American Fisheries Society Symposium 18:639–644, 1996

Detecting Population Impacts from Oil Spills: A Comparison of Methodologies

RAY HILBORN
University of Washington, School of Fisheries WH-10, Seattle, Washington 98195, USA

Abstract.—Five alternative methods for determining oil spill impacts at the population level are compared: (1) counts of dead animals, (2) pre- and postspill comparison of abundance, (3) oiled versus non-oiled comparison of abundance, (4) oiled versus non-oiled comparison of vital rates, and (5) direct experimental oiling. Counts of dead animals do not provide evidence of population-level impact. Pre- and postspill comparisons of abundance have very low statistical power and require substantial baseline data. Oiled versus non-oiled comparisons suffer from lack of randomization of oiling treatments. Experimental oiling must be shown to be comparable to the actual infield oil exposure. A strong case for oil impacts will usually require several of these types of data. It is suggested that traditional hypothesis tests are inappropriate for determination of oil spill impacts and that explicit calculation of the likelihood of different levels of impact is more appropriate.

The *Exxon Valdez* oil spill occurred on 24 March 1989. Tens of millions of dollars were spent to determine the extent of damage caused by the spill. Although the damage from the spill takes many forms, most of the scientific research has been aimed at detecting evidence of population-level impacts on the invertebrates, fishes, birds, and marine mammals in the spill area. There is a widespread feeling that the studies are not nearly as conclusive as many had hoped, and this has raised concern about the ability to detect oil spill impacts by existing methods.

The purpose of this article is to review methodological approaches used to explore population-level impacts resulting from the *Exxon Valdez* oil spill, which have a general application to the problems of detecting any sort of impact, positive or negative, on natural populations. First, I discuss alternative experimental designs for assessment of impacts; next, I consider the nature of statistical tests that should or could be employed in such analyses. Finally, I consider the implications of the different methods to the design of baseline studies before disturbances and of monitoring studies to follow recovery of disturbed systems.

Alternative Methods

There are two general ways that oil spills affect the abundance of a population: through direct mortality or through impacts on reproduction and survival. In each case the impacts might be followed by recovery to pre-impact levels or by a permanent change in abundance. Although most normal models of population dynamics predict recovery once the oil-induced changes disappear, permanent habitat change or a change in competitive or predation pressure could result in a long-term change in the abundance of the species. Holling (1973) provides a review of the types of system behavior that could result from perturbations such as an oil spill. In this section I consider the experimental design of detecting changes in abundance.

There are at least five ways to detect population-level impacts: (1) direct body counts of the number of animals killed, (2) pre- versus postspill comparison of population sizes, (3) oiled versus non-oiled spatial comparison of abundance after a spill, (4) direct measurement of vital rates in oiled versus non-oiled sites, and (5) experimental oil treatments. Each of these methods has advantages and disadvantages and may be the appropriate method for detecting impacts in some circumstances.

Body Counts

Sometimes the number of individuals killed by the spill and can be measured or estimated directly. Such counts are available for several bird and marine mammal species killed as a result of the *Exxon Valdez* oil spill. However, body counts, by themselves, do not provide any evidence of population-level impact. The body counts must be considered in relation to population abundance, natality, mortality rate, and behavior. For instance, Bowman et al. (1995) estimated that 902 bald eagles *Haliaeetus leucocephalus* (11% of the population) were killed directly by the spill, but they could detect no differences between pre- and postspill abundance. This finding could be a result of the high variance in the surveys, or it could be attributable to the fact that a one-time mortality of 11% is not detectable given the background variation in year-to-year recruitment and survival, or it could result from compen-

satory changes in births and deaths. In contrast, DeGange et al. (1994) estimated that 4,028 sea otters were directly killed by the spill, which, combined with pre- and postspill surveys, indicated that the sea otter *Enhydra lutris* population dropped by roughly 50% because of the spill.

Any assessment of population-level impacts using body counts needs to be supported by either direct comparisons of pre- and postspill abundance, oiled versus non-oiled comparisons of abundance, or a population dynamics model that accounts for recruitment and survival.

Pre- and Postspill Comparison

When prespill abundance surveys are available, comparison of pre- and postspill numbers can be used to assess the change in population. The statistical power of such comparison will depend upon the reliability of the census method, the natural variability of the population, and the magnitude of change induced by the spill. This method clearly can not be used when no prespill abundance data are available, as was the case with many fish species affected by the *Exxon Valdez* oil spill. However, pre- and postspill comparisons were effective in showing changes for many species, including sea otters (Garrott et al. 1993), and pigeon guillemots *Cepphus columba* (Oakley and Kuletz 1996, this volume).

Even when prespill data are available, the comparison may be of little value. Pink salmon *Oncorhynchus gorbuscha*, for instance, show very high year-to-year variability, and unless a spill is catastrophic, there is little chance of detecting its impact. In fact, pre- and postspill comparisons may be deceiving. Geiger et al. (1996, this volume) estimated a loss of several million pink salmon because of oiling in 1990, even though the run of pink salmon in 1990 was the largest in history.

A significant decline in abundance after a spill is not necessarily evidence of an oil impact. Populations often vary in abundance and without evidence of a mechanism for oil impact, such as direct body counts, or oiled and non-oiled comparisons, a decline in abundance is not strong proof of an impact of the spill. The disappearance of a number of killer whales from a single pod could be considered evidence of an oil impact, but without supporting mechanisms this evidence of a spill impact is subject to question.

Oiled versus Non-oiled Comparison of Abundance

Probably the most common technique used in assessing damage from the *Exxon Valdez* oil spill is postspill comparison of the abundance of a species in oiled sites to its abundance in non-oiled sites. This technique formed the basis of most intertidal and subtidal assessments (Collier et al. 1996, this volume; Highsmith et al. 1996, this volume; Jewett et al. 1996, this volume). As with pre- and postspill comparison, the power of this method depends on the reliability of the census method, the natural variability from site to site, and the magnitude of the change induced by the spill. A key problem posed by this approach is the fact that beaches were oiled as a result of physical processes, whereas in a designed experiment they would have been oiled by random assignment. Thus, postspill differences may reflect underlying habitat differences rather than the impacts of oiling. This nonrandomization of treatments is in itself not correctable by postspill analysis, but most investigators attempt to determine if there are other differences between sites and have generally tried to choose control sites (non-oiled sites) that are, to the human observer, as comparable as possible to the oiled sites.

The most convincing data from oiled versus non-oiled comparisons occur when the oiled site recovers to the same abundance as the non-oiled site during the course of the postspill evaluation, this strongly suggests that the observed differences immediately after the spill were caused by an oil impact.

Oiled versus Non-oiled Comparison of Vital Rates

An alternative approach is to measure life history parameters in oiled and non-oiled sites. The estimated parameters are used in a life history model to estimate population-level impacts. The differences in growth observed in oiled versus non-oiled sites (Hepler et al. 1996, this volume), and in egg survival for pink salmon (Bue et al. 1996, this volume) are examples of how this approach can provide evidence of damage even where population-level damages may be difficult to measure directly. The weakness of this approach is that it depends on the validity of the population dynamics models used, and in most cases the extent of damage depends on the level of compensatory mortality in the life history after the damage. If there is high density-dependent mortality, the population-level impacts will be much less than the mortality caused by oil-

ing. The potential for compensatory mortality significantly decreases the power of this approach.

Comparison of vital rates between oiled versus non-oiled sites also suffers from the weaknesses of non-randomization of treatments discussed in the previous section. Again, if vital rates in oiled sites recover to the levels of vital rates in non-oiled sites, the argument that the differences were caused by oil effects rather than pre-existing is much stronger. The continued differences in pink salmon egg survival between oiled and non-oiled sites (Bue et al. 1996) need a more complex mechanism to explain the oil impact and actually weaken the argument that the differences are attributable to oil.

Experimental Oil Treatments

A final approach is the experimental application of oil to determine impacts. Such experimentation normally takes the form of laboratory studies where individuals are exposed to oil at concentrations comparable to that found in the field. This approach was used by Carls et al. (1996, this volume) on juvenile pink salmon. The advantage of this method is that direct experimental treatment and control allow clear determination of the oil impact. The disadvantage is that even when oil impacts are determined, it must be established that the individuals in the field received the same oil exposure and that the measured impacts would have a population-level effect. These two conditions are rarely met.

A more direct approach would be field application of oil to randomly selected sites, with similar randomly selected control sites. Although such ideas were frequently suggested by scientists during the damage assessment process, it was widely felt that infield oiling was unacceptable to the public.

Appropriate Statistics

A final issue to be considered is the statistical framework for analysis of damage studies. Almost all scientists are taught a very rigorous paradigm for testing hypotheses, accepting or rejecting the null hypothesis by comparing it to a working hypothesis. This method is found in every statistics textbook. An obvious question encountered in any study of oil spill impacts is what level of significance should be used. Scientific journals normally accept the 0.05 or 0.01 level of significance. If we were to apply this to an oil spill, the obvious conclusion would be that if you could not show an impact at the 0.05 level, you should conclude that there was no oil impact. This stands in contrast to the legal criterion that damage claims need to show that there was "more likely than not" an oil impact. The statistics taught in our

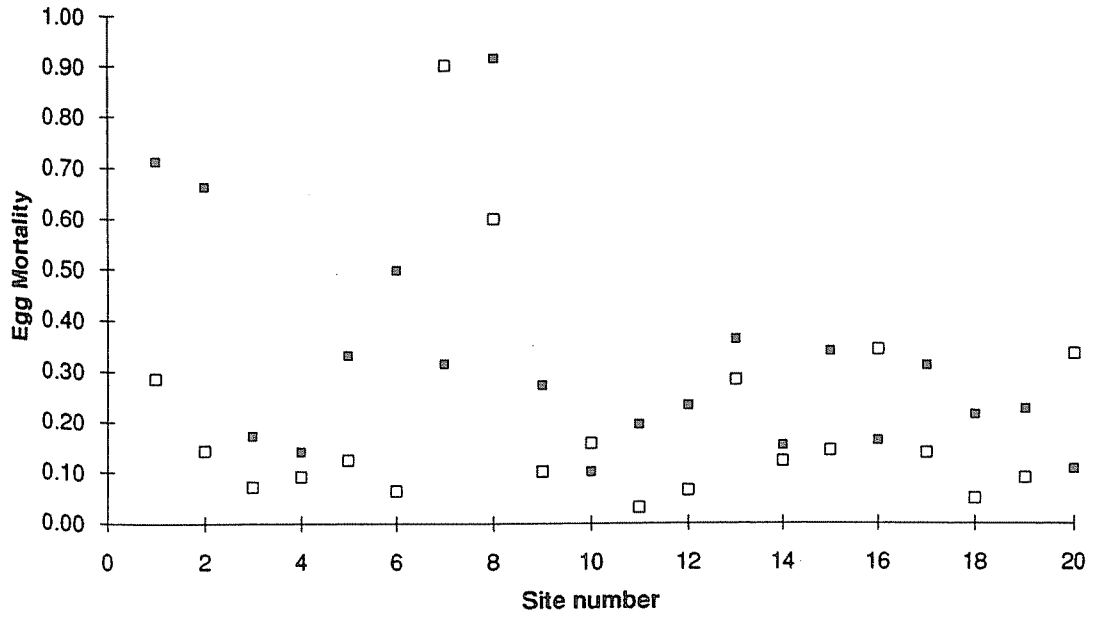

FIGURE 1.—Hypothetical data for pink salmon egg mortality in 20 paired oiled (black squares) and non-oiled (open squares) sites.

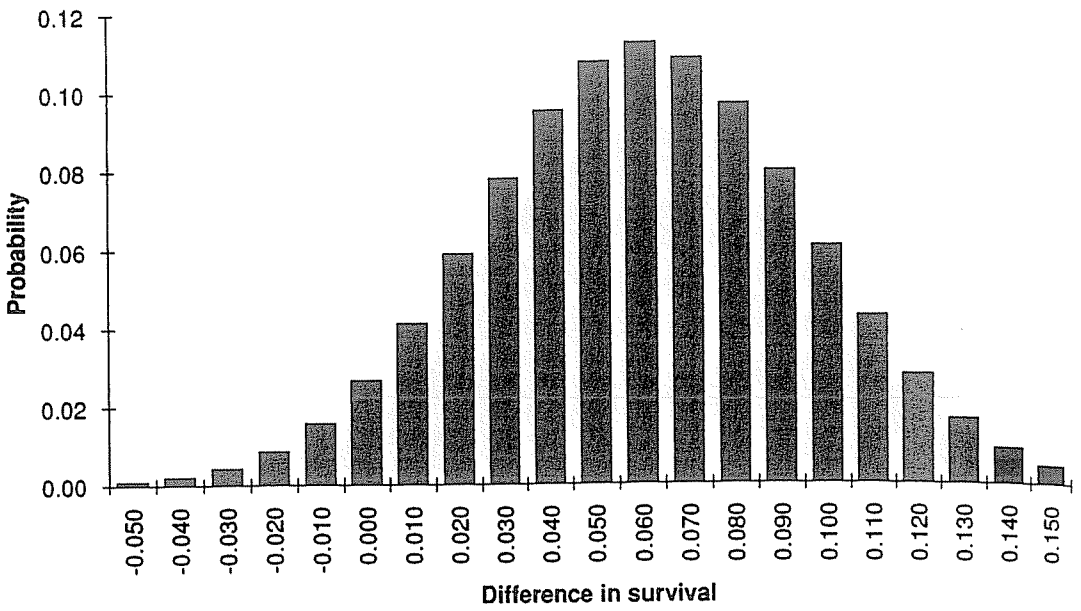

FIGURE 2.—The probability of differences in survival of pink salmon eggs between oiled and non-oiled sites from the data shown in Figure 1.

universities simply do not prepare the scientist to answer this type of question in a rigorous way.

Fortunately, almost all of the body of current statistics can be used to ask this question in a statistically rigorous fashion if we abandon the format of comparing a working hypothesis to a null hypothesis and simply look at the relative likelihood or Bayesian posterior distribution of different levels of oil impact. Methods of likelihood and Bayesian statistics are becoming more common in journal articles, but the statistics books in general use have not treated these methods in much detail. The most popular reference work on likelihood is Edwards (1972); Berger (1985) is a standard reference on Bayesian methods. The essence of likelihood is to look at the relative support provided for different values of a parameter. For instance, Figure 1 shows some hypothetical data on pink salmon egg mortality from 20 oiled (black squares) and 20 non-oiled sites (open squares). The "traditional" statistical approach would be to determine if the average mortality was different at the 0.05 or 0.01 level (see Bue et al. 1996). The relative likelihood of different levels of oil impact on mortality is shown in Figure 2. We can see that the data suggest the most likely impact of oil was a 6% increase in egg mortality, but the difference could be between −1% and 10%. Using Figure 2, a scientist could state that the most likely impact was a 6% difference in survival, but that there was a small chance (about 6%) that the average survival in oiled sites was as high or higher than in non-oiled sites. In contrast, a traditional statistical analysis would state that the null hypothesis that there was no difference between oiled and non-oiled sites could not be rejected at the 0.05 level.

There are two major advantages to be obtained by looking at likelihoods or Bayesian statistics rather than hypothesis tests. First, you avoid the common confusion between statistical and biological significance. Oiled versus non-oiled sites may be statistically significantly different, even if the difference is negligible. Any difference can be statistically significant if the sample size is large enough. When you look at the likelihood, you can see the magnitude of the difference (the units on the x-axis). A second advantage of using likelihoods is that if restoration actions are considered, there is a need to consider the consequences under different hypotheses about the intensity of damage. Likelihood lends itself to the machinery of statistical decision theory, whereas hypothesis testing does not.

Another statistical problem posed in damage assessment is pseudo-replication (Hurlbert 1984). As an example, consider a pre- and postspill comparison of abundance. The more data points that are

available pre-and postspill, the more powerful the comparison would appear. Even if there are large differences between pre-and postspill, you really only have one observation at each site. Unless you can compare multiple sites pre- and postspill, the ability to say the differences result from oiling is weak.

Implications for Baseline Studies and Postspill Monitoring

Perhaps the biggest lesson learned from damage assessment of the *Exxon Valdez* oil spill is the usefulness of prespill baseline data. When only postspill data are available, the results are always suspect because of the nonrandomization of treatments. The most convincing case for oil damage is to show abundance in oiled versus non-oiled areas to be the same before a spill, a postspill decline in the oiled areas, and a gradual recovery to prespill levels. This is an ideal that I believe was not obtained with any study in Prince William Sound. However, prespill baseline studies on numerous birds and mammals in particular were important parts of the evidence for oil damage.

The limitations of baseline studies were illustrated in the case of salmon and herring. In both cases, prespill data existed as part of the normal commercial fisheries management. However, the high year-to-year variability in recruitment and survival meant that pre- and postspill comparisons could only detect highly significant impacts. Even if postspill abundance decreased by half, a number of other mechanisms, including climatic shift, fishery harvests, and hatcheries could be invoked to question the impact of oil. Without oiled versus non-oiled comparisons, changes in pre-and postspill abundance would be weak evidence for oil spill impacts.

In the absence of prespill baseline data, a long time series of postspill monitoring assumes great importance. If the oiled sites gradually converge toward the abundance in non-oiled sites, the case that the differences observed postspill were the result of oiling is strengthened. If the oiled sites do not converge toward the non-oiled sites, we are trapped in the dilemma that if the difference is from oil the damage is greater but our degree of belief that the damage was from oil will be weaker. In either case, we need to follow the oiled and non-oiled sites for many years after the spill to determine if convergence has occurred.

Discussion

When it is technically and politically possible, experimental oiling in the field would provide the most direct, incontestable evidence of oil impacts. Such experiments were not seriously considered in the case of the *Exxon Valdez* oil spill, but from a scientific perspective this would be the most satisfying method for providing evidence of damage.

Thus far, the discussion in this paper has concentrated on changes in population abundance. I have not considered ecosystem-level effects, that is, changes in the structure of the ecosystem. Indeed, most of the damage assessments were structured around single species; the intertidal and subtidal community studies are the major exception to this. In part this bias arose from the early legal demands for damage claims: It was desirable to assess damages for individual species where changes in abundance could be demonstrated. Resource managers have been managing species, and prespill information was on individual species, not ecosystems. The legal precedents for claiming ecosystem-level damages were not felt to be as strong, and the ecologists were less certain they knew how to measure such damage.

There is a danger, however, in ignoring ecosystem-level changes. It is not clear how to measure or value changes in the ecosystems, but it should be a subject for consideration. For instance, it would be possible to select oiled and non-oiled bays for an ecosystem-wide comparison. Such studies are fraught with difficulties, however, particularly because most of the species of public interest are quite mobile, and more individuals of these species will move in and out of any spatial area than can be readily studied.

To make a good case that oil caused changes in population abundance, no single source of evidence will be absolutely convincing. The strength of the case will depend upon the number of links that can be connected in a chain of evidence. Any single source of data, whether body counts, before and after comparisons, postspill oiled versus non-oiled comparisons, or experimentally induced oil effects will not provide a convincing case on its own. Ideally, data would show that oil reached individuals, that oil killed individuals, and that abundances in oiled sites were lower than in either non-oiled sites or in the same sites before the spill. Design of future damage assessments should be built upon consideration of these links in a chain of evidence.

References

Berger, J. O. 1985. Statistical decision theory and Bayesian analysis. Springer-Verlag, New York.

Bowman, T. D., P. F. Schempf, and J. A. Bernatowicz. 1995. Bald eagle survival and population dynamics in Alaska after the *Exxon Valdez* oil spill. Journal of Wildlife Management 59:317–324.

Bue, B. G., S. Sharr, S. D. Moffitt, and A. K. Craig. 1996. Effects of the *Exxon Valdez* oil spill on pink salmon embryos and preemergent fry. American Fisheries Society Symposium 18:619–627.

Carls, M. G., and six coauthors. 1996. Growth, feeding, and survival of pink salmon fry exposed to food contaminated with crude oil. American Fisheries Society Symposium 18:608–618.

Collier, T. K., C. A. Krone, M. M. Krahn, J. E. Stein, S.-L. Chan, and U. Varanasi. 1996. Petroleum exposure and associated biochemical effects in subtidal fish after the *Exxon Valdez* oil spill. American Fisheries Society Symposium 18:671–683.

DeGange, A. R., A. M. Doroff, and D. H. Monson. 1994. Experimental recovery of sea otter carcasses at Kodiak Island, Alaska, following the *Exxon Valdez* oil spill. Marine Mammal Science 10:492–496.

Edwards, A. W. F. 1972. Likelihood. Cambridge University Press, Cambridge.

Garrott, R. A., L. L. Eberhardt, and D. M. Burn. 1993. Mortality of sea otters in Prince William sound following the *Exxon Valdez* oil spill. Marine Mammal Science 9:343–359.

Geiger, H. J., B. G. Bue, S. Sharr, A. C. Wertheimer, and T. M. Willette. 1996. A life history approach to estimating damage to Prince William Sound pink salmon caused by the *Exxon Valdez* oil spill. American Fisheries Society Symposium 18:487–498.

Hepler, K. R., P. A. Hansen, and D. R. Bernard. 1996. Impact of oil spilled from the *Exxon Valdez* on survival and growth of Dolly Varden and cutthroat trout in Prince William Sound. American Fisheries Society Symposium 18:645–658.

Highsmith, R. C., and six coauthors. 1996. Impact of the *Exxon Valdez* oil spill on intertidal biota. American Fisheries Society Symposium 18:212–237.

Holling, C. S. 1973. Resilience and stability of ecological systems. Annual Review of Ecology and Systematics 4:1–23.

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

Jewett, S. C., T. A. Dean, and D. R. Laur. 1996. Effects of the *Exxon Valdez* oil spill on benthic invertebrates in an oxygen-deficient embayment in Prince William Sound, Alaska. American Fisheries Society Symposium 18:440–447.

Oakley, K. L., and K. J. Kuletz. 1996. Population, reproduction, and foraging of pigeon guillemots at Naked Island, Alaska, before and after the *Exxon Valdez* oil spill. American Fisheries Society Symposium 18: 759–769.