Comment

Peter A. Morris

"Combining Probability Distributions: A Critique and an Annotated Bibliography," by Genest and Zidek provides a review of methodologies for aggregating probabilistic judgments. While I would evaluate the strengths and weaknesses of the different techniques somewhat differently, I found the paper a useful compendium of an assortment of different approaches.

Some of the approaches described are quite interesting, while others appear superficial and somewhat naive. However, it is difficult to review a review paper without performing another review and I'd like to avoid that here. Instead, I'd like to comment on the process of evaluating different expert aggregation methods.

1. PURPOSE OF THE AGGREGATION

I find the most useful way of thinking about the problem of expert aggregation is to view it as helping an individual update his or her state of information based on reception of an expert's advice. These states of information are typically represented with probabilities (or some form of probability statement).

Many of the approaches described in this paper are fuzzy to me in spite of their mathematical precision because they seem to be making an attempt to form an "aggregate opinion." This, in my mind, is an ill defined concept. Probability is a measure of an individual's state of information about an uncertain event. There is no such thing as a "joint state of information." Individuals have opinions; groups do not.

In the rare situation in which each expert in a group shares precisely the same state of information and the same probability distribution, then that probability distribution might reasonably be termed the "opinion of the group." However, when the experts inevitably disagree (even after intensive interaction), any so-called "consensus" or agreement on a distribution is necessarily a group *decision*, not the reflection of a "joint state of information." In particular, there is no logical reason for a group to achieve consensus in their probabilities if they start with different opinions and have different feelings about each other's expertise.

Thus, I believe the problem of aggregation, in order to be well defined, is the problem of updating an individual's state of information based on the reception of a set of expert probabilities. The best we can do is ask for a single individual, "What is the appropriate way to update a prior probability in light of learning about others' probabilities?" I agree with Lindley that other approaches have "an element of adhockery."

Thinking in this way is not only more satisfying conceptually, but provides a device for obtaining physical insights. In evaluating each approach, we can consider the specific assumptions a single individual would have to make in order to combine expert opinions in the proposed way. For example, suppose we are considering a linear weighting formula for combining two weather forecaster's rain probabilities. We can ask specific questions, like: "Does knowledge that one expert's probability of rain is high indicate that it is likely that the other expert's probability will be high as well?" If the answer is "yes" as I think it would be in most cases, then the linear weighting scheme makes no sense.

2. FUNDAMENTAL ISSUES

The discussion in the review article is fairly mathematical, and as such provides good in-depth material for researchers in expert use. However for those who are not "experts on experts," some basic issues in expert resolution may be masked by all the mathematics. In my view there are several fundamental issues that any realistic combination methodology must address (or at least *explicitly* not address) to be viable. Testing against these basic issues often helps determine quickly whether a method is reasonable, and many techniques that appear quite sophisticated fail simple reasonability checks. Four fundamental issues are:

- Nonindependent experts
- Event probabilities and underlying frequencies
- Calibration
- Level at which aggregation is performed

Nonindependent Experts

The issue of nonindependence among experts is critically important because it significantly affects the amount of uncertainty that one associates with the group. It is the single most important issue in practical applications. Yet, it is often ignored in many expert combination formulas, probably because it is extremely difficult to think about, much less quantify.

Peter A. Morris is a Principal with Applied Decision Analysis, Inc., 3000 Sand Hill Road, Building 4, Suite 255, Menlo Park, CA 94025, and a Consulting Associate Professor with the Department of Engineering-Economic Systems at Stanford University.

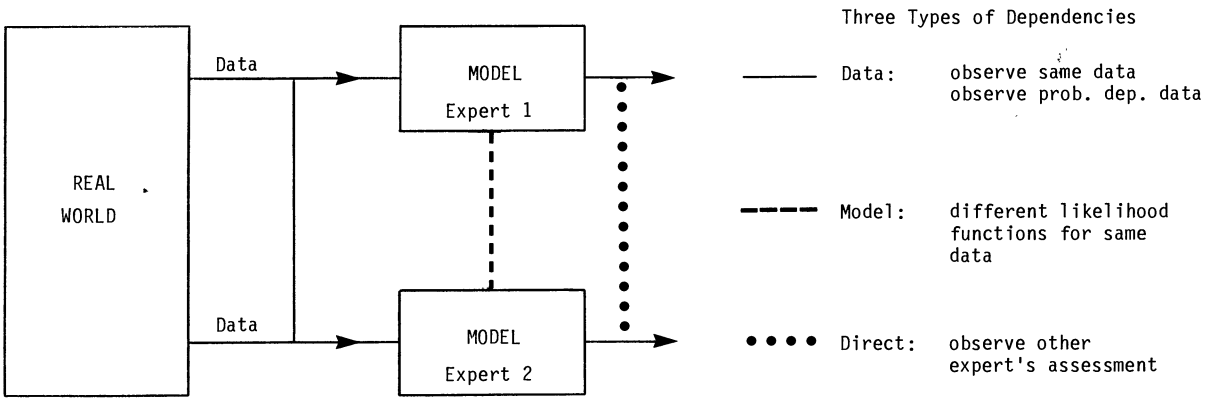

FIG. 1. *Nonindependent experts.*

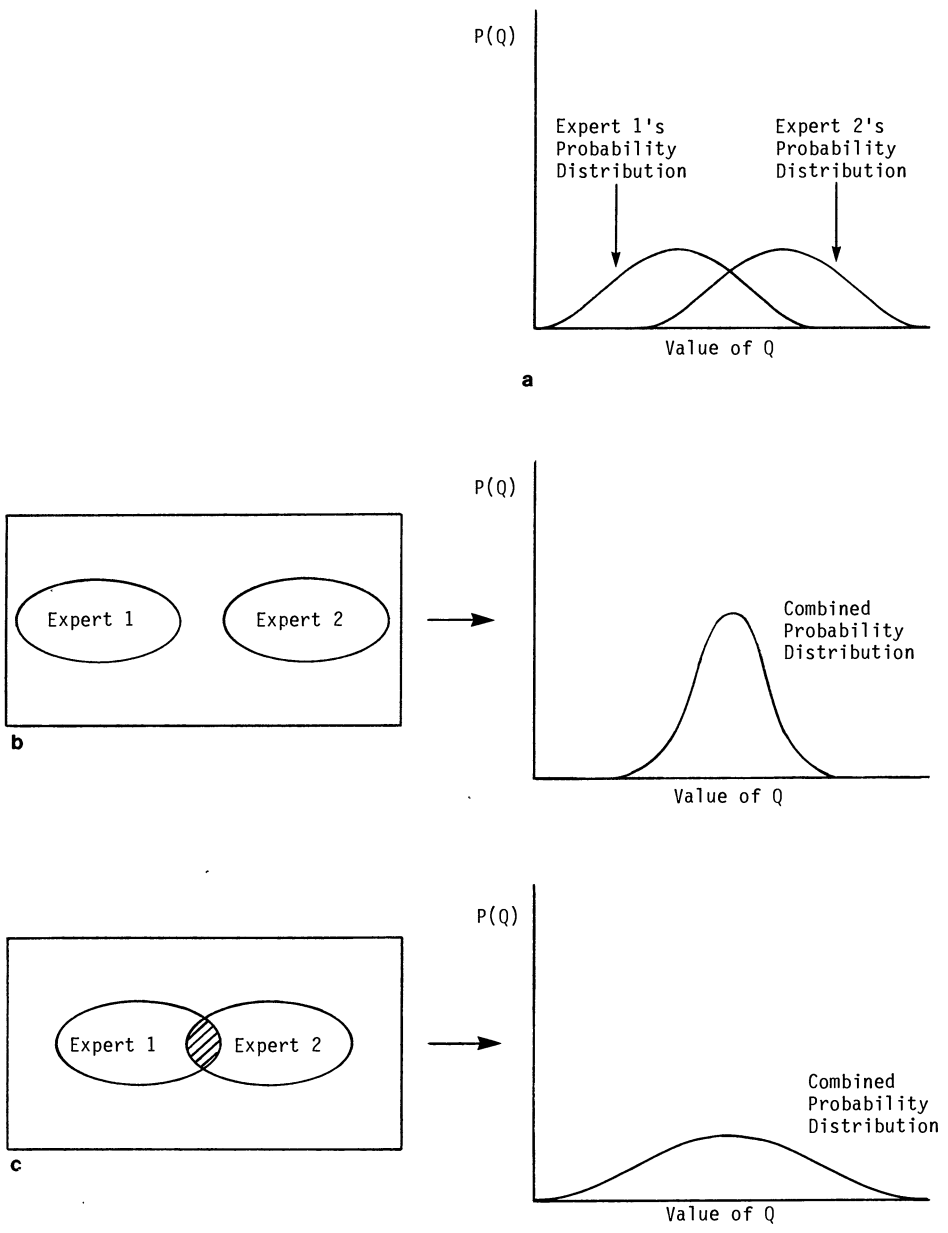

FIG. 2. (*a*) *The value of Q estimated by two experts;* (*b*) *Total independence;* (c) *Dependent experts.*

Nonindependence results from three sources: overlapping data; overlapping methodology; and direct observation and exchange of viewpoints. These relationships between expert judgments that result in nonindependence are illustrated in Fig. 1.

Overlapping data results from the fact that in most situations most experts have access to the same basic information and are basing their opinions on roughly the same body of data. Overlapping methodology may exist if experts in the field have similar academic and professional training. In this case, even if experts observe different data, they may be expected to employ many of the same modeling methods or modes of thinking. The peer review process in the scientific community also causes dependence among experts. The direct observation of other expert opinions, the presentation of public reports to the scientific community, and the open discussion of viewpoints and hypotheses will add to the overlap among expert judgments.

Fig. 2a, for example, illustrates two hypothetical probability distributions provided by two experts. The two distributions are similar and show substantial agreement on the probable value of the variable Q. If the two experts estimated these distributions independently, then the combination of their judgments would display more certainty than in either single estimate alone (Fig. 2b). However, if the two experts did not estimate these distributions independently, then the certainty of the combination may not be much greater than that of a single estimate. Fig. 2c shows that the combined distribution for variable Q has a wider range of probable values than the range that resulted from combining independent estimates.

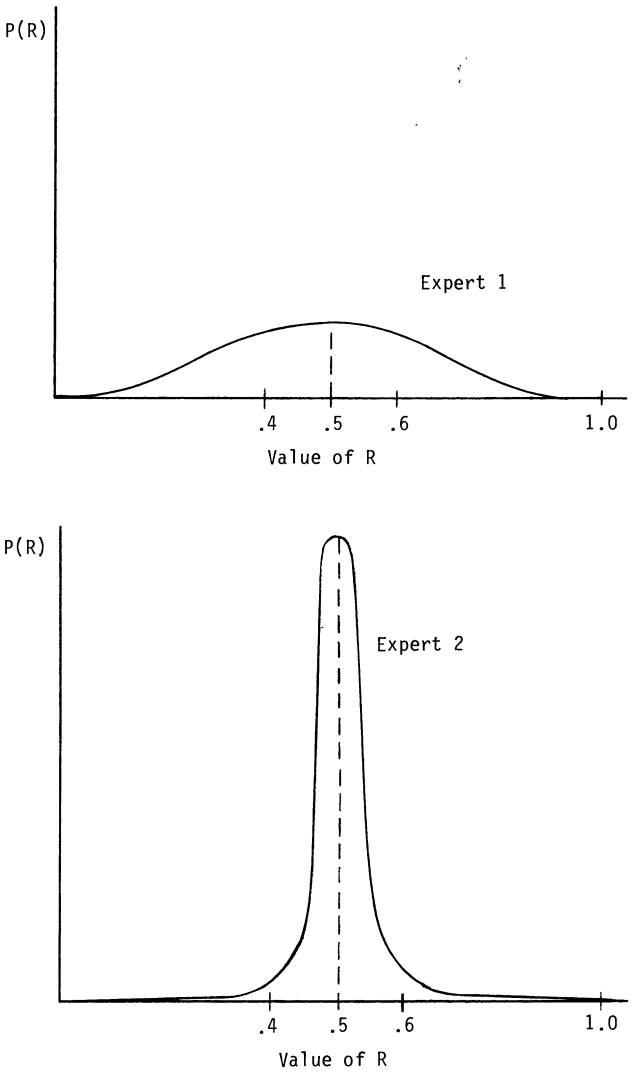

FIG. 3. *Single event probabilities versus frequency.*

Event Probabilities and Underlying Frequencies

The probability of an event describes the chance of a single occurrence. Even if this is the only statistic of interest, it may not define the appropriate level at which to combine expert judgments. Often, in combining expert probabilities, it is desirable to go one stage deeper and also consider each expert's probability distribution on the frequency of an event. This distribution more fully describes the range of possibilities of one or more occurrences over time (the event probability is simply the mean or average of the frequency distribution). *Simple event probabilities may not provide enough information to indicate how judgments should be combined.* Fig. 3, illustrates two expert probability distributions on R, which is the underlying frequency of some event. Expert 1 has estimated that the probability of this event (the mean of the distribution) is 0.5, but is very uncertain. It could easily be 0.4 or 0.6. In contrast, Expert 2 has indicated relative certainty that the probability of R is close to 0.5. The two experts agree on the estimate of 0.5, but they have very different levels of certainty about their estimates.

Common sense requires that these identical estimates of 0.5 should be processed differently when the expert judgments are combined. (I have written about this in some detail in a paper, "An Axiomatic Approach to Expert Resolution.") Although researchers disagree about how to handle the event and continuous variable cases, it is a fact that the same nonlinear weighting scheme will result in a different answer when applied to the underlying frequency distribution than when applied to the probability of an event.

Calibration

Calibration describes the difference between an expert's judgments and the observed frequency of an event. This review article mentions the issue of calibration briefly, but I think the issue is a particularly

fundamental one. It is well documented that people tend to estimate probabilities consistently too high or low. Practically, this means that their probability distributions reflect more certainty or uncertainty than they should. If experts are calibrated differently (and there is every reason to believe that experts do vary in their degree of calibration) or, worse, if there is dependence among the experts in their degree of calibration, then common sense indicates that the aggregated result should be affected.

For example, if we perform an experiment and find that one expert tends to be overconfident, and we believe that this increases the likelihood that another expert in the same field will also be overconfident, then a complicated type of dependence ensues. This is different than the physical dependence discussed above, but just as significant in determining the proper inference to make from observing a set of expert opinions. Suppose, for the sake of argument, that we believe that three experts should be equally weighted because they have the same degree of knowledge about some uncertain quantity. However, if the calibration of two of the experts is perfectly correlated but independent of the third, then equal weighting is double counting the opinions of the first two. Calibration is a bothersome issue, but lies at the heart of the expert use problem.

Level at Which Aggregation Is Performed

A key practical issue in expert combination is the level at which the aggregation should be performed. For example, suppose the variable of interest is A, and two experts agree that A depends on two more fundamental variables B and C. Should their probabilities about A be aggregated directly or should their probabilities about B and C be aggregated first and combined through probability theory to produce the composite probability distribution on A? Most would agree that the aggregation should occur at the more detailed level. However, it can be shown for most combination formulas that application at the more detailed level yields different results than application at the higher level. For example, using a linear combination formula to combine two experts' assessments of B and C is not consistent with a linear combination formula for A. The problem is compounded if the experts have different explanatory variables, as would occur, for example, if Expert 1 has a model in which A depends on B and C and Expert 2 has a model in which A depends on B and D.

This is not just a theoretical question, but a very practical issue. I was recently involved in a peer review evaluation panel of a project in which six teams of earthquake experts made probabilistic projections of the seismicity of the Eastern United States. Each of the six teams was composed of a group of individual experts. Within each team there were different models of earthquake occurrence, and among the teams there were significant differences. A fundamental practical question was whether aggregation formulas should be applied to model inputs or model outputs and whether they should be applied within each team or among teams.

3. SURVEY

The above discussion covered only some of the fundamental issues in expert resolution. Any comprehensive approach to combining expert judgments must deal with these issues either implicitly or explicitly. One useful exercise would be to take the methods discussed in this paper and measure them against how well they deal with these fundamental issues.

Another useful exercise is to apply the techniques to simplified situations in which the experts are simply observers of experimental data. For example, we can imagine experts whose sole expertise is derived from observations of thumbtack flips, balls in urns, or independent samples from normal distributions. In such cases, the problem of combining expert judgments should reduce to a standard, well understood problem in probabilistic inference. In my experience, many methods that appear to be reasonable do not produce reasonable results when applied to very simple test situations. As important, this type of test often provides insight about the underlying assumption of the method that is at fault.