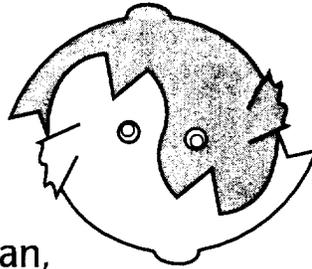


SH
379
.F56
J68
1998

FISHERY STOCK ASSESSMENT MODELS

Edited by F. Funk,
T.J. Quinn II, J. Heifetz,
J.N. Ianelli, J.E. Powers,
J.F. Schweigert, P.J. Sullivan,
and C.-I. Zhang



Proceedings of the International Symposium on
Fishery Stock Assessment Models for the 21st Century,
October 8-11, 1997, Anchorage, Alaska

Lowell Wakefield Fisheries Symposium

University of Alaska Sea Grant College Program
AK-SG-98-01 1998 Price \$40.00

Integrating Ecosystem Studies: A Bayesian Comparison of Hypotheses

Milo D. Adkison, Brenda Ballachey, James Bodkin, and Leslie Holland-Bartels

U.S. Geological Survey, Anchorage, Alaska

Abstract

Ecosystem studies are difficult to interpret because of the complexity and number of pathways that may affect a phenomenon of interest. It is not possible to study all aspects of a problem; thus subjective judgment is required to weigh what has been observed in the context of components that were not studied but may have been important. This subjective judgment is usually a poorly documented and ad hoc addendum to a statistical analysis of the data. We present a Bayesian methodology for documenting, quantifying, and incorporating these necessary subjective elements into an ecosystem study. The end product of this methodology is the probability of each of the competing hypotheses. As an example, this method is applied to an ecosystem study designed to discriminate among competing hypotheses for a low abundance of sea otters at a previously oiled site in Prince William Sound, Alaska.

Introduction

Ecosystem approaches are increasingly advocated as a way of improving the science and management of natural systems (Lackey 1998). For instance, studies of the effects of anthropogenic stressors on a species can be misleading if they ignore possible indirect effects acting through predator or prey populations (Higashi and Patten 1989). Further, natural changes in these other components of the ecosystem may cause changes in the focal population, masking or exaggerating the effects of the stressor (Piatt and Anderson 1996). Many studies of the impacts of human actions on a particular species now include research on other components of the ecosystem thought to be important to the focal species.

Nonetheless, there are practical limitations to an ecosystem approach. Because of cost and logistical constraints, not all ecosystem components can be studied and therefore some indirect impacts may be missed. Experimentation or replication may not be possible, and it may thus be difficult to unambiguously assign causes to any observed differences in populations between impacted and non-impacted sites, or before versus after an impact at a single site. It is also highly likely that among the suite of studies, some will give results that are to some degree contradictory.

For these reasons, interpreting the results of an ecosystem study requires some degree of expert judgment. Synthesizing the results of numerous studies of parts of a complex problem is difficult, and it may thus be difficult for investigators to reach conclusions in a rational fashion. Further, different scientists faced with the same evidence may arrive at different conclusions. As the subjective interpretation of results tends to be an ad hoc and poorly documented process, the sources of disagreement may be difficult to uncover and resolve. This paper presents a structured method for documenting and quantifying the expert interpretation of the results of an ecosystem study.

Proposed Methodology

The methodology presented here is designed for testing ecosystem-level hypotheses. It integrates studies of diverse components of the ecosystem, summarizing the results as the relative evidence for each hypothesis from each study and the overall evidence for each hypothesis from the ensemble of studies. Its Bayesian features consist of incorporating and quantifying the subjective step of interpreting results, and calculating a probability that each hypothesis is true.

The method consists of the following steps:

1. Generate hypotheses
2. Summarize the experiments and their results
3. Create a table of the expected results under each hypothesis if each experiment were ideal
4. Calculate the probability of the observed result under each hypothesis using statistical considerations
5. Adjust probabilities by considering potential violations of statistical assumptions
6. Adjust probabilities to account for differences between the hypotheses tested and the hypotheses of interest
7. Summarize the evidence for each hypothesis, accounting for dependencies among experiments

Table 1. Hypothetical results of a set of ideal experiments.

	Hyp. 1	Hyp. 2	Hyp. 3	Hyp. 4
Study A	Positive	Negative	Negative	Positive
Study B	Negative	Negative	Positive	Negative
Study C	Positive	Positive	Positive	Negative

Steps 3-6 deal with eliciting statements of probability from experts. Such elicitations can be problematic if experts are unfamiliar with translating their experiments into numerical probabilities (Morgan and Henrion 1990, Ch. 7). Our sequence of steps is designed to overcome such problems by sequentially considering several sources of uncertainty, progressing from the most to least familiar. At each of the seven steps, in particular those where subjective judgment is required, the rationale leading to the decision should be thoroughly documented.

Step 1. Generate Hypotheses. The first step is to have the experts identify the hypotheses that are the competing explanations for the phenomenon under investigation. It is important that the hypotheses be both exhaustive and mutually exclusive. If not, the confidence assigned to some hypotheses will be overstated, as the evidence for them will in some respects be counted twice.

Often, there will be reason to believe that several of the hypothesized phenomena might act simultaneously. There are two principal ways of constructing mutually exclusive hypotheses if this is a possibility. The first is to consider a "multiple causes" hypothesis. The second is to redefine the hypotheses to allow minor effects of other factors. For instance, the two hypotheses "effect is produced by factor A" and "effect is produced by factor B" can be made mutually exclusive by redefinition as "effect is principally produced by factor A" and "effect is principally produced by factor B."

Step 2. Summarize the Available Data. In this step, the studies and their results are summarized. For clarity, it is often more useful to use a short verbal description of the results. For instance, a study of differences in prey abundance between control and treatment might be summarized as "much greater abundance found at the control site."

Step 3. Consider Ideal Studies. The third step in this process is to lay out a table with the different hypotheses as the top row and the different experiments as the left-most column (Table 1). Then, have the experts fill out this table as if each study were an ideal experiment; i.e., there was no possibility of either false positive or false negative results.

Table 2. Table of likelihoods.

	Hypothesis 1	Hypothesis 2
Study A	$P(\text{Result of A} \text{Hyp. 1})$	$P(\text{Result of A} \text{Hyp. 2})$
Study B	$P(\text{Result of B} \text{Hyp. 1})$	$P(\text{Result of B} \text{Hyp. 2})$

$P(\text{Result of A}|\text{Hyp. 1})$ means the probability of getting the observed result of Study A if Hypothesis 1 were true.

In the hypothetical example in Table 1, Study A would distinguish between Hypotheses 1 or 4 and Hypotheses 2 or 3. In combination, the three studies would be able to determine which hypothesis was true.

Step 4. Statistical Considerations. While ideally the three studies would determine which hypothesis was true with 100% accuracy, in the real world misleading results may be obtained. One of the ways this may happen is through random sampling error. Often, almost any result is possible under any of the hypotheses. Nonetheless, the observed result will be more probable under some hypotheses than others.

The objective of this step is to calculate these relative probabilities, otherwise known as the *likelihoods* of each of the hypotheses (Gelman et al. 1995, Ch. 1). Often, with continuously distributed variables, the likelihood is a probability density rather than a probability per se. Likelihoods (Table 2) are usually obtained from standard statistical distributions such as the normal or binomial. The exact distribution used depends upon the assumptions made about the experimental data, such as whether each point is independent and identically distributed, whether the sampling variance is constant, etc.

Table 2 shows the first of a series of steps in which experts are asked to assign probabilities to the competing hypotheses. Some experts are unfamiliar with quantitative probability statements and scientists in particular are often uncomfortable making assertions about the relative merits of competing hypotheses without conclusive evidence. This step is important in that it introduces experts to assigning probabilities to the hypotheses, yet does so in a rigorous way using familiar statistical calculations.

Step 5. Account for Possible Biases in the Test or Experimental Results. The assumptions of statistical tests are rarely exactly met. Samples may not be completely independent, important sources of error may not be included in the statistical model (e.g., ignoring error in the measurement of the independent variable), and measurements may have some unknown biases. Historically, statistical confidence tends to overstate the certainty of scientific results (Henrion and Fischhoff 1986).

In constructing the table of likelihoods of results, this overconfidence needs to be accounted for. Generally, the effect of such errors is to make the probabilities of the result under each hypothesis more similar. Based on their knowledge of the experiment, experts should determine which assumptions of the test are likely to be violated, and to what degree. These judgments are to some extent subjective, but once made the statistical literature or computer simulations can provide guidance on their likely effects. In consultation with a statistician, the experts should adjust the table of probabilities to account for such violations.

Step 6. Account for Differences Between the Statistical Hypothesis Being Tested and the Biological Hypothesis That Is Actually of Interest. Often, an experiment to test a hypothesis tests it only indirectly. The results may thus be ambiguous if the indirect indicator could occur in several ways, some of which are not related to the hypothesis.

For example, if the hypothesis were that some population was affected by an environmental contaminant, an investigator might test the environment for the presence of the contaminant and test individuals for signs of poor health. A positive result in either case would not necessarily implicate the contaminant; the contaminant might be present yet not be causing health effects, or poor health might be due to causes other than the contaminant.

As in step 5, the effect of a difference between the hypothesis tested and the hypothesis of interest is to even further equalize the probabilities of the observed results under each hypothesis. The appropriate amount of adjustment of the table entries depends on the probability of other (possibly unknown) alternative explanations for the test results.

Such assessments are unavoidably subjective and require the judgment of experts. Hopefully, by this point in the process the experts are comfortable with assessing the relative probability of the data under each hypothesis and how violations of assumptions may result in misleading experimental results. It is crucial that they consider alternative explanations for their data yet not be paralyzed by such possibilities. They should be willing to examine data that seems to strongly favor one hypothesis and consider whether there are other, possibly unstudied ecosystem pathways that could produce similar results *and* state how probable they feel such pathways are.

Step 7. Summarize the Evidence. In this step, the table of probabilities is summarized to derive the overall weight of evidence for each hypothesis provided by the ensemble of studies. If the studies are independent, then elementary statistical theory says the joint likelihood of each hypothesis is simply the multiplication of its probability under each study (equation 1). The overall likelihood of each hypothesis is then simply the product of its column of probabilities (here R_1 , R_2 , and R_3 signify the results of experiments 1, 2, and 3, respectively).

$$\text{Likelihood of hypothesis} = P(R1|\text{hyp.}) \times P(R2|\text{hyp.}) \times P(R3|\text{hyp.}) \quad (1)$$

The different hypotheses can then be compared in terms of their relative likelihoods. This comparison is easier if the likelihoods are re-scaled so that the sum of all of the likelihoods is 1. From a Bayesian perspective, each re-scaled likelihood could then be interpreted as the probability that a hypothesis was true.

Complication A. Dependencies among Results. There are two ways that experimental results might not be independent. First, the data from two experiments may have been taken from the same random sample. Second, two experiments may measure the same ecological phenomenon two different ways. In either case, it is not appropriate to treat the results as providing independent evidence bearing on the alternative hypotheses; i.e., simply multiplying the probabilities of the two experiments together will overweight the evidence.

There are several possible methods to account for dependencies among experimental results. If experiments are highly interdependent, they should be lumped and a single probability of each hypothesis calculated for the ensemble results. If experiments are only partially dependent, the correlation of results must be accounted for. If the correlation can be calculated, probability theory provides methods for calculating a joint probability. If not, a value must be obtained from experts, although experts have been found to perform poorly at providing a numerical value for correlation coefficients (Morgan and Henrion 1990, Ch. 7).

A more intuitive method for dealing with partially correlated results is to ask investigators to provide an estimate of the "effective" number of experiments. For instance, investigators may feel that dependence between two experiments is such that they jointly provide only as much evidence as 1.5 independent experiments. Then, the appropriate adjustment would be to raise each of the probabilities to the 0.75 power (e.g., equation 2). In general, if N experiments are correlated so that the effective number is E , probabilities for hypotheses for each experiment should be adjusted by raising them to the E/N power.

$$\text{Likelihood of hypothesis} = P(R1|\text{hyp.})^{0.75} \times P(R2|\text{hyp.})^{0.75} \quad (2)$$

Complication B. Prior Probabilities. Bayesian statistics involves multiplying the likelihoods by a set of prior weights (the prior probabilities) for the hypotheses before re-scaling to calculate the posterior probabilities. In the Bayesian approach, these prior probabilities reflect the weight accorded each hypothesis *before* the experiments were conducted. Assuming the probability of each hypothesis to be proportional to the joint likelihoods treats each hypothesis as being equally likely a priori, thus letting the data determine the relative probability of each hypothesis. While this is intuitively appealing, it may not be appropriate.

For instance, if the analysis were being used in a legal proceeding, it might be appropriate to give the benefit of the doubt to the defendant by assigning small prior weights to hypotheses implicating the defendant. Similarly, in investigating current scientific theory a high prior weight might be assigned to the currently accepted paradigm, so that a novel competing theory would not get much credence unless the evidence for it was overwhelming. An alternative to using prior weights is to calculate probabilities only from likelihoods, but require a very high probability that a hypothesis is true before acting on it. Whatever the prior weights, if data strongly support one hypothesis over the others the final probabilities will reflect this.

Standard Bayesian practice is to compare the evidence for competing hypotheses using Bayes factors (Kass and Raftery 1995). The Bayes factor is simply the ratio of the posterior probabilities of two competing hypotheses divided by the ratio of the prior probabilities assigned before the experiments were conducted. When the prior probabilities of the hypotheses are equal, this is simply the ratio of the posterior probabilities.

An Example: Sea Otters after the Exxon Valdez Oil Spill

On March 4, 1989, the supertanker *Exxon Valdez* spilled nearly 42 million liters of crude oil in Prince William Sound, Alaska (Spies et al. 1996). This spill is hereafter referred to with the acronym EVOS. Sea otter populations in oiled areas suffered high mortality (Loughlin et al. 1996). Other components of the ecosystem were likewise severely affected. Five years after the spill, residual oil was present in sediments and mussel beds in some areas of the spill (Spies et al. 1996). Even today, residual oil is found in some areas.

The Nearshore Vertebrate Predator (NVP) project (Holland-Bartels et al. 1996), a multi-university and agency investigation funded by the EVOS Trustee Council, is aimed at determining whether top predators in Prince William Sound are still suffering the effects of the oil spill. The question is difficult to answer unambiguously because of the complicated nature of the ecosystem and the lack of data from the period before EVOS. The NVP project studies predator populations from several points of view, and also looks at other components of the ecosystem on which these predators depend. If a population is still being affected by EVOS, the study is designed to ascertain whether the effects are due to the continuing toxic effects of oil, a slow rate of recovery from past mortality, or an indirect effect on some critical ecosystem component.

With limited resources and such an intensive approach, few populations can be studied. Sea otter abundance at Knight Island, which was oiled in 1989, is lower than at Montague Island, which was not. The NVP

sea otter study has focused on these two populations, trying to find the reason for these differences in abundance. The principal hypotheses are:

1. **Direct toxicity of residual oil.** Residual oil is present and reducing the fecundity and/or survival of otters at the oiled site.
2. **Reduced forage due to oil effects.** The initial impact of oil or residual oil is reducing prey available to sea otters.
3. **Slow recovery due to demographic limitations.** Aside from the initial otter mortality from EVOS, residual oil is absent or does not affect otters or their food. However, limitations on the maximum growth rate of the population have prevented the population from reaching capacity yet.
4. **Natural differences in capacity.** The oiled site has poorer or less abundant otter habitat.

A variety of studies have been undertaken to determine which hypothesis is the most likely. These include:

1. **Demographic comparisons.** Population abundance, age structure, and reproductive rates were compared between islands.
2. **Individual health.** Otters were captured at both locations. Individuals were weighed and measured, and blood samples taken. In particular, blood cells and serum chemistry were examined for signals of poor health, and a specific signal of exposure to oil (the enzyme P450) was tested for.
3. **Prey abundance and foraging success.** The abundance and size distribution of major prey items of sea otters were compared among islands. In addition, foraging sea otters were observed to determine relative rates of success in obtaining prey items.

Statistical hypothesis tests were performed for many of the studies but are not reported here. We chose not to calculate likelihoods based solely on statistical distributions—step 4 of our methodology—because the limitations imposed by the design of the study tended to emphasize the considerations dealt with in steps 5 and 6. There are multiple predictions from each of the hypotheses, not all of which are distinct. Any particular study result may eliminate some hypotheses but leave several others. More likely, any particular study result would be ambiguous, as there is a small likelihood of almost any result from each hypothesis. In particular, the detection of a phenomenon does not necessarily imply that this was the cause of the difference in abundance between the two islands. For instance, oil could be present but yet not greatly affect survival. Likewise, prey abundance could differ between one site and another but be unrelated to the difference in otter abundance.

Table 3. First attempt at integrating studies.

Experiment and (result)	"A" Demographic limitation	"B" Food limitation	"C" Oil persistence	"D" Recovery has occurred
Otter density (K << M)	0.9	0.9	0.9	0.3
Repro. rates (equal)	0.9	0.5	0.7	0.9
Blood chemistry (equal)	0.9	0.7	0.3	0.9
P450 (equal)	0.7	0.7	0.1	0.9
Prey abundance (M < K)	0.9	0.1	0.1	0.1
Foraging success (M < K)	0.9	0.1	0.7	0.1
Joint likelihood	0.4133	0.0022	0.0013	0.0022
Probability of hypotheses	98.6%	0.53%	0.32%	0.52%

Top row gives hypotheses, and left column gives experiments with the results in parentheses. "M" refers to Montague Island (control), and "K" to Knight Island (oiled). The main body of the table gives the probability of obtaining each experimental result under each hypothesis. The bottom two rows summarize the result as the product of the probabilities for each hypothesis (i.e. the joint likelihood) and the probability products re-scaled to sum to 100%.

Thus, the interpretation of the results of the studies required some judgment. Our chief tool was to ask ourselves, "What is the probability we would get the result we observed from Study ___ if Hypothesis ___ was true?" We attempted to quantify our impression of the strength of each piece of evidence by filling out the table of probabilities, sequentially considering what the result would mean in an ideal world, what the statistical tests implied, how the assumptions of the tests might be violated, and what mechanisms might cause the results to be misleading.

We felt our ability to discriminate among probability levels was fairly coarse. Accordingly, we initially filled in the table of probabilities verbally, using the categories "high," "moderate-high," "moderate," "low-moderate," and "low," which we later replaced with 0.9, 0.7, 0.5, 0.3, and 0.1, respectively (Table 3).

The result of our first analysis was to assign more than a 98% probability to the hypothesis that the population differences were due to a demographic limitation in the rate of recovery of the Knight Island population from spill mortality. All other hypotheses combined had less than a 1.5%

probability of being true. We were unhappy with this result, as this high degree of confidence did not reflect our personal higher degree of uncertainty. We felt that the evidence for this hypothesis was not that strong.

In examining the reasons for this initial result, we identified three principal sources of error. First, we overstated the power of the studies to discriminate among hypotheses. For instance, we assigned a 0.90 probability of seeing greater prey abundance at the oiled site if demography was limiting recovery, but only a probability of 0.10 under any of the other hypotheses. We did not adequately address step 6 of our methodology; for instance, there would be a fairly good chance of seeing higher prey abundance at the oiled site under several alternative hypotheses.

Second, the range of hypotheses we considered was too narrow. In retrospect, we felt there was a strong possibility that all of the hypotheses might be incorrect, and some other factor might be responsible for differences between areas. This resulted in an unrealistically high probability for the hypothesis most consistent with the data.

Third, we did not adequately account for dependencies among experimental results (step 7, complication A). While we lumped most blood chemistry measures into one result, we kept the assay for the enzyme P450 (a more direct measure of exposure to oil) as a separate experiment. Since this assay could indicate the same phenomenon, and was measured on the same sample of animals, we felt the two results were effectively equivalent to only 1.5 experiments. Similarly, measures of prey size, prey abundance, and foraging success to some extent measured the same phenomenon. In retrospect, we decided to consider them as equivalent to two experiments.

We therefore revised the tabled probabilities, taking what we hoped was a more realistic look at the power of the studies and adding another alternative hypothesis to those we had listed. While we were able to think of several specific alternatives, we felt the true explanation for population differences might be something we hadn't considered. Therefore, we added only one hypothesis; an "unknown causes" category. Meanwhile, the completion of analyses of blood chemistry and the enzyme P450 suggested that residual oil might be present at the oiled site, and new information became available about the size distribution of prey species (Table 4).

The revised table again supports the hypothesis that the populations differ because the population in the oiled area has not had the time to recover fully from the losses due to the oil spill. However, it shows even greater support for the hypothesis that residual oil is still affecting the population. The hypothesis that some unknown factor accounts for the difference between populations is also quite probable.

Two hypotheses were eliminated from consideration, principally because of the forage abundance studies. Forage was more abundant and foraging success higher at the oiled site. These results were not at all

Table 4. Second attempt at integrating studies.

Experiment and (result)	"A" Demogr. limit	"B" Food limit	"C" Oil persist	"D" Recovered	"E" Unknown causes
Otter density (K << M)	0.9	0.9	0.9	0.3	0.9
Repro rates (equal)	0.9	0.5	0.7	0.9	0.9
Blood CBCs & chemistry (weak indication of liver damage at K)	0.5	0.5	0.7	0.3	0.5
P450 (M < K)	0.3	0.3	0.9	0.3	0.3
Prey abundance (M < K)	0.9	0.1	0.5	0.3	0.5
Prey size (M < K)	0.9	0.1	0.7	0.3	0.7
Foraging success (M < K)	0.9	0.1	0.7	0.3	0.7
Joint likelihood	0.1581	0.0011	0.1744	0.0040	0.0764
Probability of hypotheses	38.2%	0.3%	42.1%	1.0%	18.5%

Top row gives hypotheses, and left column gives experiments with the results in parentheses. "M" refers to Montague Island (control), and "K" to Knight Island (oiled). The main body of the table gives the probability of obtaining each experimental result under each hypothesis. The bottom two rows summarize the result as the product of the probabilities for each hypothesis (i.e. the joint likelihood) and the probability products re-scaled to sum to 100%.

consistent with the food limitation hypothesis, and were also unlikely if the population at the oiled site had recovered to its carrying capacity. However, it should be noted that the "unknown causes" hypothesis, which has a fairly high probability of being true, is not necessarily related to the spill. Thus it would be inappropriate to say the probability that the population is no longer suffering effects of the spill is only 0.01.

We will refine and expand this analysis as more data become available and more experts are consulted. These results are not our final interpretation, and should be viewed as a preliminary analysis. We provided this example solely to illustrate the use of the methodology.

Discussion

The Bayesian aspects of the proposed methodology are (1) use of subjective expert judgment in interpreting indirect tests of hypotheses, and (2) integration of experimental results and expert judgment into an overall probability for each hypothesis using Bayesian probability calculations. A large literature exists on using Bayesian methods to compare hypotheses (Kass and Raftery 1995).

Bayesian methods have been criticized from a variety of standpoints (e.g., Dennis 1996). The principal criticism is that Bayesian methods inject subjectivity into scientific analyses that should be objective. However, in extrapolating from the results of diverse studies on small aspects of a larger question, subjectivity in the form of expert judgment is unavoidable. We propose a methodology that formalizes the intuitive process experts use in interpreting the results of ecosystem studies. This approach clearly distinguishes subjective interpretation from experimental results, and clearly shows the reasoning used.

Our methodology provides a tool for investigators to organize their thinking. The ecosystem and the results of the numerous studies may be too complex to be readily grasped in their entirety. By allowing investigators to approach the synthesis of the studies one element at a time, our method increases the tractability of the process.

The methodology also facilitates openness and discussion, since subjective components of the synthesis of the studies are documented and quantified. It clearly shows why a particular conclusion was reached, and what evidence investigators felt was ambiguous or particularly strong. Areas of disagreement among investigators are also easily identified.

Our methodology is based on principles derived from other methods widely used for eliciting probabilities from experts (summarized in Morgan and Henrion 1990, Ch. 7). Examples of such methods include the Stanford/SRI protocol (Spetzler and Stael von Holstein 1975, Merkhofer 1987) and the Wallsten/EPA protocol (Wallsten and Whitfield 1986). We've tailored our methodology to the specific goal of summarizing the relative support for alternative hypotheses from an interrelated but necessarily incomplete set of studies.

Most methods for probability elicitation pay great attention to getting experts comfortable with the idea of translating their knowledge and judgment into probability statements, and to overcoming a tendency of experts to give probabilities that overstate the level of certainty (Tversky and Kahneman 1982; Morgan and Henrion 1990, Ch. 7). Our solution to these difficulties is to take experts through a specific sequence of probability elicitation steps. These start with specifying deterministic outcomes, then progress through familiar specifications of probability (likelihood calculations) to less familiar probability specifications (the effects of violation of statistical assumptions and of not directly testing the hypothesis of interest). This sequence gradually introduces the process of making

probability statements. It also sequentially introduces more and more forms of uncertainty, continually forcing the expert to reflect on whether the degree of confidence he's previously expressed is appropriate.

Our example illustrates both the utility and limitations of the methodology. The summary table lists the hypotheses and the experimental results. Probabilities within the table explicitly document the experts' interpretation of the consistency of the results of each experiment with each hypothesis. The summary probabilities excluded two hypotheses but retained three others, one of which appears to be only half as probable as the other two.

However, the 18.5% probability assigned to the "Unknown Causes" hypothesis makes interpretation of the other probabilities somewhat ambiguous. Much of the probability assigned to this hypothesis may indicate that recovery has occurred, and the differences we found are caused by some unknown factor(s) unrelated to the spill. It is also possible that "unknown causes" represents effects related to the spill such as cascading ecological effects. In either case, the results do provide guidance for further research; they suggest that continuing studies should focus on hypotheses "A," "C," and "E."

The necessity for re-evaluating our initial analysis because of unrealistic results is instructive. It reinforces the experience of others who have found that numerical statements of probability given by experts tend to be overly confident (Tversky and Kahneman 1982, Henrion and Fischhoff 1986). Our second try produced a result that we felt better reflected the strength of the evidence provided by the experiments.

There is a danger that allowing such reanalysis could result in investigators juggling numbers to arrive at a result that reflected their preconceptions. However, an honest reappraisal of each element in the table is not inappropriate. Most methods for probability elicitation do recommend that assessors return to an earlier phase in the process whenever questioning reveals that the probabilities elicited clearly don't reflect the expert's judgment (Kadane et al. 1980; Morgan and Henrion 1990, Ch. 7; Laskey 1995). We found the reanalysis of the table caused us to re-examine the basis of our interpretations; rather than reinforcing our preconceptions, it tended to make us change them.

Use of our methodology will make it easier to examine the source of differences in interpretation of a study. For example, a scientist who disagreed with our conclusions might find that the basis of his difference was the weight placed on the blood chemistry results. A sensitivity analysis to alternative interpretations would be easy to perform by replacing the disputed probability with an alternative value to see if this affected the conclusions.

This method is not proposed as a substitute for good experimentation. With scarce, poor quality, and ambiguous data the conclusion reached after applying this method will be that considerable uncertainty remains. However, in such situations this methodology may identify areas of major

uncertainty and suggest fruitful lines of investigation. The major benefit of this approach is the explicit documentation and quantification of the unavoidable subjective interpretation of ambiguous results that arise in many ecosystem investigations. In contrast, when strong experimental designs are available that produce clear evidence, subjective interpretation will be minimized and investigators should reach consensus.

Acknowledgments

The authors wish to thank Tom Dean, Jennifer DeGroot, George Esslinger, Steve Jewett, Dan Monson, Chuck O'Clair, Alan Rebar, Paul Snyder, and Glenn VanBlaricom for their contributions. The EVOS Trustee Council provided financial support for this study.

References

- Dennis, B. 1996. Discussion: Should ecologists become Bayesians? *Ecol. Appl.* 6:1095-103.
- Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 1995. *Bayesian data analysis*. Chapman and Hall, New York.
- Henrion, M., and B. Fischhoff 1986. Assessing uncertainty in physical constants. *Amer. J. Physics* 54:791-798.
- Higashi, M., and B.C. Patten. 1989. Dominance of indirect causality in ecosystems. *Am. Nat.* 133:288-302.
- Holland-Bartels, L., B. Ballachey, J. Bodkin, T. Bowyer, T. Dean, L. Duffy, D. Esler, S. Jewett, L. McDonald, C. O'Clair, A. Rebar, D. Roby, P. Snyder, and G. VanBlaricom. 1996. Mechanisms of impact and potential recovery of nearshore vertebrate predators. *Exxon Valdez Oil Spill Restoration Project Annual Report*. (Restoration Project 95025), National Biological Service, Anchorage, AK.
- Kadane, J.B., J.M. Dickey, R.L. Winkler, W.S. Smith, and S.C. Peters. 1980. Interactive elicitation of opinion for a normal linear model. *J. Amer. Stat. Assoc.* 75:845-854.
- Kass, R.E., and A.E. Raftery. 1995. Bayes factors. *J. Amer. Stat. Assoc.* 90:773-795.
- Lackey R.T. 1998. Seven pillars of ecosystem management. *Landscape and Urban Planning* 40:21-30.
- Laskey, K.B. 1995. Sensitivity analysis for probability assessments in Bayesian networks. *IEEE Trans. on Systems, Man, and Cybernetics* 25:901-909.
- Loughlin, T.R., B.E. Ballachey, and B.A. Wright. 1996. Overview of studies to determine injury caused by the *Exxon Valdez* oil spill to marine mammals. In: S.D. Rice, R.B. Spies, D.A. Wolfe, and B.A. Wright (eds.), *Proceedings of the Exxon Valdez oil spill symposium*. *Amer. Fish. Soc. Symp.* 18:798-808.
- Merkhofer, M.W. 1987. Quantifying judgmental uncertainty: Methodology, experiences, and insights. *IEEE Trans. on Systems, Man, and Cybernetics* 17:741-52.

- Morgan, M.G., and M. Henrion. 1990. *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge University Press, New York.
- Piatt, J.F., and P. Anderson. 1996. Response of common murrelets to the *Exxon Valdez* oil spill and long-term changes in the Gulf of Alaska marine ecosystem. In: S.D. Rice, R.B. Spies, D.A. Wolfe, and B.A. Wright (eds.), *Proceedings of the Exxon Valdez oil spill symposium*. Amer. Fish. Soc. Symp. 18:720-737.
- Spetzler, C.S., and C.-A.S. Stael von Holstein. 1975. Probability encoding in decision analysis. *Management Sci.* 22:340-352.
- Spies, R.B., S.D. Rice, D.A. Wolfe, and B.A. Wright. 1996. The effects of the *Exxon Valdez* oil spill on the Alaskan coastal environment. In: S.D. Rice, R.B. Spies, D.A. Wolfe, and B.A. Wright (eds.), *Proceedings of the Exxon Valdez oil spill symposium*. Amer. Fish. Soc. Symp. 18:1-16.
- Tversky, A., and D. Kahneman. 1982. Judgment under uncertainty: Heuristics and biases. In: D. Kahneman, P. Slovic, and A. Tversky (eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press, Cambridge, U.K., pp. 3-20.
- Wallsten, T.S., and R.G. Whitfield. 1986. Assessing the risks to young children of three effects associated with elevated blood-lead levels. ANL/AA-32, Argonne National Laboratory, Argonne, IL.