Journal of Environmental Management (1995) **43**, 67–85

Statistical Considerations in the Design and Analysis of Environmental Damage Assessment Studies

John R. Skalski

Center for Quantitative Sciences, School of Fisheries, University of Washington, Seattle, Washington 98195, U.S.A.

Received 27 *November* 1993

An environmental assessment of chemical and oil-spill effects presents unique statistical problems for both design and analysis. Because of the unplanned and accidental nature of a chemical spill, baseline data are usually absent and the quintessential elements of replication and randomization found in true experiments are absent. Alternative approaches to the design of damage assessment studies are presented that investigate spatial and/or temporal trends in biological responses centered about the spill event. Inferential and logistical limitations of alternative designs are discussed along with consideration of the spatial scale for pseudo-replication and the power of tests of impact. Meta-analysis is suggested to improve both the inferential capabilities and statistical power of damage assessments.

Keywords: additivity, damage assessment, dose–response, *Exxon Valdez*, impact assessment, interversion analysis, Kriging, meta-analysis, NRDA, oil spill, profile analysis, pseudo-replication, sample size, statistical power, time series.

1. Introduction

The word accident implies an event that is unexpected and hopefully infrequent. And with regard to an accidental chemical spill or discharge, the word accident also implies an unfortunate and possibly environmentally damaging event. As a biometrician, I naturally considered working on a major oil spill as a once-in-a-lifetime experience. Certainly, "lightning wouldn't strike twice". Furthermore, the frantic pace of working on a spill assessment needs to be experienced just once for a person to be shy of working on another oil spill. Yet, since 1986 I have worked on two major oil spills. The first was the 239 000-gallon spill by the *Arco Anchorage* on 21 December 1985 in Port Angeles, Washington; the other was the 11 million gallon spill by the *Exxon Valdez* on 24 March, 1989 in Prince William Sound, Alaska. By the very nature of a rare event, lessons learned during the first experience rarely bear fruit in the future. However, having now worked on two oil spills, I have had the opportunity to consider some of the unique problems commonly faced by statisticians attempting to assess chemical-spill effects. The purpose of this paper is to discuss some of the fundamental difficulties

0301–4797/95/010067+19 $08.00/0

statisticians face in environmental damage assessment and some possible solutions. It is my hope that statisticians, biologists and resource managers may benefit from this discussion should lightning strike again.

There are many aspects of a damage assessment requiring statistical involvement. Reconnaissance sampling may be among the first of these activities following a spill, the purpose of which is to determine the locations and concentrations of the contaminant in the environments. Smith (1979) discusses some of the principles of oversampling and Rhode (1979), Elder *et al.* (1981) and Gilbert (1987, pp. 71–88) discuss aspects of sample compositing that may be used to improve spill characterization. Later, sample surveys and hotspot detection (Gilbert, 1987, pp. 119–131) may be used to determine the relative success of clean-up activities and remediation. But throughout these phases, physical, chemical and biological studies may be conducted to assess the effects of oil or other contaminants on the environment and its inhabitants. I am going to restrict my discussions in this paper to the very challenging area of damage assessment. I am going to use examples from the pelagic and subtidal studies conducted by scientists working on the *Exxon Valdez* oil spill to illustrate various statistical issues in the design and analysis of assessment studies.

2. NRDA and statistics

Government regulations outline procedures by which a Federal or State agency acting as a trustee can determine damages and compensation for injuries to the environment from an oil or hazardous waste spill. These procedures for Natural Resource Damage Assessments (NRDA) are called for in the Comprehensive Environmental Response, Compensation, and Liability Act of 1980 (CERCLA) and the Oil Pollution Act of 1990 (UPA). The process has three basic phases: (1) an injury determination phase whose purpose is to determine whether natural resources had been injured; (2) a quantification phase to determine the extent of the injuries; and (3) a damage determination phase that is to establish a dollar amount for the loss of services associated with the injuries.

From a statistical point of view, the first phase of injury determination corresponds to a test of the null hypotheses of no impact. The second phase, that of quantification, is then associated with estimating the magnitude of effects given the null hypotheses of no impact is rejected. These regulations prompt the first of many choices consulting statisticians must make. The assessment study could either be designed to: (1) provide information concurrently for both hypothesis testing and estimation of effects; or (2) be sequentially structured to first test for impact, and only if effects are detected, go on to estimating the magnitude of effects.

The choice of either concurrent or sequential use of hypothesis testing and estimation of injury might be straightforward were it not for the fact these sampling objectives are largely in opposition of one another. To begin, there is never an overabundance of resources for detailed environmental sampling, whether the federal government or a large multinational corporation is conducting the study. Sampling costs and the logistics of conducting pelagic studies are inevitably overwhelmed by the imprecision of available sampling techniques and inherent environmental variability. Acknowledging finite resources for sampling, either the test of impact or estimation of the magnitude of injury is sacrificed to a varying degree by focusing on the alternative.

When designing a study for the sole purpose of testing for impact, the optimal sampling design would concentrate on the extreme conditions. The worst case scenario of the heaviest contaminated areas compared with unaffected sites would be the focus

of the sampling program. The statistical power of the test of impact would be greatest with this dichotomous sampling scheme. Should the absence of effects be adequately demonstrated, the study design itself would be adequate for inferring the lack of effects throughout the entire region of the spill. However, should effects be detected, the very design used to find the effects is now handicapped in the next phase of estimating the magnitude of injury. To estimate the magnitude of the injury, samples must be spatially distributed among all levels of contamination and all locales in order to estimate the regional extent of the effects.

The alternative strategy of estimating the magnitude of the injury focuses on sampling the range and frequency of environmental effects. With this approach, numerous samples will be collected in benign environments and fewer in the most severely stressed sites. The consequence is that the power of the test of impact is sacrificed to a degree in lieu of making inferences to the amount of injury throughout the region of inference.

Engaging in an environmental damage assessment, consulting statisticians must be aware of the duality of objectives and the need to address the competing requirements of detection and estimation of injury. It appears that most assessment studies to date have focused on the test of impact, leaving the estimation of injury to later consideration. An exception is presented by Smith (1979), who combined oversampling to characterize spill location with subsampling of biological responses to estimate the environmental injury resulting from an oil spill. Using grid sampling, Smith (1979) estimated the effects of the spill using a contrast between oiled and non-oiled areas after adjusting for unequal strata weights in a post-stratification based on habitat criteria. The usual emphasis on the test of impact is motivated by the practicalities and difficulties of impact assessment and the overwhelming desire to first and foremost identify if any environmental problems may exist. Given the inherent difficulties of substantiating the existence of an impact, detection of accident-related effects is often considered paramount and adequate in itself for the purpose of injury assessment.

By itself, a biological assessment study of any one species or environmental component will typically be limited in satisfying both phase 1 and phase 2 objectives. However, prospects for successful phase 1 and phase 2 studies is greatly increased by co-ordinating the various biological, physical and chemical sampling programs present in a typical damage assessment. Reconnaissance sampling of sediment and water can be used to characterize contaminant levels and their distributions in the environment. Coupled with dose–response relationships developed from the biological assessment studies, the environmental loads can be used to project the cumulative amount of injury sustained. However, co-ordination of the various aspects of the damage assessment are necessary from the onset, if the required information is going to be collected and the biological significance of effects interpreted.

It may be difficult to have the clarity of thought to balance and integrate the designs of multiple tasks during the frantic activities associated with an event such as an oil-spill clean-up. Nevertheless, it may also be the greatest contribution a statistician can make to the assessment process.

3. Considerations in assessment designs

3.1. INHERENT LIMITATIONS

An assessment of a chemical or oil spill is *not* an experiment. The very nature of an

accidental chemical spill precludes the quintessential elements of randomization and replication found in a true experiment. There is only one spill, and the objective of the assessment is to make inferences to that one event in time and at that specific site. Even if the spill could be replicated, no one desires to replicate the event. Furthermore, the legal mandate (i.e. NRDA) is to assess the effect of that site-specific event and not to make inferences to the general process or the population of historical spills. What makes the situation even more difficult is that the spill site is not randomly selected, nor can it be assumed the contaminant will randomly or uniformly distribute itself across the seascape. Because statistical inferences are intended to be site-specific, as well as, event-specific, investigators using analysis of variance (ANOVA) often advocate the use of fixed-effect models in tests of impact hypotheses (Skalski and McKenzie, 1982; Green, 1984; Millard *et al.*, 1985).

Some locations by their very nature may be more susceptible to contamination following a spill event than others. A simple example is an oil spill in a riverine environment. In such circumstances, the oil will tend to distribute itself downstream of the spill site. Yet the downstream sites may be inherently different from non-oiled upstream sites. In Prince William Sound, the confounding may be more subtle but the potential still exists. For example, north-facing bays and inlets of Knight Island typically received greater oil exposure than south-facing bays that received reduced amounts or no oiling (Figure 1). Hence, oiled versus non-oiled bays are potentially confounded with the geography, aspect and any hydrologic difference that may have influenced both oil transport and the distribution and abundance of marine larvae. Another potentially confounding factor is that oiled sites will tend to be clustered in close proximity while the non-oiled sites will be dispersed about the periphery of the spill (Figure 2). The different dispersion patterns for oiled and non-oiled sites will also contribute to different levels of heterogeneity (i.e. spatial variance) between treatment designations. The National Research Council (1985, pp. 89–93) used a fictitious data set to illustrate the potential confounding of oiling with tidal height and biomass in benthic communities.

Because of the unanticipated nature of an accidental spill, the "when" and the "where" of the eventual event is unknown in advance. This lack of prior knowledge necessarily precludes detailed baseline environmental data at either contaminated or reference sites in the vicinity of the spill zone. The inherent lack of baseline data differentiates chemical or oil spill assessment from more traditional environmental impact studies. Green (1979) describes an "optimal" impact design as having both spatial and temporal controls, wherein reference and potentially impacted sites are monitored prior to and after the environmental event. The lack of temporal baseline data is one feature that uniquely defines an accident assessment and the possible approaches to quantifying effects. Another defining feature is that reference sites are not what one would traditionally call "controls". As discussed earlier, reference and contaminated designations are not randomly assigned to sites. Hence, treatment designations and site-specific differences are completely confounded. Even in the absence of the chemical spill, there is no expectation that the mean response would be the same at reference and the potentially impacted sites. Environmental differences no matter how subtle will always exist between sites, and, in the absence of randomization, these differences will be completely confounded with any effects of the chemical or oil spill. Loehle and Smith (1990) and Loehle *et al.* (1990) discuss the added inferential dangers of trying to assess impacts in the presence of short-term successional charges.

In virtually any statistical test of effects based on analysis of variance (ANOVA),

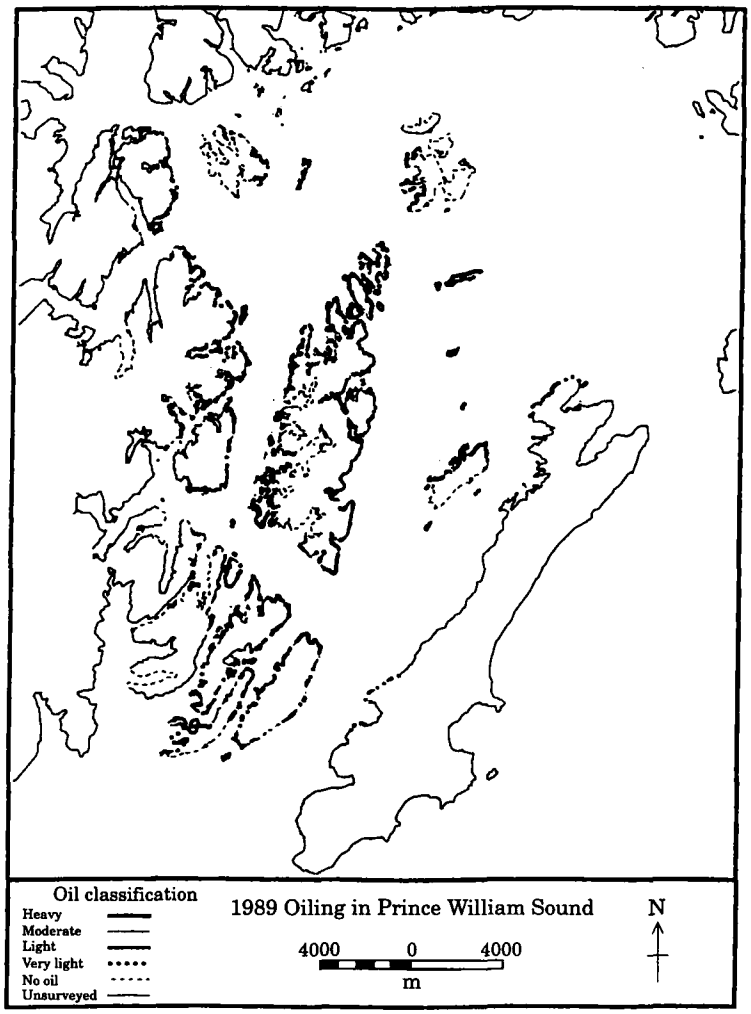

Figure 1. Aerial map of Prince William Sound and distribution of spilled oil along coastal environments.

covariance (ANODEV) or regression, the replicate reference and contaminated sites are assumed to be independent. However, the replicate study sites in a damage assessment are pseudo-replicates (Eberhardt, 1976; Skalski and McKenzie, 1982; Hurlbert, 1984; Stewart-Oaten *et al.*, 1986) because the only true level of replication is at the level of the spill event itself. An experimental unit is typically defined as the smallest geographic unit upon which the treatment was applied. In the case of a damage assessment, that smallest unit is the entire spill or discharge zone. To apply standard statistical techniques, an alternative paradigm for the selection of sampling locations must be used in damage assessments.

The pseudo-replicates by necessity must be defined by the spatial scales that biological responses are operating, and the geographic level that environmental contamination is measured. The choice of spatial scale will have a crucial effect on the degree of

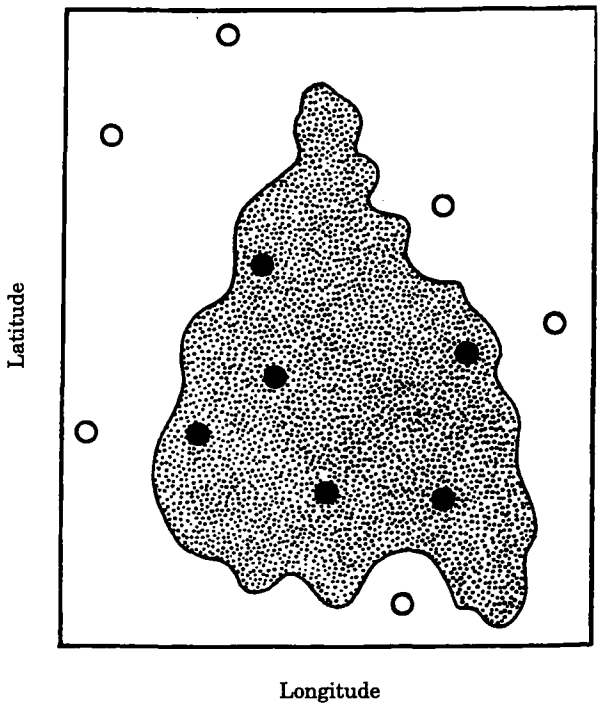

Figure 2. Schematic of a stylized oil spill and the distribution of oiled (○) and non-oiled (●) sites.

independence and how environmental contamination is characterized. A few examples may help clarify these considerations.

In an assessment of the effects of the *Exxon Valdez* oil spill on pink salmon escapement, the natural choice of pseudo-replicate in Prince William Sound is at the level of the stream. The reason is that the smallest scale that escapement occurs is at the level of the stream, not subsections of the stream. The degree of oiling in this case would also be determined at the level of the stream. Alternatively, if the response variable was proportion of salmon eggs in the gravel that hatched, replication could be at the level of a sampling station within the stream if the level of oiling was also measured at such sites. The reason is that the toxic response of salmon eggs to oil might be anticipated to be dominated by local conditions at the level of the sampling station and be little effected by hatching success and contamination elsewhere.

Alternatively, in surveys of crustacean abundance, sampling stations within a bay would not be independent because of the wider-ranging movements of the animals than just about the stations. Furthermore, oil concentrations at individual sampling stations would not be necessarily representative of the exposure of these wide-ranging animals collected at a station. Consequently, in the case of crustacean surveys, responses and level of contamination are best measured at the geographic unit of an oiled or non-oiled bay. A similar argument can be made for tests of effects on sea-bird populations. The wide ranging movements of these species would suggest bays or rookeries as a basic unit of measurement. Using an inappropriate scale for a test of impact can result in biased tests because of the dependences within the data (Millard *et al.*, 1985), and a false sense of precision because of inflated degrees of freedom.

TABLE 1. Hypothetical data on egg mortality of salmon at reference and contaminated sites. Sample size (n), number of dead eggs (x) and observed proportions (p) presented

Replicate	Reference		Contaminated	
	x/n	p	x/n	p
1	1/5	0·20	4/5	0·80
2	3/25	0·12	12/25	0·48
3	180/900	0·20	135/900	0·15
4	15/60	0·25	42/60	0·70
5	1/10	0·10	7/10	0·70
Mean ($\bar{p}$)		0·174		0·566
p pooled	200/1000	0·20	200/1000	0·20

Green (1984) advocates that tests of impacts be based on the between-year variance in response as found in Skalski and McKenzie (1982), and elsewhere. The necessity to base spill assessments on estimates of between-year variability, however, also dictates impact assessments be multi-year studies. Green (1984) further argues that long-term effects should be the focus of spill assessments rather than short-term and readily apparent effects. Certainly statistics are not needed to verify that sea birds died from acute oiling immediately following a major oil spill. Instead, careful design and statistical analysis is essential to assess longer-term effects of an oil spill on the dynamics of a bird population, and these changes in population dynamics must be based on the normal variability in population abundance over time.

Count, binary and binomial data are common in assessment programs. Taking into account not only the natural variability in response but also the distributional nature of the data is important if proper statistical inferences are to be made. The use of data analyses such as GLIM (generalized linear models) is extremely helpful because data can be analyzed using realistic response models and the most appropriate choice of statistical distribution (Numerical Algorithms Group, 1985). Typically, instead of assuming all data are normally distributed and using standard analyses such as analysis of variance (ANOVA), data are analyzed using a variety of statistical distributions (e.g., Poisson, binomial, gamma, etc.) and analysis of deviance (ANODEV) (McCullagh and Nelder, 1983: pp. 26–28, 201).

A simple example illustrates how the choice of statistical analysis influences test results. Table 1 lists hypothetical mortality data for samples of salmon eggs from replicate reference and contaminated streams. Using standard data analysis based on a t-test and logit-transformation of the observed proportions (i.e., p_{ij}, $i=1,2$ and $j=1, \ldots, 5$), mortality is significantly higher at contaminated sites than at reference sites [$P(t_8>3{\cdot}1747)=0{\cdot}0066$]. The average proportion of dead eggs at reference sites is $\bar{p}=0{\cdot}174$ while $\bar{p}=0{\cdot}566$ at contaminated sites. Closer examination of the data reveals however, that for both reference and contaminated sites, the overall proportion of dead eggs is the same $p_{pool}=200/1000=0{\cdot}20$. Using GLIM based on the more appropriate binomial error structure and a logistic link-function, analysis of deviance (ANODEV) results in an asymptotic t-statistic that is not at all significant [$P(t_8>0{\cdot}0)=0{\cdot}50$] for a one-tailed test. Unlike the GLIM analysis, the logistic transformation of the data and subsequent use of normal theory statistics ignores sample sizes for the estimated proportions and the more appropriate way of summarizing the overall test results. Normal approximations become more disparate from the more appropriate binomial

analysis as sample sizes (n) and responses (p_{ij}) vary from replicate to replicate. In environmental studies both sample size and responses can be anticipated to vary greatly, making GLIM-type analysis essential for proper interpretation of assessment data.

In the absence of replication, baseline data and true controls, it has been argued that a damage assessment is *not* a statistically definable problem, and, as such, should be avoided by statisticians. The reasoning is that because the problem cannot be structured in a traditional Fisherian framework, the inferential problem is neither tractable nor desirable to pursue. However, as an environmental statistician, I argue that because of the environmental, social and political consequences of an environmental event as potentially catastrophic as the *Exxon Valdez* oil spill, we cannot limit ourselves to safe and traditional inferential problems. By law, a damage assessment *will* be conducted, and I strongly believe the quality of inferences to site-specific effects can be greatly improved by rigorous statistical guidance. In the next section, I describe various approaches for making sound inferences to the presence or magnitude of environmental effects following chemical spills or discharges.

3.2. DESIGN ALTERNATIVES

Alternatives to the design and analysis of chemical and oil spill assessment studies fall along a continuum of options. The options include the possibility of incorporating either or both of the temporal and spatial dimensions of a spill event. As in classical experimental designs, the more information incorporated in the design of the study, the greater the power of the study to detect effects. The desire for conclusive data is balanced however against the costs, logistics and physical layout of the spill event.

3.2.1. *Spatial trend*

If environmental effects occur as a result of a chemical spill or discharge, it is possible that a spatial gradient of effects radiating from the spill zone may become evident (Figure 3). Hence, grid sampling may be used to collect environmental data throughout the spill zone including both contaminated and uncontaminated sites. The sampling locations would include a gradient from the heaviest to the lightest of contaminated sites. Isopleths of biological response could then be constructed and superimposed on maps of the spill zone to identify spatial trends. Kriging (Journel and Huijbregts, 1978; Verly *et al.*, 1984) procedures could be used to estimate the isopleths and estimate the sampling variance associated with their locations. Variance estimates for the isopleths can be used to assure the observed patterns are real and not random artifacts of the sampling process.

A potential inferential weakness of this spatial trend approach is that the observed biological pattern of responses may have existed prior to the spill. Although remote, there is a possibility that a naturally-occurring spatial pattern may have predated the spill. Without baseline data, or auxiliary information, the potential for a pre-existing pattern cannot be wholly discounted. A simple example illustrating this point is that of a small oil spill resulting from a tanker running aground on a rocky reef. The depth contours about the reef would result in a naturally-occurring pattern of subtidal or pelagic communities that may parallel an oil gradient. Auxiliary information regarding oil burdens would be necessary to differentiate biological responses the result of the oil from the effects of depth.

The detection of a spatial trend may work well if the normal change in biological

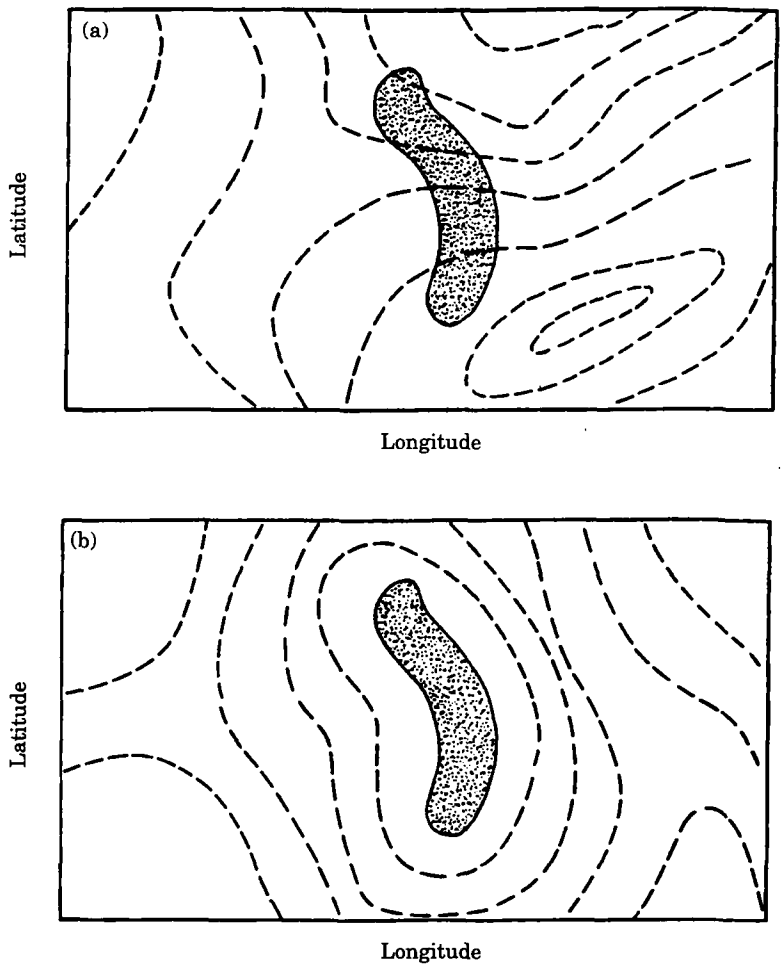

Figure 3. Schematics of possible outcomes of a spatial trend analysis under: (a) null; and (b) alternative hypotheses.

response is relatively small, data are collected using a fine-enough grid and broad regional patterns are desired (Gilbert, 1987, p. 103). The approach will have difficulties on the other hand if spatial correlation is irregular, or sampling is sparse. Consequently, problems may arise if the pattern of the spill is broken or irregular, and the spill zone occurs over a wide range of biological communities. Sampling costs may also prohibit the fine-grained sampling required to characterize a spatial pattern. The application of spatial trend analyses may be best for small spills or subareas of larger events. In these scenarios, the spatial regularity and sampling intensity necessary to precisely characterize spatial patterns in biological response are likely to be fulfilled.

3.2.2. *Spatial–temporal trend*

The inferential weaknesses in asserting the existence of spill effects from an observed spatial trend of biological response can be largely rectified by adding a temporal

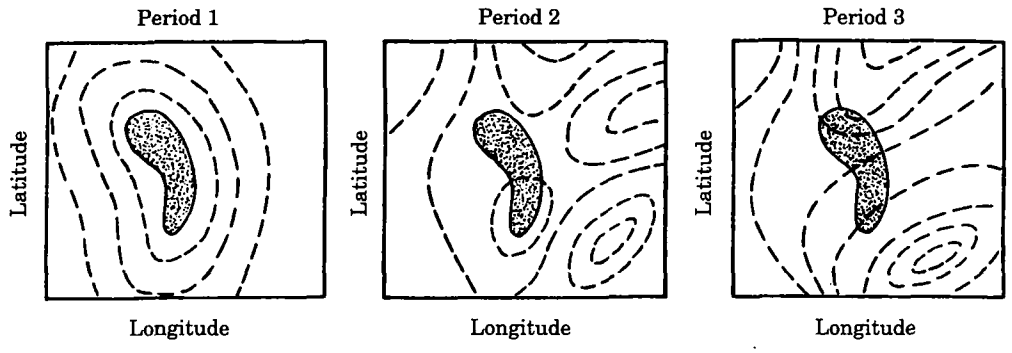

Figure 4. Schematic of an anticipated spatial–temporal trend at an oil site under the alternative hypothesis of an oil-spill effect.

dimension to the data. Although baseline data will typically be non-existent, repeated sampling following the spill can be used to investigate the time trajectory of response in and about the spill zone. The effect of an oil spill will likely be transitory, acute toxicity lasting days to weeks after the spill (Riley *et al.*, 1980; Ward *et al.*, 1980), followed by possible subacute effects lasting months or years (Sanders *et al.*, 1980; Clark, 1982). Under this scenario, any spatial pattern originating from the spill will dissipate over time. Repeated sampling may be used to document a spatial–temporal trend in biological responses (Figure 4) under the alternative hypothesis of spill effects. A preponderance of inferential evidence can be generated by this approach. The likelihood of a spatial–temporal trend in biological response being centered about the the spill event by chance alone is remote, making inferences to causation plausible.

Although more conclusive findings of environmental damage can be documented based on spatial–temporal trends, the statistical limitations previously mentioned regarding the detection of spatial trends still apply. The approach is additionally hampered by the added costs of repeated sampling over time. To date, neither spatial nor spatial–temporal trend analyses have been applied to a major oil-spill assessment study. The lack of the use of either approach is unfortunate for both approaches provide an opportunity to simultaneously gather information for both the injury determination and quantification phases of an oil-spill assessment.

3.2.3. *Time-by-treatment interaction*

The strong inferential capabilities of a spatial–temporal trend analysis suggests retaining both spatial and temporal components in an assessment design. However, the high costs of sampling for pattern suggests reducing the scale of the sampling program to bring costs within reason. One approach is to eliminate the use of a gradient of contaminated sites, focusing instead on a simple time-by-treatment interaction between extreme conditions. Skalski and Robson (1992) proposed a repeated measures study where reference and potentially impacted sites are sampled through time. The contaminated sites would be selected from among the heaviest of polluted sites under this scenario while reference sites would be selected from outside the zone of potential impact.

Lettenmaier *et al.* (1978) demonstrated that the optimal sampling design for assessing impacts resulting in a sudden shift followed by an exponential decline in effects (i.e.

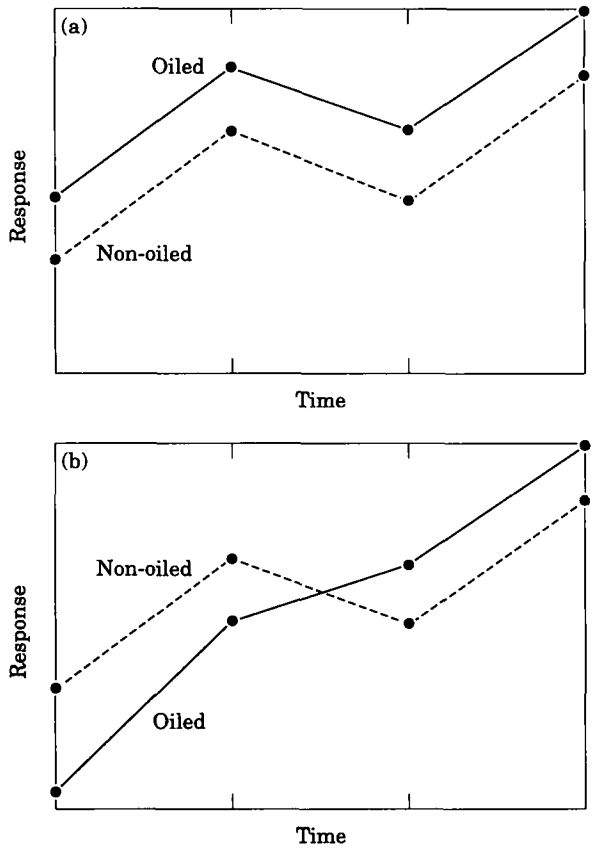

Figure 5. Schematic of possible response profiles for oiled and non-oiled sites in a test of a time-by-treatment interaction under: (a) null hypothesis of no impact; (b) alternative hypothesis of impact.

impulse decay model) is a series of post-event sampling periods. Baseline data under this scenario adds little or no improvement in statistical power. However, the proposed intervention analysis (Box and Tiao, 1973; Hipel *et al.*, 1975), a special case of Box–Jenkins (Box and Jenkins, 1970) models, requires a very large number of samples (i.e. >50) collected uniformly over time in the absence of seasonality to test for effects. For this reason, the alternative of using multiple locations sampled less frequently over time is preferred in practice.

The objective of the test of a time-by-treatment interaction (Skalski and Robson, 1992) is to determine whether the mean time profiles (Morrison, 1976, pp. 205–216) for reference and contaminated sites are parallel (Figure 5). Under the null hypothesis of no impact, the mean time profiles would be parallel but not necessarily coincident [Figure 5(a)]. Because control and treatment designations (i.e. contaminated, respectively) are not randomly assigned, mean response levels would not be expected to be equal even under the null hypotheses of no impact. However, responses to regional climatic conditions would tend to induce similar patterns of biological response over time in the absence of spill or discharge effects. Under the alternative hypothesis of impact, the time profiles at the two types of sites would not track (i.e. not be parallel) until spill effects diminished [Figure 5(b)]. Sequential testing using profile analysis

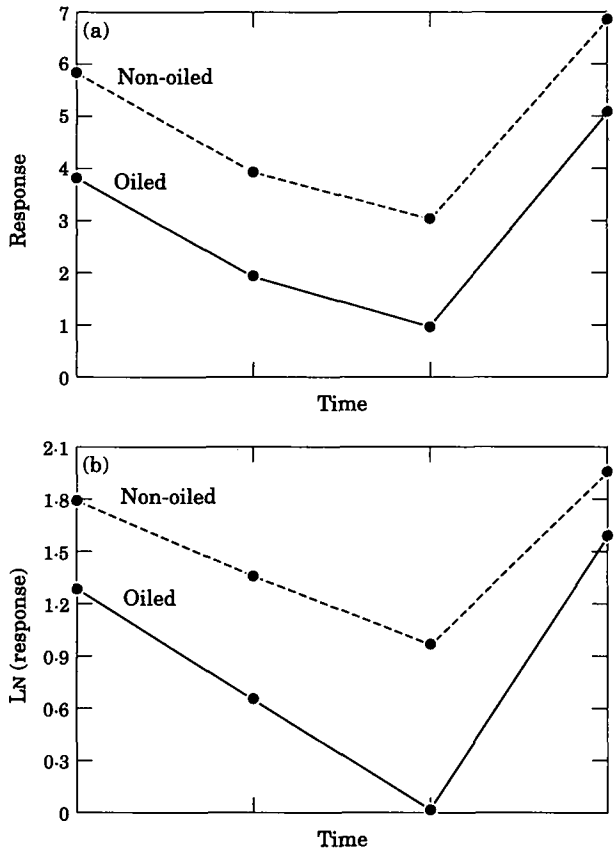

Figure 6. Example of inducing an interaction between time and oil designations by analyzing data on wrong scale. (a) Additive scale; (b) multiplicative scale.

(Skalski and Robson, 1992) would be used to identify the sampling period beyond which time profiles returned to parallel and spill effects no longer exist. Although the test for a time-by-treatment interaction can be used to help identify the temporal extent of the impact, it cannot be used to estimate the spatial extent or magnitude of effects as can the spatial–temporal trend analysis discussed above. Walters *et al.* (1988) recommended a staircase design with a staggered entry of sampling stations into the study to help differential treatment effects from temporal trends across the region. However, applicability of a staircase design is restricted to manipulative experiments.

Skalski and Robson (1992) illustrate that statistical inferences in impact and accident assessments are inherently model-dependent. The necessity for model-dependent inferences is because extraneous and spatial effects are not randomized between treatments as occurs in true experiments. The analysis of a time-by-treatment interaction is an excellent case in point. Typically, a test of interactions is not the focal point of an experiment. However, in a test of spill effects, the test of impact is based on determining whether reference and contaminant profiles are parallel over time (i.e. no interaction between treatment designations and time). Parallelism will be exhibited if reference and contaminant sites respond similarly to climatic events and the data are plotted in the appropriate scale under the null hypothesis of no impact (Figure 6). For instance, if

climatic effects have an additive effect on a biological response, then the time profiles will be parallel only if the data are plotted on the arithmetic scale [Figure 6(a)]. However, if climatic effects have a proportional effect, then the time profiles will be parallel if and only if the data are plotted on the log scale. Analysis of the assessment data on the wrong scale (i.e. use of an inappropriate data transformation) can be misinterpreted as an impact when the time profiles are non-parallel [Figure 6(b)]. Hence, the correct scale of analysis is crucial if a test of interactions is to be properly interpreted. Eberhardt (1978) concluded that the most important consideration in a choice of a data transformation is to achieve additivity. In accident assessment, finding the appropriate scale for data analysis is essential. Tukey's one-degree-of-freedom test of non-additivity in an analysis of variance for a randomized block design (Snedecor and Cochran, 1980, pp. 283–285) should be applied to the sample data from reference sites to determine the proper scale for data analysis. The extreme dependence of valid tests of impact on proper data transformations is a consequence of model-based analyses attempting to compensate for the lack of randomization.

3.2.4. *Dose–response*

Laboratory bioassays have repeatedly shown that crude and refined oil and other chemical products can cause both acute and chronic effects on marine organisms (for a review see Beynon and Corvell, 1974; NRC, 1985, pp. 369–547). By itself, knowing that a contaminant can produce a toxic response is insufficient. For virtually any compound at high enough concentrations will cause acute mortality. Instead, laboratory bioassays provide important information on levels of contamination that may promote toxic response. However, laboratory dose–response results may not accurately mimic environmental effects.

In situ bioassays and regressions of biological responses measured in the environment against contaminant concentrations are necessary to determine whether environmental concentrations, routes of exposure and weathered states of the pollutant produced effects in nature. The purpose of modeling dose–response relationships from biological surveys and environmental chemistry are to: (1) determine whether environmental exposure conditions actually affect biological responses; and (2) demonstrate that sites within the spill zone exist that have contaminant levels and associated biological responses that indicate injury. Detections of a significant dose–response relationship from the biological survey data satisfies the first phase of injury determination.

The real importance of establishing a dose–response relationship comes when combining the toxicity results with surveys of environmental concentrations of the contaminant. With this step, the potential magnitude of the injuries can be assessed. To be useful, wide-spread sampling of concentrations in both suspected contaminated and uncontaminated locations must occur. The detection of only a few heavily contaminated and impacted sites [Figure 7(a)] would suggest regional effects are small. On the other hand, numerous areas at high contaminant levels would suggest widespread effects from the spill or discharge [Figure 7(b)]. Systematic or probabilistic sampling of contaminant concentrations would be necessary to make inferences to the spatial extent of affects.

In any regression analysis of observational data (Cochran, 1983), the sampling units vary because of self-assigned differences in traits that may be systematically or haphazardly assigned. To study a single causal influence such as degree of contamination,

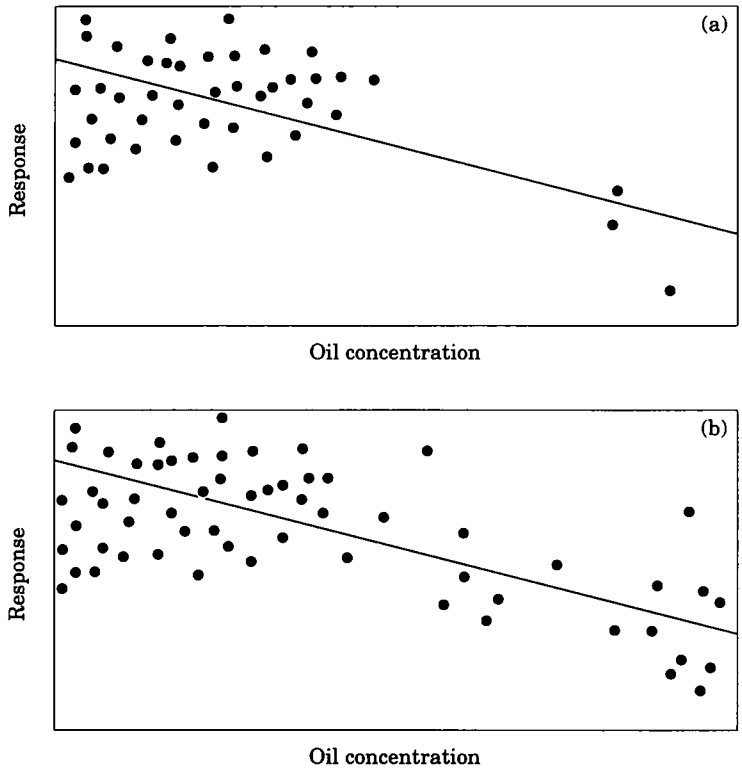

Figure 7. Scatter plots of biological response versus oil concentration under the alternative hypothesis of impact with: (a) few highly contaminated sites; (b) numerous contaminated sites.

the observations must be first adjusted for the effects of all other confounding environmental factors using hierarchical regression analysis [i.e sequential regression based on subject matter considerations, see partial or sequential F-tests (Draper and Smith, 1981, pp. 101–102)]. These adjustments are model-dependent once again and may be no better than our understanding of the natural system being investigated. Inferences to a single causal influence are therefore dependent on adjusting for all other important factors that may systematically affect responses. Because regression models are typically based on simplified relationships between explanatory variables and dependent variables, inferences are less rigorous than in randomized experiments and caution is advised.

3.2.5. *Internally-controlled designs*

The absence of a set of true experimental sites suggests the need to find yet other approaches to assess chemical spill or discharge effects. The biological systems being studied may themselves suggest responses that inherently have control standards. One obvious example is tissue burdens of petrogenic compounds following an oil spill. In pristine environments, contaminant levels would typically be very low. Elevated concentrations of various petrogenic compounds and their metabolites in such circumstances would be self-evident of exposure and potential injury from an oil spill. However, background levels of petrogenic compounds resulting from regional and

global pollution or prior contamination may cloud interpretation of body burdens. Actual assessment of injuries at the organismic, population or community cannot be determined from tissue concentrations alone. Rather, data on tissue burdens provide a valuable additional source of information, that together with other biological and environmental chemistry data, can serve as a building block towards inference of injury.

Another class of potentially useful data is information on the size–frequency distribution of individuals within populations or the age structure of a population. Many pelagic species are characterized by occasions of both extraordinarily strong recruitment as well as failure of age classes (e.g. for salmon see Royer, 1990; Cooney *et al.*, 1992). Consequently, failure of an age class by itself would not constitute conclusive evidence of a spill effect. However, failure of one or more age classes of a species coincident in time and space with a spill event would strongly suggest population effects. A sampling program to test for a time-by-treatment interaction in recruitment would help establish a causal connection with the spill that simple failure could not. Although NRDA regulations preclude the use of stock assessment techniques (Gulland, 1983; Hillborn and Walters, 1992) for measuring injury, direct analysis of age or size structures of populations may provide a useful test of effects, and an important building block in developing an inferential argument for or against a spill affect.

3.2.6. *Other options*

A design option recommended in NRDA regulations is the simple comparison of contaminated and reference (i.e. control) sites. Data analysis might be envisioned as an analysis of covariance (ANOCOV) to test the main effect of contamination designation after adjustment for environmental covariates measured at each site. As mentioned earlier, such designations (e.g. oiled and non-oiled at an oil spill) are completely confounded with site effects because of the lack of randomization. Adjustment for covariates may partially account for site differences, but are model-dependent, and do not guarantee total adjustment for site differences. Even in the absence of contaminant effects, mean response levels for reference and potentially impacted sites are not anticipated to be equal under the null hypothesis.

With repeated measures at each site over time, a factorial analysis of variance or covariance might be envisioned to test the significance of the main effect of contaminant designation. However, any spill effects are still confounded with any site differences plus now the added complication of temporal correlation among the data. Millard *et al.* (1985) discusses the effects of temporal correlation on the significance level of *F*-tests. The profile analysis of Skalski and Robson (1992), on the other hand, explicitly incorporates the temporal correlation to provide a valid test of impact. Irregardless, simple tests of the main effects of contaminant designation are inappropriate in assessing the impacts of a chemical or oil spill. Interpretation of test results whether or not the null hypothesis of no impact is rejected is impossible because of the absence of randomization. Consequently, despite recommendations in NRDA regulations, simple or adjusted comparisons of the mean responses of reference and potentially impacted sites do not provide unequivocal assessments of the effects of a chemical or oil discharge. More complex analyses that try to sort out confounding influences as discussed in previous sections are required to interpret the effects of a chemical spill using quasi-experimental or observational studies.

4. Performance of assessment studies

A study design capable of differentiating effects of a hazardous chemical spill from natural variations in biological response is a necessary but not sufficient requirement for damage assessment. An assessment design must also have sufficient statistical power (i.e. $1-\beta$) to detect biological effects of ecological, economic or social importance if they occur. However, NRDA regulations do not recommend levels of statistical power ($1-\beta$), significance level (α) or the magnitude of effects (Δ) that assessment studies should be designed to detect. For the designs and subsequent analyses to be meaningful, the statistical power of the tests of effects should be used *a priori* in sample size calculations, and *post hoc* to interpret the results of statistical tests that failed to reject the null hypothesis of no injury.

Choice of significance level (α) for tests of impact must be balanced against the chance of not detecting environmental effects of importance. The α-level (type I error) is the probability of falsely declaring an effect when none has occurred. This risk is borne by the responsible party for the chemical spill or discharge. On the other hand, the type II error rate (β) is the probability of not detecting an effect that exists. This type II risk is borne by the public trustees of the environment (i.e. state and federal agencies) and the resource itself. In the absence of regulatory guidance, it seems reasonable for both parties to bear equal risk ($\alpha=\beta$) for a prescribed level of effect (Δ). Both Millard (1987) and Rotenberry and Wiens (1985) recommend balancing α and β error rates if relative "costs" of committing errors can be established. However, environmental costs associated with a type II error may be more difficult to ascribe than the Type I error of a false damage assessment. A decision theoretic approach to establishing study performance may be helpful in these circumstances (Berger, 1985).

There are at least three approaches commonly used in selecting a level of impact considered important to detect (Δ). The first and most commonly used approach is to balance α, β and Δ against available sampling effort. Because of typically high environmental variability and the variance of sampling techniques, values of $\alpha=\beta$ tend to be set high (i.e. $\alpha=0{\cdot}10$ or $0{\cdot}20$). The sensitivity of the test (Δ) is then based on available sampling effort. The second approach is to consider the magnitude of the spatial or temporal fluctuations in response under baseline or reference conditions (i.e. σ^2) and select a Δ that is of the same magnitude as σ or greater. The reasoning behind this choice of Δ is that the environment is accustomed to fluctuations of size σ, and, as such, impacts of that magnitude are not catastrophic, but, rather, populations are accustomed to and resilient to changes of that size. The third and rarely used approach is to seek guidance from demographic models. One can use demographic models to determine how many years would a population take to recover from a decline of size Δ. A value of Δ corresponding to more than 1 or 2 years of recovery would be a candidate for sample-size calculations. Alternatively, demographic models could be used to determine the risk of population failure following a sudden decline of size Δ. A value of Δ could be selected corresponding to a moderate failure rate (i.e. 5, 10 or 20%). In all cases, the economics of recovery time or population failure can be taken into account when selecting a value of Δ. Skalski and Robson (1992) provide power calculations for a repeated measure study of a test of time-by-treatment interactions for accident assessment.

A unique aspect of sample size calculations for accident assessment is that the spill scenario may be self-limiting. For example, in crustacean surveys following the *Exxon Valdez* oil spill, there were a limited number of oiled bays for investigation (i.e. ≤ 8).

To compensate, a greater number of trawl sampling stations within bays was used to reduce the overall variance of a treatment mean ($\bar{x}$), where:

$$\mathrm{Var}(\bar{x}) = \frac{\sigma_B^2}{n} + \frac{\sigma_s^2}{nm},$$

and where σ_B^2 = between bay variance in crustacean density; σ_s^2 = within bay variance in crustacean density, n = number of bays sampled; m = number of trawl samples taken within bays.

However, statistical performance can still be stymied. Increasing the number of sampling stations within bay decreases the contribution of the second variance component (i.e. σ_s^2) but does not decrease the contribution of the between-bay variability (i.e. σ_B^2). At some point, increased subsampling (m) has little or no effect on test performance. Furthermore, in some bays, the substrate limited the number of trawlable areas to 1 or 2 stations per bay. The performance of tests of effects in such circumstances may be preordained by the size of the spill and its location. Increasing the number of non-oiled sites and number of sampling periods can mitigate such limitations only to a degree.

Compensating for limitations of any one investigation is the multiplicity of response variables that can be investigated within and between species and communities. For example, in an investigation of subtidal invertebrates, the abundance, tissue burdens, size frequency and fecundity of several different species may be investigated concurrently. Each response provides an opportunity to detect effects. Just as there are test-wise and experimental-wise type I error rates there are also test-wise and experimental-wise statistical powers to detect effects. In the case of k independent tests of effects each with power $1-\beta_i (i=1 \cdots k)$, the experimental-wise power is:

$$\mathrm{Power}_{\mathrm{EXP}} = 1 - \prod_{i=1}^{k} \beta_i.$$

Because of the potential limitation in the power of any one analysis, the results of related studies need to be combined. Statistical methods of integrating the results of separate studies and investigations are referred to as meta-analysis (for reviews see Hedges and Olkin, 1985; Wolf, 1986). Interpretation of assessment study results should take into account both an overall significance level of the multiple tests and the empirical pattern of the findings. For example, Rotenberry and Wiens (1985) explored the pattern of P-values from 91 different pair-wise correlations between bird densities when interpreting the results of a study of interspecies competition. The purpose is to not only increase the statistical power of the investigations but to determine the internal consistency of the findings. In this way, a cogent argument supporting a conclusion of injury or no injury can be developed.

5. Conclusion

Damage and injury assessment following a chemical or oil spill is an intriguing problem in statistical inference of social and environmental consequence. However, the general public may not appreciate the inferential difficulties of establishing a causal relationship between biological response and a chemical spill or discharge. The difficulties arise because of the absence of replication, randomization and baseline data. Any effects of

the spill are totally confounded with effects of time and location. Hence, classical experimental approaches to injury assessment are not appropriate. Instead alternative approaches tailored to damage assessment must be employed. These approaches may include model dependent inference, and the incorporation of both spatial and temporal data in the tests of effects.

A successful damage assessment has many of the features of a well-crafted detective investigation. The statistician's role is to design investigations that can serve as building blocks in constructing an inferential argument. The individual component investigations as well as the overall assessment requires careful attention to detail. Multifaceted investigations of biological communities and environmental chemistry provide the opportunity to evaluate alternative routes of exposure and effects, increase the overall power of the assessment, judge the internal consistency of test results, and develop a coherent argument for rejecting or not rejecting the null hypotheses of no injury. An important aspect of the design and analysis of assessment studies is associated sample size and power calculations.

The non-Fisherian nature of damage assessment can be argued as a reason for statisticians to avoid participation in such studies. On the other hand, the inevitable need to summarize results and draw sound conclusions from damage assessments following an oil spill or chemical release argue for the need to more thoroughly investigate the inferential nature of quasi-experiments (Cook and Campbell, 1979). Similar inferential problems also exist in assessing global warming, acid rain, impacts of nuclear and coal-fired power plants, and the wildfires in Yellowstone National Park. All are examples of non-replicated, non-randomized investigations of environmental effects of public concern. Greater discussion and instruction of non-classical statistical inference in our colleges and universities is needed to improve the performance of environmental assessment studies (Millard, 1987; Ross, 1987; Price, 1987).

References

Berger, J. O. (1985). *Statistical Decision Theory and Bayerian Analysis.* New York: Springer-Verlag.

Beynon, L. R. and Cowell, E. B. (1974). *Ecological Aspects of Toxicity Testing of Oils and Dispersants.* New York: J. Wiley & Sons.

Box, G. E. P. and Tiao, G. C. (1973). *Intervention Analyses with Applications to Economic and Environmental Problems.* Technical Report 355. Department of Statistics. University of Wisconsin, Madison.

Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control.* San Francisco: Holden-Day.

Clark, R. B. (1982). The long-term effects of oil pollution on marine populations, communities and ecosystems. Proceedings: A Royal Society Discussion meeting held on 28 and 29 October 1981. London: The Royal Society.

Cochran, W. G. (1983). *Planning and Analysis of Observational Studies.* New York: J. Wiley & Sons.

Cook, T. D. and Campbell, D. T. (1979). *Quasi-experimentation: Design and Analysis Issues for Field Settings.* Boston: Houghton Miffin.

Cooney, R. T., Willette, T. M., Sharr, J., Shaup, D. and Olsen, J. (1992). The effect of climate on North Pacific pink salmon production: examining the details of a natural experiment. In *International Symposium on Climate Change and Northern Fish Populations.* 13–16 October 1992, Victoria, B.C.

Draper, N. and Smith, H. (1981). *Applied Regression Analysis.* New York: J. Wiley & Sons.

Eberhardt, L. L. (1976). Quantitative ecology and impact assessment. *Journal of Environmental Management.* **4**, 27–70.

Eberhardt, L. L. (1978). Appraising variability in population studies. *Journal of Wildlife Management.* **42**, 207–238.

Elder, S. R., Thompson, W. O. and Myers, R. H. (1981). Properties of composite sampling procedures. *Technometrics* **22**, 179–186.

Gilbert, R. O. (1987). *Statistical Methods for Environmental Pollution Monitoring.* New York: Van Nostrand Reinhold.

Green, R. H. (1979). *Sampling Design and Statistical Methods for Environmental Biologists.* New York: J. Wiley & Sons.

Green, R. H. (1984). Statistical and nonstatistical considerations for environmental monitoring studies. *Environmental Monitoring Assessments* **4**, 293–301.
Gulland, J. A. (1983). *Fish Stock Assessment: A Manual of Basic Methods*. New York: J. Wiley & Sons.
Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta-analysis*. San Diego: Academic Press.
Hilborn, R. and Walters, C. J. (1992). *Quantitative Stock Assessment: Choice, Dynamics and Uncertainty*. New York: Chapman and Hall.
Hipel, K. W., Lennox, W. C., Unny, T. E. and McLeod, A. I. (1975). Intervention analysis in water resources. *Water Resources Research* **11**, 855–861.
Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecology Monographs* **54**, 187–211.
Journel, A. G. and Huijbregts, C. H. J. (1978). *Mining Geostatics*. New York: Academic Press.
Lettenmaier, D. P., Hipel, K. W. and McLeod, A. I. (1978). Assessment of environmental impacts part two: data collection. *Environmental Management* **2**, 537–554.
Loehle, C. and Smith, E. P. (1990). An assessment methodology for successional systems. II. Statistical tests and specific examples. *Environmental Management* **14**, 259–268.
Loehle, C., Gladden, J. and Smith E. (1990). An assessment methodology for successional systems. I. Null models and the regulatory framework. *Environmental Management* **14**, 249–258.
McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*. New York: Chapman and Hall.
Millard, S. P., Yearsley, J. R. and Lettenmaier, D. P. (1985). Space-time correlation and its effects on methods for detecting aquatic ecological change. *Canadian Journal of Fisheries and Aquatic Sciences* **42**, 1391–1400.
Millard, S. P. (1987). Environmental monitoring, statistics, and the law: Room for improvement. *Journal of the American Statisical Association* **41**, 249–253.
Morrison, D. F. (1976). *Multivariate Statistical Methods*. New York: McGraw-Hill.
National Research Council (1985). *Oil in the Sea: Inputs, Fates and Effects*. Washington, D.C.: National Academy Press.
Numerical Algorithms Group. (1985). *The GLIM System Release 3.77 Manual*. Downers Grove, Illinois: Numerical Algorithms Group, Inc.
Price, B. (1987). Comment. *American Statistician* **41**, 257–259.
Riley, R. G., Thomas, B. L. Anderson, J. W. and Bean, R. M. (1980). Changes in the volatile hydrocarbon content of Prudhoe Bay crude oil treated under different simulated weathering conditions. *Marine Environmental Research*. **4**, 109–119
Rohde, C. A. (1979). Batch, bulk and composite sampling. In *Sampling Biological Populations* (R. M. Carmach, G. P. Patil and D. G. Robson, eds). Fairland, Maryland: International Cooperative Publishing House.
Ross, N. P. (1987). Comment. *American Statistician* **41**, 254–256.
Rottenberry, J. T. and Wiens, J. A. (1985). Statistical power analysis and community-wide patterns. *American Nature* **125**, 164–168.
Royer, T. C. (1990). High latitude oceanic variability associated with the 18.6 year nodal tide. In *Proceedings of the International Conference on the Role of the Polar Regions in Global Change*, 11–15 June, 1990. Fairbanks: University of Alaska.
Sanders, H. L., Grassie, J. F., Hampson, G. R., Morse, L. S., Garner-Price, S. and Jones, C. C. (1980). Anatomy of an oil spill: Long-term effects from the grounding of the barge *Florida* off West Falmouth, Massachusetts. *Journal of Marine Research* **38**, 265–380.
Skalski, J. R. and McKenzie, D. H. (1982). A design for aquatic monitoring programs. *Journal of Environmental Management*. **14**, 237–251.
Skalski, J. R. and Robson, D. S. (1992). *Techniques for Wildlife Investigations: Design and Analysis of Capture Data*. San Diego: Academic Press.
Smith, W. (1979). An oil spill sampling strategy. In *Sampling biological populations* (R. M. Carmach, G. B. Patil and D. G. Robson, eds), pp. 355–364. Fairland, Maryland: International Cooperative Publishing House.
Snedecor, G. W. and Cochran, W. G. (1980). *Statistical methods*. 7th edn. Aimes, Iowa: Iowa St. University Press.
Stewart-Oaten, A., Murdoch, W. W. and Parker, K. R. (1986). Environmental impact assessment: "Pseudo-replicator" in time? *Ecology* **67**, 929–940.
Verly, G., David, M., Journel, A. G. and Marechal, A. (1984). *Geostatistics for Natural Resources Characterization. Parts 1&2. NATO ASI Series C: Mathematical and Physical Sciences*, Vol. 122. Boston: Reidel.
Walters, C. J., Collie, J. G. and Webb, T. (1988). Experimental designs for estimating transient response to management disturbances.*Canadian Journal of Fisheries and Aquatic Sciences* **45**, 530–538.
Ward, D. M., Atlas, R. M., Boehm, P. D. and Calder, J. A. (1980). Biogradation and chemical evolution of oil from the *Amoco* microbial spill. *Ambio* **9**, 227–283.
Wolf, F. (1986). *Meta-analysis: Quantitative Methods for Research Synthesis*. Beverly Hills, California: Sage Press.