contrasted the strengths and weaknesses of possible approaches, including the Control-Impact, Before-After, and Before-After–Control-Impact designs. We advocate the use of Before-After–Control-Impact Paired Series design because it can be applied to spatially unreplicated interventions, is site-specific, does not require random assignment of sites, and yields defensible estimates of the effect of the restoration activity (rather than *P*-values from null-hypothesis tests). We illustrated advantages and limitations of different approaches through a discussion of studies of marine protected areas and closed by proposing future directions, including the use of more appropriate designs that will address current shortcomings and enhance the practice of restoration ecology.

Acknowledgments

We thank Ray Hilborn, René Galzin, and Caroline Vieux for helpful discussion; Tom Adam and Jackie Wilson for help compiling the MPA literature; Margaret Palmer, Don Falk, and

Joy Zedler for useful comments; and the Minerals Management Service, Florida and National Sea Grant programs, and the NSF (OCE-0242312) for support that contributed to the development of these ideas.

Literature Cited

Allison, G. W., J. Lubchenco, and M. H. Carr. 1998. Marine reserves are necessary but not sufficient for marine conservation. *Ecological Applications* 8:S79–S92.

Ambrose, R. F., R. J. Schmitt, and C. W. Osenberg. 1996. Predicted and observed environmental impacts: Can we foretell ecological change? In *Detecting ecological impacts: Concepts and applications in coastal habitats*, ed. R. J. Schmitt and C. W. Osenberg, 343–367. San Diego: Academic Press.

Anderson, D. R., K. P. Burnham, and W. L. Thompson. 2000. Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management* 64:912–923.

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

Bence, J. R., A. Stewart-Oaten, and S. C. Schroeter. 1996. Estimating the size of an effect from a before-after–control-impact paired series design: The predictive approach applied to a power plant study. In *Detecting ecological impacts: Concepts and applications in coastal habitats*, ed. R. J. Schmitt and C. W. Osenberg, 133–149. San Diego: Academic Press.

Bernardi, G., S. J. Holbrook, and R. J. Schmitt. 2001. Gene flow at three spatial scales in a coral reef fish, the three spot dascyllus, *Dascyllus trimaculatus*. *Marine Biology* 138:457–465.

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.

Cote, I.M., I. Mosqueira, and J. D. Reynolds. 2001. Effects of marine reserve characteristics on the protection of fish populations: A meta-analysis. *Journal of Fish Biology* 59 (suppl. A): 178–189.

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.

Earn, D. J. D., P. Rohani, B. M. Bolker, and B. T. Grenfell. 2000. A simple model for complex dynamical transitions in epidemics. *Science* 287:667–670.

Englund G., and S. D. Cooper. 2003. Scale effects and extrapolation in ecological experiments. *Advances in Ecological Research* 33:161–213.

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

Gurevitch, J., and L. V. Hedges. 1999. Statistical issues in ecological meta-analyses. *Ecology* 80:1142–1149.

Halpern, B. S. 2003. The impact of marine reserves: Do reserves work and does reserve size matter? *Ecological Applications* 13:S117–S137.

Halpern, B. S., and R. R. Warner. 2002. Marine reserves have rapid and lasting effects. *Ecology Letters* 5:361–366.

Johnson, D. M. 1999. The insignificance of statistical significance testing. *Journal of Wildlife Management* 63:763–772.

Lubchenco, J., S. R. Palumbi, S. D. Gaines, and S. Andelman. 2003. Plugging a hole in the ocean: The emerging science of marine reserves. *Ecological Applications* 13:S3–S7.

Magnuson, J. J., B. J. Benson, and T. K. Kratz. 1990. Temporal coherence in the limnology of a suite of lakes in Wisconsin, U.S.A. *Freshwater Biology* 23:145–159.

Mapstone, B. D. 1995. Scalable decision rules for environmental impact studies—Effect size, type-I, and type-II errors. *Ecological Applications* 5:401–410.

Melbourne, B. A., and P. Chesson. 2005. Scaling up population dynamics: Integrating theory and data. *Oecologia*145:179–187.

Michener, W. K. 1997. Quantitatively evaluating restoration experiments: Research design, statistical analysis and data management considerations. *Restoration Ecology* 5:324–337.

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*. San Francisco: California Coastal Commission.

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

Murtaugh, P. A. 2003. On rejection rates of paired intervention analysis: Reply. *Ecology* 84:2799–2802.

Norse, E. A., C. B. Grimes, S. V. Ralston, R. Hilborn, J. C. Castilla, S. R. Palumbi, D. Fraser, and P. Kareiva. 2003. Marine reserves: The best options for our oceans (Forum). *Frontiers in Ecology and the Environment* 1:495–502.

Osenberg, C. W., O. Sarnelle, S. D. Cooper, and R. D. Holt. 1999. Resolving ecological questions through meta-analysis: Goals, metrics and models. *Ecology* 80:1105–1117.

Osenberg, C. W., and R. J. Schmitt. 1996. Detecting ecological impacts caused by human activities. In *Detecting ecological impacts: Concepts and applications in coastal habitats*, ed. R. J. Schmitt and C. W. Osenberg, 3–16. San Diego: Academic Press.

Osenberg, C. W., R. J. Schmitt, S. J. Holbrook, K. E. Abu-Saba, and A. R. Flegal. 1994. Detection of environmental impacts: Natural variability, effect size, and power analysis. *Ecological Applications* 4:16–30.

Osenberg, C. W., R. J. Schmitt, S. J. Holbrook, and D. Canestro. 1992. Spatial scale of ecological effects associated with an open coast discharge of produced water. In *Produced water: Technological/environmental issues and solutions*, ed. J. P. Ray and F. R. Englehardt, 387–402. New York: Plenum Publishing.

Osenberg, C. W., C. M. St. Mary, R. J. Schmitt, S. J. Holbrook, P. Chesson, B. Byrne. 2002. Rethinking ecological inference: Density dependence in reef fishes. *Ecology Letters* 5:715–721.

Palumbi, S. R. 2000. The ecology of marine protected areas. In *Marine ecology: The new synthesis*, ed. M. D. Bertness, 509–530. Sunderland, MA: Sinauer Associates.

Piltz, F. M. 1996. Organization constraints on environmental impact assessment research. In *Detecting ecological impacts: Concepts and applications in coastal habitats*, ed. R. J. Schmitt and C. W. Osenberg, 335–345. San Diego: Academic Press.

Planes, S. 1993. Genetic differentiation in relation to restricted larval dispersal of the convict surgeonfish *Acanthurus triostegus* in French-Polynesia. *Marine Ecology Progress Series* 98 (3): 237–246.

Planes, S., P. Romans, and R. Lecomte-Finiger. 1998. Genetic evidence of closed life-cycles for some coral reef fishes within Taiaro Lagoon (Tuamotu Archipelago, French Polynesia). *Coral Reefs* 17:9–14.

Roberts, C. M., J. A. Bohnsack, F. Gell, J. P. Hawkins, and R. Goodridge. 2001. Effects of marine reserves on adjacent fisheries. *Science* 294:1920–1923.

Russ, G. R. 2002. Yet another review of marine reserves as reef fishery management tools. In *Coral reef fishes: Dynamics and diversity in a complex ecosystem*, ed. P. F. Sale, 421–443. San Diego: Academic Press.

Russ, G. R., and A. C. Alcala. 1996. Marine reserves: Rates and patterns of recovery and decline of large predatory fish. *Ecological Applications* 6:947–961.

Russ, G. R., and A. C. Alcala. 2003. Marine reserves: Rates and patterns of recovery and decline of predator fish, 1983–2000. *Ecological Applications* 13:1553–1565.

Russ, G. R., A. C. Alcala, A. P. Maypa, H. P. Calumpong, and A. T. White. 2004. Marine reserve benefits local fisheries. *Ecological Applications* 14:597–606.

Sanchez Lizaso, J. L., R. Goni, O. Renones, J. A. Garcia Charton, R. Galzin, J. T. Bayle, P. Sanchez Jerez, A. Perez Ruzafa, and A. A. Ramos. 2000. Density dependence in marine protected populations: A review. *Environmental Conservation* 27:144–158.

Scheiner, S. M., and J. Gurevitch, editors. 2001. *Design and analysis of ecological experiments*. New York: Oxford University Press.

Schindler D. W. 1998. Replication versus realism: The need for ecosystem-scale experiments. *Ecosystems* 1:323–334.

Schmitt, R. J., and C. W. Osenberg, editors and contributing authors. 1996. *Detecting ecological impacts: Concepts and applications in coastal habitats*. San Diego: Academic Press.

Schmitz, O. 2005. Scaling from plot experiments to landscapes: Studying grasshoppers to inform forest ecosystem management. *Oecologia* 145:225–234.

Schreuder, H. T., R. Ernst, and H. Ramirez-Maldonado. 2004. *Statistical techniques for sampling and monitoring natural resources*. General Technical Report RMRS-GTR-126. Rocky Mountain Research Station, Fort Collins: USDA, Forest Service. 111 pp. http://www.treesearch.fs.fed.us/.

Schroeter, S. C., J. D. Dixon, J. Kastendiek, R. O. Smith, and J. R. Bence. 1993. Effects of the cooling system for a coastal power plant on kelp forest invertebrates. *Ecological Applications* 3:331–350.

Stewart-Oaten, A. 1996a. Goals in environmental monitoring. In *Detecting ecological impacts: Concepts and applications in coastal habitats*, ed. R. J. Schmitt and C. W. Osenberg, 17–27. San Diego: Academic Press.

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

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

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

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

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

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

Underwood, A. J. 1994. On beyond BACI: Sampling designs that might reliably detect environmental disturbances. *Ecological Applications* 4:3–15.

Underwood, A. J. 1997. *Experiments in ecology: Their logical design and interpretation using analysis of variance*. New York: Cambridge University Press.

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

Osenberg, C.W., B.M. Bolker, J.S. White, C. St. Mary, and J.S. Shima. 2006. Statistical issues and study design in ecological restorations: lessons learned from marine reserves. Pages 280-302 in: Foundations of Restoration Ecology, DA Falk, MA Palmer, and JB Zedler, eds. Island Press.