A RAND NOTE

A Research Agenda for Assessment and Propagation of Model Uncertainty

David Draper, James S. Hodges, Edward E. Leamer, Carl N. Morris, Donald B. Rubin

November 1987

RAND

This publication was supported by The RAND Corporation as part of its program of public service.

Published by The RAND Corporation
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90406-2138

A RAND NOTE

N-2683-RC

A Research Agenda for Assessment and Propagation of Model Uncertainty

David Draper, James S. Hodges, Edward E. Leamer, Carl N. Morris, Donald B. Rubin

November 1987

RAND

PREFACE

This document is about assessing the uncertainty that arises in the modeling step of statistical analyses and propagating that uncertainty through to the final inferences drawn or decisions made. Using I. J. Good's convenient device of referring to a person making modeling judgments as You, the point is that

- Usual practice consists of selecting a single model after what amounts to a search in the space of all possible models, and then acting as if Your favorite model chosen in this way is "right" in forming estimates, predictions, and uncertainty assessments; and that
- This tends to understate Your actual uncertainty, with the result that Your actions, both inferentially in science and predictively in decisionmaking, are not sufficiently conservative.

The reaction of one's colleagues to this problem is interesting and a bit strange, given its evident importance: Everybody knows exactly what is meant when the problem is described, many people have thought about it, but few people seem to be doing much about it.

Recently a team consisting of David Draper and James Hodges in the Department of Economics and Statistics at RAND, Edward Leamer in the Department of Economics at UCLA, Carl Morris in the Center for Statistical Sciences at Texas, and Donald Rubin in the Department of Statistics at Harvard put together a grant proposal to the Decision, Risk, and Management Sciences Program at the National Science Foundation on this topic. Draper and Hodges, the co-Principal Investigators, wrote and revised the proposal based on detailed comments from Leamer, Morris, and Rubin. The project description section of the proposal forms the body of this document, which is being distributed to interested colleagues for comment and criticism. The proposal describes work in progress that we plan to summarize in future publications.

Three points on style and scope of this effort are worth bearing in mind while reading the material below.

- The intended audience for the proposal was not statisticians but decision scientists, mainly operations research analysts, so the tone is fairly expository and there are some references to topics like fuzzy set theory that are out of the statistical mainstream.
- The emphasis here is on prediction and its role in decisionmaking rather than inference in scientific contexts. Much of what may be said about the effects of model uncertainty on prediction is also relevant to inference, but the two activities differ in basic ways.
- The main methodology we advocate for assessment and propagation of model uncertainty is Bayesian. The idea is to put a prior distribution on the space of all possible models that is retrospectively reasonably well calibrated, in the sense that uncertainty assessments are produced that accord fairly well with past prediction errors, and then integrate model uncertainty out essentially like a nuisance parameter. However, we are not entirely dogmatic about this Bayesian position; we also discuss several frequentist alternatives, including bootstrapping the modeling process and mixture likelihoods. The issue is not statistical ideology but finding methods that work--methods that lead to decisions You will regret retrospectively less often than those made with current techniques. We welcome comments from interested readers on other major frequentist and Bayesian approaches to the problem and on the relationships among these various approaches.

We also welcome descriptions of problems and settings in which model uncertainty is paramount, anecdotes on how people attempt informally to include model uncertainty in the "give-or-takes" they attach to predictions and estimates, and citations to key references we may have forgotten. Discussion, comments, and so on should be sent to David Draper at RAND.

SUMMARY

Decision analyses rely on predictions of future outcomes; good analyses incorporate assessments of likely predictive accuracy. Most statistical methods permit assessments of uncertainty only about parameters of a chosen model; few allow uncertainty about models and scenarios--*structural uncertainty*--to be counted explicitly. In many problems, structural uncertainty dominates; if left uncounted, the resulting decisions may not hedge sufficiently against uncertainty. Improvement on current practice requires quantitative methods for assessing such uncertainty and propagating it through to overall measures of predictive accuracy. This Note describes a research agenda for assessment and propagation of structural uncertainty.

Bayesian statistical theory, which is routinely used for assessing and propagating other sources of uncertainty, extends naturally to account for structural uncertainty. Once prior probabilities are assigned to each of a collection of plausible models, posterior probabilities can be computed and used to mix the predictive distributions arising from each of the models, to form an aggregate measure of predictive uncertainty. The idea is not new, but new methods are needed to implement it. The approach advocated here requires new techniques for selecting the prior probability distribution across the models, for doing the necessary computations, and for presenting the contributions of within- and between-model uncertainty. This Note, which is based on the project description section of a grant proposal to the National Science Foundation, outlines how to supply these methods by developing retrospective calibration techniques, extending existing Bayesian computational methods to allow propagation of model uncertainty, and adapting presentation methods designed for summarizing simple Bayesian analyses.

Successful research of this type will provide new general-purpose tools for decisionmaking that will improve the assessment of how much hedging against uncertainty should be built into one's actions.

CONTENTS

A RESEARCH AGENDA FOR ASSESSMENT AND PROPAGATION OF MODEL UNCERTAINTY

INTRODUCTION

Most decision analysis, including risk assessment and management, relies on predictions about what would happen in the future under given sets of conditions. Inaccurate predictions are unavoidable--nobody has a crystal ball. However, it is *not* acceptable to be overly optimistic about how accurate these predictions are likely to be, because then the actions indicated by the decision analysis will not be sufficiently conservative.

Three sources of uncertainty generally need to be assessed in determining the likely accuracy of a prediction. If we define a *scenario* to be a set of assumptions about how the available predictor variables will behave in the future, and a *model* to be a set of assumptions about how the outcomes of interest and the predictors are linked, these three sources of uncertainty are:

- *Structural* uncertainty about the *scenario* and *model*;
- *Estimation* uncertainty about *parameters*, given the model and scenario; and
- *Prediction* uncertainty about *future outcomes* given the scenario, model, and parameters, because unexplained stochastic fluctuation is built into most models even if scenario, model, and parameters are known.

The attention paid to these three sources of uncertainty varies by discipline and type of decision problem. In most fields, standard practice is to concentrate on estimation uncertainty. Probabilistic risk assessments are largely about prediction uncertainty, but in many other applications this source of uncertainty is ignored or downplayed. Although it is frequently discussed, structural uncertainty about models and scenarios, which often dominates the other kinds of uncertainty, almost never is incorporated into the "give-or-take" attached to a

prediction. The result can be a substantial overstatement of predictive accuracy, leading to actions that in retrospect did not hedge enough against alternative ways the future could have turned out. This project will focus on structural uncertainty and how it can be assessed, represented, and combined with estimation and prediction uncertainty to produce better overall measures of predictive uncertainty for use in decisionmaking.

As an example of these issues, consider the problem of forecasting oil prices.

> Predictions of oil prices critically affect decisionmaking both in government and in the private sector. Such predictions have a substantial influence on policymakers in designing tax and energy policies, and on evaluations of and choices among investment strategies in industry and banking. Numerous forecasts of 1986 oil prices were made by reputable institutions and individuals in the late 1970s and early 1980s predicting oil prices of over $40 a barrel, more than three times the actual mid-1986 price of $11 a barrel. These large errors have had major consequences for the oil and banking industries. For example, the oil industry overinvested in almost all sectors, and suffered idle capacity, low profits, and reduced demand for their services. International banks, also anticipating rising oil prices, increased their long-term lending to developing oil countries, with Mexico's current acute debt predicament a direct consequence. A study by the Cambridge Energy Research Associates estimates that one half trillion dollars was invested in 1980-81 on the expectation that oil prices would continue to rise (Syme, 1987).

It is instructive to examine how forecasts of this type were made. For this purpose the efforts of the Energy Modeling Forum (EMF) at Stanford University are typical. In the early 1980s EMF's 43-person working group examined oil price predictions from 10 different econometric models for 12 scenarios: a "reference," or "most likely" scenario, and 11 others, including various combinations of reduced demand, disrupted supply, technological breakthrough, reduced price elasticity of demand, and low economic growth (EMF, 1982). EMF presented detailed forecasts for each of the 10 models, but only for the "reference" scenario. Point forecasts for 1986 with this scenario, for instance, varied across the 10 models from roughly $30 to $50 a barrel, with an unweighted average of about $43.

EMF did not attach explicit "give-or-takes" to their predictions, but people using these and similar forecasts to make decisions did so implicitly. In retrospect these implicit uncertainty assessments were far too small. There were also a few explicit uncertainty assessments published at about the same time; even these showed unjustified optimism. In 1981, for example, the Energy Information Administration (EIA) predicted a 1986 price of $36 a barrel (in constant 1981 dollars), with low and high limits of $25 and $42 a barrel (EIA, 1982).

Thus in retrospect decisionmakers basing their actions on analyses like EMF's failed to accurately assess how wrong their predictions were likely to be. The reason, again in retrospect, is clear: People acted as if something quite similar to EMF's "reference" scenario would come to pass, and it did not. Even though groups like EMF thought a number of models and scenarios, implying radically different futures, were plausible enough to be singled out for attention in their analysis, many consumers of their findings, and other similar analyses, did not incorporate this structural uncertainty correctly into the "give-or-takes" attached to their predictions. If they had, their actions would have reflected substantially more hedging against uncertainty, and the mistakes described in the quotation above would not have been so large. This observation may seem like easy hindsight, but it is not necessary to go beyond what was available to readers of the EMF report to improve one's decisionmaking: In this and other examples we have examined, scenarios different from the analysts' favorite, but quite similar to what actually happened, *were* entertained but were not permitted to influence the overall prediction and uncertainty assessment.

This phenomenon is not a fluke; in varying degrees, it describes almost every application of statistical or probabilistic methods--Bayesian or classical--in which a model is selected and used to make a prediction and assess its uncertainty. Standard practice is to examine a variety of potential models and scenarios and then make a single choice, forming estimates, predictions, and uncertainty assessments internal to the chosen model-scenario combination. When models and scenarios are used in this way, the analyst acts, in effect, as if the model and scenario were known to be true, when in fact model-scenario

selection is almost always attended by substantial uncertainty. The result is often a retrospective realization that the standard uncertainty assessment was far too small; in other words, in retrospect the prediction process is seen to be not well *calibrated*. This Note describes a research program to develop an approach to the propagation of model and scenario uncertainty that will be retrospectively better calibrated than standard methods that ignore such uncertainty, thereby permitting some confidence that prospective calibration will also be improved. The work to be undertaken builds on material laid out in section 2.1 of Hodges (1987) and in section 2 of Draper (1987).

GOALS AND OBJECTIVES

At present, most careful analysts consider structural uncertainty through *sensitivity analyses* (Dempster, 1985; Leamer, 1985; Skene, Shaw, and Lee, 1986; von Winterfeldt and Edwards, 1986). With this technique the investigator varies the model and scenario assumptions about which he is most unsure, noting how the predictions and "give-or-takes" change. If outcomes on the scales that inform action are fairly stable, the analyst ignores this structural uncertainty. But in the more usual case where the assumptions do matter, typical practice is to acknowledge this extra uncertainty qualitatively, by describing in words how someone believing one set of assumptions might discount the findings arising from another set. In purely scientific contexts, where no action need be taken other than the choice of the next experiment to be performed, this summary is often adequate. But for decisionmaking purposes this approach is deficient, because it does not suggest how to combine between-model/scenario and within-model/scenario uncertainty to obtain an overall quantitative measure of uncertainty that accurately reflects the analyst's ignorance about which of the possible futures will actually occur.

We advocate improving on standard practice through a Bayesian extension of sensitivity analysis, in which predictions and uncertainty assessments from some or all of the separate model-scenario combinations examined in the sensitivity analysis are combined quantitatively. As commonly used, Bayesian methods are like classical methods in that they are executed as if some single model were true, even when other models

are almost as plausible as the favorite. But Bayesian theory extends naturally to incorporate uncertainty about models: Once prior probabilities are placed on each of a collection of models, they can be updated to posterior probabilities, and this post-data model uncertainty can be propagated through to the overall measure of predictive uncertainty. It is not unusual for a Bayesian to expand a given model (Box, 1980) by adding parameters. The extension of Bayesian technique that we advocate is similar, except that instead of adding parameters to a model, it involves adding distinct models. Otherwise the theory is no different.

The practice, however, is different, for although this idea is in the minds of many, few have used it and few appropriate tools exist. We propose to develop tools that will permit analysts to execute this extension of Bayesian technique and to present their analyses with appropriate explicitness. Specifically, we propose to develop readily coded and computed methods with which analysts can

- improve their current practice of sensitivity analysis by more readily identifying *high-leverage* alternative models and scenarios--those plausible alternative modeling specifications that have the greatest influence on predictions, uncertainty assessments, and actions;
- assess prior probabilities for each of a collection of models and compute posterior probabilities, in much greater generality than is now possible;
- propagate model uncertainty through to an aggregate predictive measure of uncertainty (using approximations of various degrees of computer intensity as needed); and
- report an accounting of the portions of that aggregate assessment of uncertainty that arise from uncertainty *about* models as well as *conditional on* models.

SIGNIFICANCE OF THE WORK

Models are ubiquitous tools in policy analysis, in making and enforcing regulations, in business decisionmaking, in scientific research, and elsewhere. Such models are used to assess the uncertainty of conclusions drawn from the modeling process, assessments which in turn affect choices by individuals, government agencies, and businesses. Many researchers (surveyed, for instance, in Covello, von Winterfeldt, and Slovic, 1986) have examined how decisionmakers and citizens perceive and use the products of the modeling exercise. Work of this type takes modeling and its products as a given; our proposed work examines the process and results of modeling.

As the oil price example indicates, an overly optimistic judgment of the likely accuracy of a prediction can lead to actions that are not conservative enough, with dramatic practical consequences. Currently, no readily usable technology exists for incorporating all of the important sources of uncertainty into the overall "give-or-take" attached to a prediction. We propose to develop the most promising candidate.

RELATIONSHIP TO PREVIOUS WORK

Model Expansion

Box and Tiao (1962) introduced the Bayesian analysis of a *model expansion*. Instead of assuming that the data in their example were normally distributed with unknown mean and variance, as earlier writers had, they expanded that normal model by adding a third unknown parameter, for kurtosis. Their main interest was the mean, for which they derived a marginal posterior distribution by integrating out the variance and kurtosis in the joint posterior distribution of the three parameters. This marginal posterior distribution reflected greater uncertainty--was more spread out--than the posterior distribution that followed from the usual normal model, because it incorporated the uncertainty about the kurtosis of the distribution. The kurtosis in this case might be termed the *expansion parameter*, and this approach to capturing model uncertainty might be called the *parametric expansion* method.

Another type of Bayesian model expansion is to be found in *hierarchical Bayesian procedures* (Lindley and Smith, 1972). In simple Bayesian applications, such as the unexpanded Box and Tiao model above, prior distributions on the parameters in the model are themselves indexed by parameters. For instance, in standard location problems where the data are assumed to be normal with mean μ, the prior distribution for μ is often itself taken to be normal with some prior mean τ. In hierarchical Bayesian methods, an additional layer of uncertainty is added by introducing prior distributions, indexed by *hyperparameter(s)*, on the parameter(s) of the prior distribution. This is sometimes done to reflect the analyst's uncertainty in his prior specification (Berger and Berliner, 1986; Good, 1980), but is also a way to motivate many so-called *empirical Bayes* or *shrinkage* estimators (Morris, 1983), which have increased accuracy under certain conditions when compared with frequentist alternatives. (DuMouchel and Harris, 1983, provide an interesting example of a different use of hierarchical Bayesian methods, to capture interspecies and interexperimental model uncertainty in the context of human and animal dose-response models in cancer studies.) We examine hierarchical methods in more detail below.

Model expansion has also been adapted to non-Bayesian contexts, although in such cases it is typically used differently. In these contexts, the model is expanded by adding parameters, and then a significance test is usually performed to determine whether the new parameters should be retained (see McCullagh and Nelder, 1983, for example). (Cook, 1986, suggests an alternative method based on the change in the maximum likelihood estimate of the parameters of the unexpanded model.) After the decision to include or exclude the new parameters, an inference is made about the parameters of interest by any of a variety of standard techniques. These preliminary test methods are known to have deficiencies (Sclove, Morris, and Radhakrishnan, 1972) and can be substantially dominated by formal Bayesian methods (Leamer, 1978) or shrinkage methods.

Model-Mixing

The idea of attaching prior probabilities to each of a collection of models and updating them to form posterior probabilities has been a part of Bayesian theory for decades (e.g., Jeffreys, 1961), and continues to be applied in particular cases (e.g., Box and Hill, 1967; Geisel, 1974; and Zellner 1984). Once these posterior probabilities have been computed, it is a natural extension of Bayesian theory to suggest that they be used to compute posterior distributions for quantities that are common to all of the models, such as a parameter capturing a common effect, or a prediction on the original scale.

Many authors have suggested such "model-mixing"--see, for example, Berger (1984), Clayton, Geisser, and Jennings (1986), Harrison and Stevens (1976), Leamer (1978), Smith (1983), Stewart and Davis (1986), West (1986) and Zellner (1984)--and Box and Tiao's parametric expansion can be viewed as a special case. As we will discuss further below, hierarchical Bayesian procedures can also be thought of as model-mixing techniques in which uncertainty in the modeling process takes a particular form. But the emphasis in much of this previous work has been on estimation of unobservable quantities (parameters), rather than prediction of observables relevant to decisionmaking, with the result that the existing techniques of model-mixing are not necessarily well suited to prediction.

Practical applications of model-mixing have so far been limited almost exclusively to the multi-process Kalman filter (Harrison and Stevens, 1971) and its extension to non-normal observation processes (West 1986). In these cases, the emphasis has been on detecting shifts in an underlying mechanism from one process to another. For example, Smith and West (1983) used the multi-process Kalman filter to monitor kidney transplant patients, but they used the probabilities of the different models (the "processes") only to detect a patient's turn for the worse. Before Hodges (1987) and Draper (1987), little attention had been given in practice to model-mixing as a means of propagating uncertainty about the model through to a measure of predictive uncertainty. Correspondingly, little effort has been put into tools for propagating model uncertainty in this way. Stewart and Davis (1986) and

Hodges (1986) give approximations for mixed predictive distributions in general cases, but neither approximation has been appreciably examined or used.

Some Theory for Model-Mixing

It will be useful in what follows to describe in more detail some of the theory behind model-mixing. A particular scenario and model, together with specific parameter values conditional on the model, define a joint probability distribution for the data and any quantities to be predicted, so that one can conceive of the space of all possible model-scenario combinations as the collection of all such distributions. (Diaconis, 1977, and Meeden, 1986, give an explicit representation of this space of all models for a simple case.) What people usually refer to as a "model" with unknown parameters, which might in our context be better thought of as a "model-scenario combination," is a low-dimensional curve in this space, indexed by the parameters. Choosing a single "model" corresponds in a Bayesian sense to putting a prior distribution on model space that concentrates all its mass on this curve. In what follows we will use "model-scenario," or just "model" for short, to denote a subspace of model space, indexed parametrically in this way.

In a typical analysis in which model uncertainty is not propagated, the Bayesian approach requires data, a single "model" specifying a likelihood function $p(\text{data}|\theta)$ for the parameters θ given the data, a prior distribution for θ, $p(\theta)$, and a probability distribution for the future outcomes given the data and parameters, $p(\text{future}|\text{data},\theta)$. The prior for θ is updated through Bayes' Theorem to a posterior $p(\theta|\text{data})$; and the posterior predictive distribution (ppd) for future observables is calculated by integrating uncertainty about θ out of $p(\text{future}|\text{data},\theta)$, as follows:

$$(1) \qquad p(\text{future}|\text{data}) = \int p(\text{future}|\text{data},\theta)p(\theta|\text{data})d\theta.$$

All of the distributions in (1) are in fact conditional on the "model," but this is usually (incorrectly) suppressed in the notation. This calculation accounts for estimation and prediction uncertainty--the former through the integration in (1) and the latter through $p(\text{future}|\text{data},\theta)$--but "model-scenario" uncertainty is missing.

When there is structural uncertainty about the "model," as there almost always is, priors on model space like the one just discussed may not realistically reflect this uncertainty. To improve on usual practice it is necessary to entertain the possibility of more than one "model." The new ingredient in this more complete analysis of predictive uncertainty is a prior on model space, p("model"), that is a more accurate reflection of model and scenario uncertainty than the point mass on a single "model" implicit in (1). It is of course necessary to specify all of the usual distributions for each "model" given nonzero probability by the prior on model space--p(parameters|"model"), p(data|parameters,"model"), and p(future|data,parameters,"model")--but this would have to be done in any case as part of the sensitivity analyses one would ordinarily conduct.

Given these ingredients, the calculation of the posterior predictive distribution now requires an additional layer of integration, over uncertainty in the "model." In effect, one ends up mixing the ppd's conditional on each "model," p(future|data,"model"), to arrive at the overall ppd, using as mixing weights the posterior probabilities of the "models" given the data, p("model"|data). Symbolically,

$$(2)\quad p(\text{future}|\text{data}) = \int p(\text{future}|\text{data},\text{"model"})\,p(\text{"model"}|\text{data})\,d\text{"model"}.$$

When a small number of judiciously chosen "models" suffices, as will often be the case in practice, the integration in (2) becomes a simple summation. The "models" involved in calculations of this type typically have unknown parameters, so the process described by (2) also involves an integration over uncertainty in the parameters, which is not visible in the formula. In practice this integration is difficult to perform exactly; fast, accurate approximations to multidimensional integrals are needed to implement this idea.

Hierarchical Bayesian and Parametric Expansion Methods as Special Cases of Model-Mixing

Standard hierarchical Bayesian procedures can be viewed as one form of model-mixing, in which the prior distribution on model space takes a special form: Instead of putting all the prior mass on a single

"model," alternatives to that "model" indexed smoothly by one or more hyperparameters are also given nonzero prior weight. A similar phenomenon occurs with parametric expansion methods such as Box and Tiao's: Instead of all the prior mass resting on a single "model," an entire neighborhood around that "model" traced out continuously by the expansion parameters gets nonzero weight in the prior. But there is an important distinction between these methods and the techniques we propose to investigate.

With standard hierarchical Bayesian and parametric expansion procedures, the goal typically is estimation of some core parameters in the single, unexpanded "model." The introduction of expansion parameters, and the hierarchical placing of prior distributions on the parameters of the priors in such "models," both correspond to distributions on model space that are well suited to estimation activities of this type. In decisionmaking, by contrast, the goal is prediction of quantities relevant to action rather than inference about unobservable parameters. In practical contexts like the oil price example with its many models and scenarios, it is usually not possible to start with a single "model" and index all relevant departures from it with a small number of hyperparameters or expansion parameters. The prior distributions on model space whose predictive usefulness we propose to investigate are thus rather different from those in previous use in hierarchical Bayesian procedures and other model expansion methods. In such procedures the integration in Eq. (2) above really is an integration, over the hyperparameters or expansion parameters. We conjecture that in practical predictive contexts it will be more useful to perform a summation, over a range of discrete "models" (such as EMF's 10 models and 12 scenarios), chosen to span multidimensional departures from the assumptions implicit in an analyst's favorite "model."

Choosing the Prior Distribution on Model Space

It is clear with the above formulation of the problem that everything comes down to the choice of the prior on model space. How is this prior to be specified? Gaining experience with this process is an integral part of the research we propose, so it is difficult to speak to the point definitively now, but two general remarks can be made.

Analysts Already Use Priors on Model Space. Analysts reveal implicit priors on model space in their sensitivity analyses. To the extent that these implicit priors cover important departures from the favorite model, they are already well chosen, and analysts need only take the additional step of model-mixing to produce better-calibrated measures of predictive uncertainty. All that is needed to gain some improvement over current practice is modest success at "staking out the corners in model space"--finding the plausible variations on the model and scenario that strongly influence what actions would be taken.

Locating such high-leverage models and scenarios is partly context-specific, but the process could clearly benefit from some general methodological advances. Leverage, as we use the term here, depends on two things: prior plausibility of alternative modeling specifications, and the effect of such alternative specifications on predictions and uncertainty assessments. The first of these components depends intimately on the context. The second in effect requires calculating the derivative of the mapping from the prior on model space to the posterior predictive distribution, followed by a search for "directions" of departure from an analyst's favorite model with large derivatives in this sense. Freedman and Diaconis (1986) and Leamer (1982) have made some relevant progress, the former by working out such derivatives in some simple estimation settings and the latter by calculating bounds for posterior means when the prior comes from a class of alternatives, again in an estimation context. We propose to extend their work and related efforts to predictive settings.

Retrospective Calibration and Prospective Validation. The main theme of this proposal is the idea of assessing how wrong a prediction is likely to be. When this is done successfully, the prediction process is said to be *well calibrated*. Calibration can be checked by an exercise in which predictions--and estimates of how big the prediction errors are likely to be--are made, and resources are set aside to observe how big the prediction errors actually are once the future unfolds. Such exercises can be undertaken retrospectively as well as prospectively, by making predictions, using only data available at some point in the past, and examining the data from that point to the present

to see how accurate these forecasts would have been. This has two advantages: No waiting is needed for the future to reveal itself, and a number of points in the past can be examined, allowing a frequency history of prediction errors to be accumulated for calibration purposes.

When people realize they are out of predictive calibration, they are discovering that they should have had a different prior on model space. In other words, the trick in practice is to choose a prior on model space that will not be regretted in retrospect, where regret is measured by the quality of both the predictions and the "give-or-take" attached to the predictions. Retrospective calibration is a way to choose a prior on model space that would have performed well had it been used in the past. There is no free lunch here--this form of learning from experience requires an explicit judgment that the future is *exchangeable* (de Finetti, 1974/1975; Lindley and Novick, 1981) with relevant aspects of the past--but one should at least attempt to avoid repeating old forecasting mistakes, and retrospective calibration combined with prospective validation provides a framework for doing so.

Alternative Approaches

Other approaches can be considered for assessment and propagation of structural uncertainty. Several of these alternative approaches use ideas common in current statistical theory and practice (Alho and Spencer, 1985, and Stoto, 1983, have used ad hoc analogues of model-mixing in a frequentist context, for example, in assessing the uncertainty of population forecasts). Others involve schemes for representing uncertainty outside the probability calculus. We will discuss them in that order.

Frequentist Methods. (1) *Mixture Likelihoods.* An analyst might be willing to postulate that the model that will produce the outcome to be predicted is chosen at random from a collection of possible models. In the simplest case, for instance, the number of such models is finite, say m. This is a frequentist formulation that induces a *mixture likelihood*: Instead of a single likelihood $p(\text{data}|\theta)$, as in the discussion preceding Eq. (1) above, each model being mixed would have its own likelihood $p_i(\text{data}|\theta_i)$, where θ_i is a vector of parameters for model i, and the overall mixture likelihood would be

$$p(\text{data}|\pi,\theta) = \Sigma\, \pi_i p_i(\text{data}|\theta_i). \tag{3}$$

Here θ represents $\{\theta_i,\ i = 1, 2, \ldots, m\}$ and $\pi = (\pi_1, \ldots, \pi_m)$ is a vector of mixing parameters for the m models, where $0 < \pi_i < 1$ and the π_i sum to 1. Some care is needed in the specification of these models and parameters to avoid *identifiability* problems--to prevent situations where the data provide no basis to distinguish between two or more different model-parameter combinations. When such care is exercised, within the usual likelihood framework (see Edwards, 1972, for example) it is possible to estimate the mixing parameters and the θ_i and attach standard errors to those parameters. It may also be possible to construct frequentist predictive intervals that capture both within-model and between-model uncertainty; but difficulties with the frequentist predictive approach when applied to much simpler problems (Geisser, 1980a) make it clear that, if this is possible, it will not be easy.

(2) *Bootstrapping the Modeling Process*. Diaconis and Efron (1983) and Efron and Gong (1983) describe an interesting frequentist attempt to propagate model uncertainty involving the *bootstrap* (Efron 1979). To a dataset linking survival for 155 chronic hepatitis patients to 20 clinical and demographic covariates, Efron and Gong fitted a logistic regression model to predict mortality that had an internally assessed predictive misclassification rate of about 16 percent. They then drew 500 bootstrap samples from this dataset and fitted separate logistic regression models on each bootstrap sample, using the same modeling algorithm that had led them to their original model. They could have propagated model uncertainty at this point by constructing separate predictive intervals from each bootstrap dataset and combining them. Instead, on each bootstrap dataset, they calculated a quantity estimating the "overoptimism" of the 16 percent overall misclassification rate, and averaged these overoptimism estimates. The result was a revised misclassification estimate of about 20 percent, implying that the failure to propagate model uncertainty resulted in an overoptimism of about 30 percent.

By analogy with the usual relation between frequentist and Bayesian methods, we conjecture that both the mixture likelihood and model-bootstrapping frequentist approaches approximate the posterior predictive distribution given by the Bayesian model-mixing analysis outlined above, for a particular prior distribution on model space. This distribution places noninformative priors on the parameters conditional on the models, and a flat prior on the subregions in model space over which model uncertainty is being propagated. We will test this conjecture in Task 1 of the Workplan.

(3) *Cross-validation*. A third frequentist technique, *cross-validation* (Stone, 1974), has proven useful in selecting and evaluating prediction methods (see Breiman et al., 1984, for instance) and should also be helpful in model uncertainty propagation. In a cross-validation, various subsets of the full dataset are omitted in turn, and the competing prediction methods are applied to the remaining data and used to predict the omitted values. Then a measure of the predictive performance of the competing methods is aggregated across the omitted subsets, and the best method is selected. Typically, calibration considerations do not arise: The cross-validation permits comparison of the predictive methods only by producing point predictions and uncertainty estimates for these predictions, without an attempt to evaluate the quality of these uncertainty assessments. However, it is possible to apply cross-validation in our problem: If the competing predictors are formed from the possible mixtures of predictive distributions from a collection of models, then cross-validation could be used to evaluate the "give-or-take" produced by each mixture as well as the point prediction. We propose the use of such a procedure in Task 1.

Specifying Uncertainty Outside the Probability Calculus. Many authors have argued that the probability calculus is inadequate because it is difficult to do the necessary computations, or because probabilities of events are difficult to assess (Spiegelhalter, 1987, summarizes these arguments). In response to this perceived inadequacy, several other tools have been put forward for expressing and manipulating uncertainty; the three most prominent are *fuzzy set theory*

(Schmucker, 1984), *certainty factors* (Barr and Feigenbaum, 1982), and *belief functions* (Shafer, 1976, 1982). It is well known (de Finetti, 1974/1975; Lindley, 1987) that any system of expressing uncertainty and updating it to reflect learning from new data is incoherent (will produce inconsistencies) unless it yields results identical to those of standard probability theory. Because these three alternatives produce answers different from those of probability in some instances, these inconsistencies are always a hazard. Moreover, all of the problems that attend the use of probability are present with these alternatives (Spiegelhalter, 1987).

Fuzzy Set Theory. Fuzzy set theory has been suggested for many uses other than the representation of uncertainty; we discuss it only in that role. In traditional set theory, any object either belongs or does not belong to a given set: If the function m(e) denotes the membership status of an object e in a given set, traditionally it can only take the values 0 and 1, denoting nonmembership and membership respectively. Fuzzy set theory is intended to extend set theory to cases for which set membership is not clear cut. Thus the function m(e) is allowed to take values in the closed interval from 0 to 1, where 0 means that e is certainly not in the set, 1 means that it certainly is, and numbers between zero and one represent intermediate degrees of membership.

For example, proponents of fuzzy set theory argue that it is often difficult for risk analysts to specify a precise probability p for some event (in a fault tree, say), so that probabilistic methods are inadequate (Unwin, 1986; Zadeh, 1984). To apply fuzzy set theory, the analyst gives a natural language assessment of the probability ("extremely low," for example), and either the analyst (Unwin, 1986) or an automated risk assessment system (Schmucker, 1984) attaches to that natural language assessment a membership function m(q), which represents the degree to which q belongs to the fuzzy set "p's that are extremely low." Fuzzy set theory then supplies rules for combining these membership functions for unions and intersections of fuzzy sets.

In fact, for this application, and all others involving the representation of uncertainty, fuzzy set theory fills no role that probability cannot fill with equal or greater ease (Lindley, 1987), for the fuzzy set membership function is nothing more than an unnormalized

probability distribution. In the example above, uncertainty about p can be represented by a probability distribution, as in a standard Bayesian analysis of data arising as proportions, and manipulated as a probability distribution. Thus fuzzy sets do not represent a real alternative to probability theory.

Certainty Factors. Certainty factors (CFs) were developed by the creators of the medical expert system MYCIN (Barr and Feigenbaum, 1982; Shortliffe, 1976), who found that doctors had difficulty assessing and interpreting probabilities. CFs measure the confidence in the truth of an implication from a premise to a consequence, with -1.0 representing complete confidence that the implication is false and 1.0 complete confidence that it is true. MYCIN specified rules for combining CFs for several uncertain propositions. Recently, some of the MYCIN workers have modified their rules for combining CFs so that they more closely resemble those of belief functions (Gordon and Shortliffe, 1984, 1985). In view of the general considerations noted above, and given that its own originators consider the approach of belief functions to be a more promising route for future development, we see little reason to pursue CF theory further.

Belief Functions. Belief functions (Dempster, 1966; Shafer, 1976, 1982) grew out of the same idea as fuzzy set theory: that Bayesian theory asks for too much. Here, though, the criticism of Bayesian methods takes a different form. Sometimes a given piece of evidence, where "evidence" could be an analyst's subjective judgment, allows the analyst to specify probabilities for only a small number of sets out of the relevant group of possible sets. For example, for a problem involving an uncertain inference about the positive integers, the evidence may allow the specification of probabilities only for the sets {1, 2, ..., 10}, {11, 12, ..., 20}, and so on, rather than for each integer individually. If the analyst has several such pieces of evidence, all specified for the same sets, and he either can assert that the pieces of evidence are independent or he can specify the dependence, in the usual probabilistic fashion, then Bayesian theory can be used. Otherwise, some other theory is needed.

Belief functions are intended to fill this gap. The theory of belief functions supplies rules for combining incompletely specified probability distributions. As such, belief function theory can be viewed as an extension of Bayesian theory and not a dramatic departure (Dempster, 1987). However, belief functions are not yet successful at performing this extension. A user of belief functions must still either assert that the pieces of information are independent, or specify the form of the dependence; but the theory of belief functions contains no definition of dependence and independence (Watson, 1987). Also, belief functions offer no computational advantages over probability methods (Lindley, 1987; Spiegelhalter, 1987). Thus, although the theory of belief functions has promise, at this stage it is only being given its first tests in applications and does not yet rival standard probability theory.

Applicability of the Proposed Approach

Bayesian theory appears to be the best available tool for making progress on the problem described in the Introduction. Our proposed solution is not a panacea; it is clearly not appropriate for some problems. For example, our proposal will not apply to problems for which data are not clearly defined and specifiable (von Winterfeldt and Edwards, 1986, Chapter 6). Similarly, it will not help in situations in which the analyst does not have specific measurements of input variables that are presumed to be causally related, either directly or through proxies, to the outcome measure of interest (Hahn, McRae, and Milford, 1986). It is not likely to patch up a gross deficiency of conception like that underlying fault-tree analysis (see Kamins, 1975).

Our proposed solution *will* be useful in cases in which an analyst has clearly defined and specifiable measures of input and output variables, but lacks a sure specification for the scenario and the functional form relating the input variables to the probability distribution of the outcome variables. This is the case in many statistical applications (Andrews and Herzberg, 1985, give many examples fitting this description) and in many examples of risk analysis (e.g., Hattis and Smith, 1985). It will be especially applicable to situations

in which the analyst can apply the rich collection of statistical diagnostic methods (e.g., for linear regression, Atkinson, 1985; Cook and Weisberg, 1982; Leamer, 1978; Weisberg, 1985).

Within this range of usefulness, the proposed approach has some apparent and some real limitations. We first discuss the apparent limitations and then those that are of greater concern.

Apparent Limitations

"Black Box" Criticisms. Mazur (1980), Speed (1985), and Spencer, Diegert, and Easterling (1987), among others, have criticized the use of probabilistic methods in nuclear risk assessment. They argue that in the conventional approach to such risk assessments, all manner of probabilities--posterior probabilities from statistical analyses, expert opinions, unjustified conventions--are combined without heed to their different natures, the product being an unqualified and unfounded assertion of one single probability of a catastrophic release of radiation.

These criticisms reflect an old fear about Bayesian methods: that the subjective element (in the prior distribution) will be so mixed up with the data element that consumers of the analysis will not be able to distinguish one from the other. Nothing makes this necessary; proponents of Bayesian methods have argued repeatedly that the proper method of presentation allows the consumer to insert his own prior, or at least presents the analyst's prior so that the consumer can see if he agrees. (Diamond and Forrester, 1983, for instance, give a graphical method that allows a user to insert his own prior.) Unfortunately, in probabilistic risk assessment, this warning has not been heeded. This has occurred partly because the sort of presentation method mentioned above is applicable to cases where one has a likelihood and a prior and wishes to distinguish the information conveyed by each, and probabilistic risk assessment does not fit this formula. In the contexts in which our methods show promise, it is straightforward to present both mixed and unmixed results so that readers can mix with their own priors. We will demonstrate this in all three of our specific tasks in the Workplan.

Biases in Judgment and Elicitation. Since the early 1970s, psychologists have learned a great deal about how people assess probabilities. (Kahneman, Slovic, and Tversky, 1982, is a survey of this research.) Tversky and Kahneman (1974) showed that people use three heuristic techniques to assess probabilities, each of which creates predictable biases. These biases show some correction with training, if the probability assessor is given sufficient time to do the assessment (Alpert and Raiffa, 1982; Fischoff, Slovic, and Lichtenstein, 1980; von Winterfeldt and Edwards, 1986). Tversky and Kahneman (1981) found similar effects arising from the method of elicitation: A person's probability assessment is conditioned by irrelevant particulars of the elicitation method (by the "framing" of the elicitation, in other words). There is evidence that this, too, can be overcome if the probability is elicited by more than one method (von Winterfeldt and Edwards, 1986).

Some have interpreted these studies as meaning that any method relying on probability elicitation is founded on sand. The authors of this literature contradict this inference: The issue is not whether people can judge probabilities but how they can be helped to do it better (von Winterfeldt and Edwards, 1986). In fact, the approach we propose can be understood as a contribution to "debiasing." In most statistical problems, model diagnostic methods (Atkinson, 1985; Cook and Weisberg, 1982; Leamer, 1978; Weisberg, 1985) provide a variety of ways to seek plausible variations on one's model, and emerging methods may soon allow some automation in the generation of alternative models (Gale, 1986a, 1986b; Glymour et al., 1987). Conscientiously applied, the methods we propose to investigate can counteract the framing problem at least partly by forcing the analyst to consider alternative models.

Real Issues

Too Many Models? Once Pandora's Box is open and more than one model is entertained, how is the number of models to be limited? Fortunately, putting a prior on the space of all models is not the same as putting priors on the parameters of every conceivable "model." We conjecture (and it will be a central goal of Tasks 1 and 2 to verify or

disprove this conjecture) that there is no need to include all possible models in the model-mixing exercise. Instead, the ensemble of models need cover only the range of *a priori* plausible models differing in ways that give substantially different predictions and "give-or-takes" for the outcome variable. For example, in a linear regression problem, it may not be necessary to consider all possible transformations of the outcome variable (as in the Box-Cox method, for instance; see Weisberg, 1985), as long as the ensemble of models includes the logarithmic transformation, which can introduce a sharp change in predictions from the resulting model when back-transformed to the original scale. In the few examples we have begun to explore, mixing even three or four well-chosen alternative models with one's favorite model produces better retrospective calibration and thus more appropriate hedging against uncertainty.

Infinite Regress: You're Only as Good as Your Prior on Model Space. If the analyst is to assess and propagate uncertainty about the model, then why not uncertainty about the uncertainty about the model, and so on in an infinite regress (Geisser, 1980b; Hodges, 1987)? It is inescapable as a subjectivist that at some point You must express a judgment as a probability distribution without further qualifications (Hodges, 1987). Current practice is to choose a single model and use it as if it were known to be true. This is a particular choice about the level of judgment: Stop with a single model. The failures of this approach in situations in which no model can be asserted with certainty reflect another hard fact: As noted in the section on retrospective calibration and prospective validation, You're only as good as Your prior on the space of possible models. The selection of a single model is a strong piece of information; its use has consequences.

The proposed method is a way to weaken that piece of information, by making more diffuse the distribution of probability on the space of all models. As described previously, we conjecture that it will often be possible to evaluate candidate distributions on model space by a retrospective calibration. This exercise will provide some basis for believing that the chosen distribution on model space will give good prospective assessments of the uncertainty of predictions. Modest as it sounds, there is no reason to believe that it will ever be possible to

make any stronger promises than this, since (as noted before) the success of this form of borrowing from past experience rests on an explicit subjective exchangeability judgment that may or may not prove to be correct in practice. In general, self-calibrating priors do not exist (Oakes, 1985).

THE WORKPLAN

The goal of the work proposed here is to provide practical tools for implementing the approach described above, and to gain experience in the use of this approach. Task 1 concentrates on computational tools, and application in an uncomplicated situation. The subsequent tasks provide progressively more demanding applications of the model-mixing method.

TASK 1: MODEL UNCERTAINTY ASSESSMENT AND PROPAGATION IN LINEAR AND LOGISTIC REGRESSION

For several years in the 1970s, two statistics professors in the RAND Graduate School, Carl Morris and Dan Relles, used their Ph.D. students in Policy Analysis as subjects in an informal study. They gave each student a small sample from a large dataset gathered in a designed experiment (Newhouse et al., 1981), in which the experimental units were families, the outcome variable was the annual expenditure by each family on health care, and the explanatory variables included characteristics of the family members and of the type of health insurance plan that the family had been assigned in the experiment. Each student was asked to apply to his or her dataset the statistical techniques the class had studied, including all the standard diagnostic techniques. The products were to be a fitted model, a prediction of the annual expenditure by a family that had been chosen at random and held out of all of the small samples, and an assessment of the accuracy of that prediction. Some of the more sophisticated students used cross-validation to avoid overfitting in the selection of their models.

The results were striking: Each year, the variation in the errors of the students' predictions was about twice as large as the internal assessments of accuracy indicated it should be. The standard, conventionally sanctioned methods that the students used to assess the

variability of their predictions systematically understated the actual size of their predictive errors.

We propose to use this situation--a dataset in hand, collected by one of the members of the project team, straightforwardly amenable to the standard statistical modeling and diagnostic tools--and others like it, including Efron and Gong's (1983) logistic regression example, to develop the computational and presentation tools discussed above. These situations are uncomplicated and will allow us to concentrate on the more technical aspects of the proposed program of research.

Subtask 1. Write and test computer code to execute and compare five methods of computing approximate predictive distributions, namely (in rough order of increasing accuracy and computer intensity) those of Hodges (1986), Tierney and Kadane (1986), Morris (1987), Stewart and Davis (1986), and Smith et al. (1985). As noted in the discussion of the theory of model-mixing, approximations are usually necessary because the integrals involved often cannot be computed in closed form. Hodges' method is based on approximating the low-order moments of the posterior predictive distribution and using the relevant member of a simple location-scale family. The Stewart-Davis approach is based on Monte-Carlo integration with judiciously chosen weight functions. Both of these approximation methods have been coded for special cases but not for more general situations. Tierney and Kadane's method, based on the Laplace approximation to integrals, has been embodied in an S function (Becker and Chambers, 1984), but as yet it produces predictive distributions only for single models, not for collections of models. We will do the extension to the model-mixing case. Morris' method extends Tierney and Kadane's by generalizing the Laplace approximation; it has been coded only in a few special cases so far. Smith et al. offer the most accurate and computer-intensive method of the five. Their method integrates out the parameters to form a predictive distribution by using Gauss-Hermite approximations to the distributions of carefully selected re-parameterizations. Like Tierney and Kadane's method, it has so far been used only for predictive distributions for single models, and we will extend it to the model-mixing case.

Subtask 2. Redo the model uncertainty analyses of small samples of the health insurance experiment and Efron-Gong logistic regression data. Application of the standard statistical modeling and diagnostic techniques will produce a collection of models, of varying but non-trivial degrees of plausibility, that can then be mixed as discussed above. For each of the small samples on which this is performed, the resulting model mixture can be "retrospectively" calibrated on the small sample itself. Then each of the model mixtures can be validated prospectively, by predicting a part of the larger dataset that was selected at random and held out from the sampling that produced the smaller datasets. This will provide a test of the model mixing on grounds as favorable as possible: The prospective validation will use data that are exchangeable, by construction, with those in the small samples. In this setting we will also explore the connections between frequentist and Bayesian model-mixing procedures discussed above, and investigate the identification of high-leverage models by computing derivatives of the mapping from prior to posterior predictive distributions.

TASK 2: A RETROSPECTIVE CASE STUDY--FORECASTING OIL PRICES

The first task will provide technical tools and experience using them in a favorable situation. The next two tasks will apply these tools in situations that are less favorable in that, unlike Task 1, there is substantially less reason to treat the future (the data used for prospective validation) as exchangeable with the past (the data used for modeling, prediction, and retrospective calibration). Thus these last two tasks will permit an examination of the relation between retrospective calibration and prospective validation.

In Task 2 of this project, we propose to work jointly with EMF investigators to use the forecasting of oil prices as a retrospective case study of model-scenario uncertainty assessment and propagation.

Subtask 1. Data acquisition and preparation for analysis. We need to understand the models, scenarios, and data used by EMF for the modeling and predictions described in the Introduction, and to acquire the corresponding data that they have accumulated since those

predictions were made. We also need to obtain the internal assessments of predictive uncertainty each produced, which have not yet been published. We will not need to have the actual EMF models running at RAND; for our purposes, it will be enough to have the predictions and internal assessments of uncertainty for each model-scenario combination. We have secured the cooperation of the director of EMF, John Weyant, and we propose to spend sufficient time at EMF to become familiar with the models, how they were used, and so on.

Subtask 2. Assessment and propagation of model-scenario uncertainty and retrospective calibration using the tools developed in Task 1. Using the predictions and internal assessments of predictive uncertainty from the EMF models, and the actual oil price data for the period for which those models were estimated, in collaboration with EMF researchers we will examine various specifications of prior probabilities of the models to find those that are, retrospectively, the best calibrated. This will involve dividing the past up into multiple time periods and developing a frequency history of predictive errors, and it will be necessary to explore various ways of doing this.

In addition, we will make a "prospective" validation for the period subsequent to the EMF forecasts discussed in the Introduction (1980 to 1987), using the retrospectively calibrated model mix discussed above. This will require specifications of the probabilities of the different scenarios; in conjunction with the EMF researchers, several such specifications will be generated and considered. We conjecture that any of these specifications will provide a measure of predictive uncertainty that is better calibrated than that provided by using a single scenario.

TASK 3: PROSPECTIVE CASE STUDIES

Task 2 will be a more demanding test of the model mixing approach than Task 1; Task 3 will be more demanding still. In this task, we will apply the model-mixing approach prospectively in collaboration with researchers at RAND and elsewhere on current research--that is, to predictions where we do not know the right answer.

Subtask 1. Choice of appropriate substantive problems and familiarization with the substantive issues. We will select a RAND project for this task. One candidate is a project to evaluate the

desirability of various schemes for regulating chlorofluorocarbon emissions; another is ongoing work on deregulating bulk power transfers among electric utilities. In addition, we will consider applying our methods to other appropriate problems that arise. One of our team (Carl Morris) consults with a computer research firm that is optimizing its process for chip design. Several factors are involved in this optimization, and the apparent theoretical optimum depends on the statistical model used. Model uncertainty thus needs to be included in a decision-theoretic solution to this problem. In another project, Prof. Morris is estimating a production function relating team winning percentages and other factors to the attendance revenues of professional sports franchises. Uncertainty, including model uncertainty, must be properly assessed so that, in the application of these estimated equations to salary bargaining, the bargaining range is limited properly.

Subtask 2. Choice of the models and scenarios to be used in the propagation of model-scenario uncertainty; assessment and propagation of model-scenario uncertainty; retrospective calibration and prospective prediction and validation. This subtask will be performed in collaboration with the relevant substantive researchers for the problems selected in subtask 1, and the activities will be essentially the same as those in subtask 2 of Task 2. The details will depend on the context of the problems chosen.

BIBLIOGRAPHY

Alho, J.M. and Spencer, B.D. (1985). Uncertain population forecasting. *Journal of the American Statistical Association*, June 1985, Vol. 80, No. 390, pp. 306-314.

Alpert, M. and Raiffa, H. (1982). A progress report on the training of probability assessors. In *Judgment Under Uncertainty: Heuristics and Biases* (D. Kahneman, P. Slovic, A. Tversky, eds.), Cambridge University Press, Cambridge, pp. 294-305.

Andrews, D.F. and Herzberg, A.M. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker.* Springer-Verlag, New York.

Atkinson, A.C. (1985). *Plots, Transformations, and Regression.* Clarendon Press, Oxford.

Barr, A. and Feigenbaum, E.A. (1982). *The Handbook of Artificial Intelligence*, vol. II. William Kaufman, Los Altos, California.

Becker, R.A. and Chambers, J.M. (1984). *S: An Environment for Data Analysis and Graphics.* Wadsworth, Belmont, California.

Berger, J.O. (1984). The robust Bayesian viewpoint (with discussion). In *Robustness of Bayesian Analysis* (J.B. Kadane, ed.), North-Holland, New York, pp. 63-144.

Berger, J.O. and Berliner, L.M. (1986). Robust Bayes and empirical Bayes analysis with ε-contaminated priors. *Annals of Statistics*, Vol. 14, pp. 461-486.

Box, G.E.P. (1980). Sampling and Bayes inference in scientific modeling and robustness (with discussion). *Journal of the Royal Statistical Society*, Series A, Vol. 143, No. 4, pp. 383-430.

Box, G.E.P. and Hill, W.J. (1967). Discrimination among mechanistic models. *Technometrics*, Vol. 9, No. 1, Feb. 1967, pp. 57-71.

Box, G.E.P. and Tiao, G.C. (1962). A further look at robustness via Bayes' Theorem. *Biometrika*, Vol. 49, No. 3 and 4, pp. 419-432.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, California.

Clayton, M.K., Geisser, S., and Jennings, D.E. (1986). A comparison of several model selection procedures. In *Bayesian Inference and Decision Techniques* (P.K. Goel, A. Zellner, eds.), North-Holland, New York, pp. 425-439.

Cook, R.D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society*, Series B, Vol. 48, No. 2, pp. 133-169.

Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.

Covello, V.T., von Winterfeldt, D., and Slovic, P. (1986). Risk communication: A review of the literature. *Risk Abstracts*, Vol. 3, No. 4, pp. 171-182.

de Finetti, B. (1974/1975). *The Theory of Probability*, Volumes 1 and 2. Wiley, New York.

Dempster, A.P. (1966). New methods for reasoning toward posterior distributions based on sample data. *Annals of Mathematical Statistics*, Vol. 37, pp. 355-374.

Dempster, A.P. (1985). Probability, evidence, and judgment. *Bayesian Statistics*, Vol. 2, pp. 119-131.

Dempster, A.P. (1987). Discussion of papers by Lane and Cooper on "Probability and artificial intelligence in medicine," presented to the annual meetings of the American Statistical Association, San Francisco, August 17, 1987.

Diaconis, P. (1977). Finite forms of de Finetti's theorem on exchangeability. *Synthese*, Vol. 36, pp. 271-281.

Diaconis, P. and Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, Vol. 127, pp. 116-130.

Diamond, G.A. and Forrester, J.S. (1983). Clinical trials and statistical verdicts: probable grounds for appeal. *Annals of Internal Medicine*, Vol. 98, pp. 385-394.

Draper, D. (1987). On exchangeability judgments in predictive modeling, and the role of data in statistical research. Comment on C.R. Rao, "Prediction of Future Observations in Growth Curve Models," *Statistical Science*, in press.

DuMouchel, W.H. and Harris, J.E. (1983). Bayes methods for combining the results of cancer studies in humans and other species (with discussion). *Journal of the American Statistical Association*, Vol. 78, pp. 293-315.

Edwards, A.W.F. (1972). *Likelihood*. Cambridge University Press, Cambridge.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, Vol. 7, pp. 1-26.

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, Vol. 37, No. 1, pp. 36-48.

EIA (1982). *Outlook for World Oil Prices*. Energy Information Administration, U.S. Department of Energy.

EMF (1982). *World Oil: Summary Report*. EMF Report 6 (February 1982), Energy Modeling Forum, Stanford University.

Fischoff, B., Slovic, P., and Lichtenstein, S. (1980). Labile values: a challenge for risk assessment. In *Society, Technology, and Risk Assessment* (J. Conrad, ed.), Academic Press, New York, pp. 57-66.

Freedman, D. and Diaconis, P. (1986). On the consistency of Bayes estimates (with discussion). *Annals of Statistics*, Vol. 14, No. 1, pp. 1-67.

Gale, W.A. (1986a). REX review. In *Artificial Intelligence and Statistics* (W.A. Gale, ed.), Addison-Wesley, Reading Massachusetts.

Gale, W.A. (1986b). Student phase I: A report on work in progress. In *Artificial Intelligence and Statistics* (W.A. Gale, ed.), Addison-Wesley, Reading Massachusetts.

Geisel, M.S. (1974). Bayesian comparisons of simple macroeconomic models. In *Studies in Bayesian Econometrics and Statistics* (S.E. Fienberg and A. Zellner, eds.), North-Holland/American Elsevier, New York, pp. 227-256.

Geisser, S. (1980a). A predictivistic primer. In *Bayesian Analysis in Econometrics and Statistics* (A. Zellner, ed.), North Holland, New York.

Geisser, S. (1980b). Predictive sample reuse techniques for censored data (with discussion). In *Bayesian Statistics* (J.M. Bernardo et al., eds.), University Press, Valencia, Spain, pp. 430-468.

Glymour, C., Scheines, R., Spirtes, P., and Kelly, K. (1987). *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling*. Academic Press, New York.

Good, I.J. (1980). Some history of the hierarchical Bayesian methodology. In *Bayesian Statistics* (J.M. Bernardo et al., eds.), University Press, Valencia, Spain.

Gordon, J. and Shortliffe, E.H. (1984). The Dempster-Shafer theory of evidence. In *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project* (B.G. Buchanan, E.H. Shortliffe, eds.), Addison-Wesley, Reading, Massachusetts.

Gordon, J. and Shortliffe, E.H. (1985). A method for managing evidential reasoning in hierarchical hypothesis spaces. *Artificial Intelligence*, vol. 26, pp. 323-358.

Hahn, R.W., McRae, G.J., and Milford, J.B. (1986). The role of uncertainty in environmental policy design. Carnegie-Mellon University, unpublished manuscript.

Harrison, P.J. and Stevens, C.F. (1971). A Bayesian approach to short-term forecasting. *Operational Research Quarterly*, Vol. 22, No. 4, pp. 341-362.

Harrison, P.J. and Stevens, C.F. (1976). Bayesian forecasting. *Journal of the Royal Statistical Society*, Series B, Vol. 38, No. 3, pp. 205-247.

Hattis, D. and Smith, J.A. (1985). What's wrong with quantitative risk assessment? Presented at the Conference on Moral Issues and Public Policy Issues in the Use of the Method of Quantitative Risk Assessment, Georgia State University, September 26-27, 1985. To appear in *Biomedical Ethics Reviews*.

Hodges, J.S. (1986). Parsimony and prediction. Paper delivered at the Seminar on Bayesian Inference in Econometrics, University of California, Riverside, October 26, 1986.

Hodges, J.S. (1987). Uncertainty, Policy Analysis, and Statistics. *Statistical Science*, Vol. 2, No. 3, August 1987, in press.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Oxford University Press, London.

Kahneman, D., Slovic, P., and Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge.

Kamins, M. (1975). A reliability review of the Reactor Safety Study. The RAND Corporation, P-5413, April 1975.

Leamer, E.E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Wiley, New York.

Leamer, E.E. (1982). Sets of posterior means with bounded variance priors. *Econometrica*, Vol. 50, pp. 725-736.

Leamer, E.E. (1985). Sensitivity analyses would help. *American Economic Review*, Vol. 75, pp. 308-313.

Lindley, D.V. (1987). The probability approach to the treatment of uncertainty in artificial intelligence and expert systems. *Statistical Science*, Vol. 2, No. 1, Feb. 1987, pp. 17-24.

Lindley, D.V. and Novick, M.R. (1981). The role of exchangeability in inference. *Annals of Statistics*, Vol. 9, pp. 45-58.

Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates in the linear model (with discussion). *Journal of the Royal Statistical Society*, Series B, Vol. 34, pp. 1-41.

Mazur, A. (1980). Societal and scientific causes of the historical development of risk assessment. In *Society, Technology, and Risk Assessment* (J. Conrad, ed.), Academic Press, New York, pp. 151-157.

McCullagh, P. and Nelder, J.A. (1983). *Generalized Linear Models*. Chapman and Hall, New York.

Meeden, G. (1986). Sufficiency and partitions of the class of all possible discrete distributions. *The American Statistician*, Vol. 40, pp. 42-44.

Morris, C.N. (1983). Parametric empirical Bayes inference: Theory and Applications. *Journal of the American Statistical Association*, Vol. 78, pp. 47-65.

Morris, C.N. (1987). Approximating posterior distributions and posterior moments. Technical Report #51, Center for Statistical Sciences, University of Texas. To appear in *Bayesian Statistics 3*, proceedings of the Third Valencia (Spain) International Meeting on Bayesian statistics, with discussion.

Newhouse, J.P., Manning, W.G., Morris, C.N., Orr, L.L., Duan, N., Keeler, E.B., Leibowitz, A., Marquis, K.H., Marquis, M.S., Phelps, C.E., and Brook, R.H. (1981). Some interim results from a controlled trial of cost sharing in health insurance. *The New England Journal of Medicine*, Vol. 305, pp. 1501-1507.

Oakes, D. (1985). Self-calibrating priors do not exist (with discussion). *Journal of the American Statistical Association*, Vol. 80, No. 390, pp. 339-342.

Schmucker, K.J. (1984). *Fuzzy Sets, Natural Language, Computations, and Risk Analysis*. Computer Science Press, Rockville, Maryland.

Sclove, S., Morris, C.N., and Radhakrishnan, R. (1972). Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution. *Annals of Mathematical Statistics*, Vol. 43, No. 5, pp. 1481-1490.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton.

Shafer, G. (1982). Belief Functions and parametric models. *Journal of the Royal Statistical Society*, Series B, vol. 44, No. 3, pp. 322-352.

Shortliffe, E.H. (1976). *Computer-based Medical Consultation: MYCIN.* American Elsevier, New York.

Skene, A.M., Shaw, J.E.H., and Lee, T.D. (1986). Bayesian modeling and sensitivity analysis. *The Statistician*, Vol. 35, pp. 281-288.

Smith, A.F.M. (1983). Bayesian approaches to outliers and robustness. In *Specifying Statistical Models: From Parametric to Non-parametric, Using Bayesian or Non-Bayesian Approaches* (J.P. Florens, M. Mouchart, J.P. Raoult, L. Simar, A.F.M. Smith, eds.), Springer-Verlag, New York.

Smith, A.F.M., Skene, A.M., Shaw, J.E.H., Naylor, J.C., and Dransfield, M. (1985). The implementation of the Bayesian paradigm. *Communications in Statistics: Theory and Methods*, Vol. 14, pp. 1079-1102.

Smith, A.F.M. and West, M. (1983). Monitoring renal transplants: An application of the multiprocess Kalman filter. *Biometrics*, Vol. 39, pp. 867-878.

Speed, T.P. (1985). Probabilistic risk assessment in the nuclear industry: Wash-1400 and beyond. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L.M. LeCam and R.A. Olshen, eds.), Wadsworth, Belmont, California, pp. 173-200.

Spencer, F.W., Diegert, K.V., and Easterling, R.G. (1987). Statistically based uncertainty assessments in nuclear risk analysis. Paper presented to the annual meetings of the American Statistical Association, San Francisco, August 20, 1987.

Spiegelhalter, D.J. (1987). Probabilistic expert systems in medicine: Practical issues in handling uncertainty. *Statistical Science*, Vol. 2, No. 1, Feb. 1987, pp. 25-30.

Stewart, L. and Davis, W.W. (1986). Bayesian posterior distributions over sets of possible models with inferences computed by Monte Carlo integration. *The Statistician*, Vol. 35, pp. 175-182.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society*, Series B, Vol. 36, No. 2, pp. 111-147.

Stoto, M.A. (1983). The accuracy of population projections. *Journal of the American Statistical Association*, March 1983, Vol. 78, No. 381, pp. 13-20.

Syme, J. (1987). "Forecast models and policy analysis: The case of oil prices." The RAND Corporation, N-2524-RC, forthcoming.

Tierney, L. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, Vol. 81, pp. 82-86.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, Vol. 185, Sept. 27, 1974, pp. 1124-1131.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, Vol. 211, Jan. 30, 1981, pp. 453-458.

Unwin, S.D. (1986). A fuzzy set theoretic foundation for vagueness in uncertainty analysis. *Risk Analysis*, Vol. 6, No. 1, pp. 27-34.

von Winterfeldt, D. and Edwards, W. (1986). *Decision Analysis and Behavioral Research*. Cambridge University Press, Cambridge.

Watson, S.R. (1987). Comment on papers by Shafer, Lindley, and Spiegelhalter on "The Calculus of Uncertainty in Artificial Intelligence and Expert Systems," *Statistical Science*, Vol. 2, No. 1, pp. 30-32.

Weisberg, S. (1985). *Applied Linear Regression*, second edition. Wiley, New York.

West, M. (1986). Non-normal multi-process models. University of Warwick Department of Statistics Research Report # 81.

Zadeh, L.A. (1984). Foreword to Schmucker (1984).

Zellner, A. (1984). Posterior odds ratios for regression hypotheses: general considerations and some specific results. In *Basic Issues in Econometrics* (A. Zellner, ed.), University of Chicago Press, pp. 275-305.