Written comments on EE white paper

Roger Cooke
Resources for the Future
Feb 25, 2009

General:

Plus

- That the document appears at all
- Recognize the importance of EE in uncertainty quantification
- Recognize that subjective probability is the primary formalism for encapsulating EE
- Findings chapter 7.

Min

- Literature is very old
- Insufficient focus on applications
- No attention for performance validation
- No attention for dependence between variables
    - How to elicit dependence information
    - How to represent dependence mathematically
- Combination given very short shrift
    - No mention of scoring rules
    - No mention of EE data and expert performance
- The process of transferring expert distributions on observable quantities to parameters of a model (probabilistic inversion) is very important in applications and is missing
- Costing is unrealistic

There are more Mins than Plusses, but the Plusses are more important.

Detailed:

p.22 "The goal of an EE is to characterize, to the degree possible, each expert's beliefs (typically expressed as probabilities) about relationships, quantities, events, or parameters of interest." Rational decision requires utilities and probabilities. Utilities (values) are usually the province of stakeholder preference, but may also be elicited from experts. The techniques for this are different from those discussed in the white paper. EE is mostly employed for uncertainty quantification. I refer to quantification of uncertainty via expert judgment throughout as EJ. Uncertainty is degree of belief wrt the true state of the world. Things like "relationships" and "parameters of interest" are suitable objects for uncertainty quantification only insofar as they can be given operational meaning. We cannot use EJ to assess the specific heat of phlogiston, regardless of our interest in phlogiston. There are three plausible goals of an EJ exercise:

1. Census (survey of opinions)
2. 'Political' consensus (equilibrium of interests)
3. Rational consensus (pre-commit to method consistent with scientific principles – see documentation on classical model).

Pick one, and follow a method that is consistent with that goal

p.28 “Using mathematical methods to combine expert opinions relies on an assumption that the individual expert opinions are independent” This statement remains meaningless until a definition of “expert dependence” is given. Suppose two experts each give distributions to express their uncertainty about given uncertain quantities. How do you define the dependence or independence here? Do you envisage a correlation between distribution functions (as opposed to correlation between random variables)? How is that defined? Perhaps you mean the correlation between the experts’ mean values or median values? Well, these better be correlated, since the experts are assessing the same uncertain quantities. Perhaps you mean the *conditional* correlation of means or medians *given* the true values of the quantities in question. Perhaps you mean the correlation of the percentile realized by the true value in each experts’ distribution. Something like that makes sense, but it can only be assessed if the true values are known. The TU Delft EJ database allows expert dependence to be assessed, an article giving results is included with the submitted written material. The link between (in)dependence and combination is quite obscure. Clemen proposed a model for this, and its performance in terms of statistical accuracy was poor. (Kallen, M. J., Cooke, R.M. (2002) “Expert Aggregation with Dependence” in *Probabilistic Safety Assessment and Management* E.J. Bonano, A.L. Camp, M.J. Majors, R.A. Thompson (eds),Elsevier, 2002; 1287-1294.)

The EU-USNRC suite of expert judgment studies described in supporting material involved 69 experts spread over 9 panels, and totaling 15,422 expert-variable elicitations. To get the final results we must use distributions from each panel. If we did not combine the experts in each panel, should we consider each expert in panel A with each expert in panel B, etc? That would be 67 million combinations. (Cooke, R.M. and Kelly, G.N. Overview of Joint EU/USNRC Project on Uncertainty Analysis of Probabilistic Accident Consequence Codes (written for EE SAB))

p.29 “However in addition to their knowledge, each expert brings different biases (and experience) to the question of interest. Therefore EE practitioners must be cautious about aggregating expert judgments and presenting combined conclusions about EE results”. No one is against caution; however the slur on experts is unwarranted. How do you know that they are “biased” if you don’t know the true values of the uncertain quantities? We use expert judgment because things really are uncertain, and because they are uncertain, the experts will and should disagree. What justifies the word “biases”?

p.30 “According to Kieth (1996), combining judgments could be problematic because the methodological assumption is that the experts chosen for the elicitation represent the entire continuum of ‘truth’ with respect to the technical question” . If this means anything, it is wrong; there is no such assumption. How does the true value of an uncertain quantity become a ‘continuum’? If you know the ‘truth’ then why do EJ?

“…the fraction of experts used for the elicitation cannot be assumed to be proportional to the probability of that view or opinion being correct.” What does it mean to say that a subjective probability, or probability distribution is “correct”?

p.33 “In EE the expression of expert judgment as probabilities assumes that experts understand all alternatives so that their judgments can be compared.” The questions should be clear, and the experts should understand them – does the quote mean any more than that?

p.34 “Cooke(1990)” is Cooke(1991)??

p.35 Again the issue of “biased” Figure 3-1 fails to convey statistical accuracy; statistical accuracy can only be judged relative to a set of assessments with realizations. In statistics, an estimator of a parameter is ‘unbiased’ if, on the assumption that the samples are randomly drawn from the assumed distribution type, the expected value of the estimator equals the true value of the parameter. This interpretation is not available here, and hence only the perjorative senses of ‘biased” remain.

p.40 “Poor performance when the number of variables increases can be caused by many different factors’” How do you define and measure performance?

p.49 Table 4-1. No indication is given of the size of the EJ studies. The EU-USNRC study mentioned above involved 69 experts and 2036 uncertain quantities, and cost $4M (1990), including 15$K remuneration per expert. That study carried a peer review burden at least as great as anything at EPA. The cost figures in table 4-1 seem totally out of touch with the reality on the ground. A typical EJ study involving circa 100 variables should cost less than $100k. A recent EPA study of economic impacts of invasive species in the Great Lakes cost about half that (9 experts, 100 variables)

p.51 “For example an EE could be conducted to estimate the magnitude of a cancer slope factor (including that number’s uncertainty)”. The cardinal rule in our work is *ask experts only about outcome of possible (but infeasible) measurements/experiments.* What would you measure to estimate a CSF? Ask experts about the outcome of such measurements.

P 52 “other methods” “fuzzy methods” do not represent uncertainty. They are not a middle ground between frequentist and subjective methods. The fuzzy membership function has no operational meaning, and the combination rules are palpably wrong for representing uncertainty. At most they may claim to represent ambiguity.

p.52 sensitivity analysis – the discussion here is too brief to make sense. No reference to Saltelli?

p.54 “…there are suitable approaches to adjust data so that it is sufficient to support a specific decision”. I have no idea what this means, but it sounds like it’s against the law.

p. 62 Table 4-2. Same remark as for Table 4-1. Spending more money entails higher quality only if the possibility of wasting money is excluded.

p.63 “… alternatives to face-to-face elicitation may reduce costs” I strongly discourage that. If it is not important enough for the analyst’s time, it’s not important enough for the expert’s time. Tele-interviews are appropriate for surveys, not for EJ.

P65. “training” it is mentioned (p.74) affirming that it is good, but nowhere do you say what it is or how it is to be done. You *must not* use almanac items (how long is the Nile river) for training, it alienates experts, as it does not engage their expertise and does not emphasize the value of

informativeness. The old SRI school believed that the purpose of training was to make experts less confident. Do not do that, it will also irritate experts. Training variables must be from the experts field of expertise. You cannot do training without knowing what constitutes a 'good assessor'. That means saying how expert performance is defined and measured. Once you say that, it is very difficult to explain why you pay no attention to performance validation in an EJ study, or in the White Paper.

p.67 "clairvoyance test" . "This demands that all of the significant assumptions and conditions that could impact the expert's response are well-specified". The intention is right, but the articulation is wide of the mark. Better than saying what value an omniscient being would assign to the uncertain quantity, it is better to describe the measurement which YOU the analyst would do to determine the value of the quantity. This forces the analyst to operationalize his concepts and conditionalize them in a way appropriate to the purpose. It is very important that ALL uncertain quantities be conditionalized consistently. The notion of a clairvoyance test does not bring out this conditionalization aspect.

Example

An example from the EU-USNRC study, dry deposition velocity (ddv): It is known that ddv depends on at least 80 physical parameters ranging from the mean free path of Brownian motion to the mixing layer of the atmosphere; but it is not know *how* it depends on all these. There are many models, all of them wrong. The study in question distinguishes ddv's according to

- Chemical species
- Surface (grass, meadow, urban, peat, forest, skin)
- Aerodynamic diameter
- Windspeed

An expert is NOT asked to build a model for ddv. He is asked something like *suppose we measure the ddv of aerosols with aerodynamic diameter = 1μ on grass with windspeed 2 m/s, what will we get, please state quantiles of your distribution*?

The expert is asked to conditionalize his uncertainty in a way conformable to the model whose uncertainty is being quantified. We all know that ddv can vary by an order of magnitude according to the species of grass. The expert is not told the species of grass, rather he is told that uncertainty arising from species of grass should be 'folded into his distribution'. What about electrostatic potential, humidity, temperature, surface roughness, insolation, wind profile, etc, etc,etc? Any of these may be important, the omniscient being knows, but we don't. Hence if you really demand "that all of the significant assumptions and conditions that could impact the expert's response are well-specified" then you might as well abort the whole thing before you begin. On the other hand, it is very important, and very difficult, to maintain a consistent conditionalization across a large study. When we elicit atmospheric dispersion and wet deposition, the conditionalization must be consistent with that for dry deposition. All of this problem gets swept under the rug by talking about 'clairvoyance'.

How do expert's answer such an elicitation question? In very different ways. One may build a model, another may consult data, another may draw on his experience, another may average published models etc etc. In the end they all quantify uncertainty wrt a measurable quantity. We are combining these uncertainties, NOT combining their models.

P 68 – 84, The references for this 'how-to' part of the white paper are very old and of dubious relevance. For example the Clemen and Winkler reference (1985) concerns Winkler's normal error model, according to which expert's biases are 'corrected' by the decision maker, and expert unbiased estimates are combined as if they were point observations from a joint normal distribution – extremely unrealistic. "The practical implication is that, the more different the experts are, the more experts are needed"(p72) – don't understand that. The situation is more complicated, negative dependence is helpful, very high positive dependence with unequal variances can be helpful – but this model is not relevant for applications. A few other remarks

- The correlation coefficient ranges from -1 to 1, not 0 to 1 (note 12, p 72).
- There should be no discussion of the Delphi method without referencing (and reading) Sackman, H. (1975) *Delphi Critique, Expert Opinion, Forecasting and Group Processes,* Lexington Books, Lexington, Mass.
- The discussion of Cooke(1991)'s principles for rational consensus is garbled. "…the goal of accountability requires each judgment to be explicitly associated with a named expert" (p.81). No. Experts must be named, and the association of names and distributions must be preserved for competent peer review but need not be part of the open literature (see p. 82 of Cooke (91)). The reason is to shield experts from cherry-picking by interest groups in juridical proceedings, or from other interest conflicts. See EU Procedures Guide (p.27). The issue of use of expert names deserves more attention.
- The question of how and why to aggregate expert opinions reads like a section that didn't want to be there. Why not define the alternatives, compare their performance in practical applications?
- Measuring expert performance is necessary for performance based combination, but even without combination, good scientific practice demands that performance be defined and measured. Without this, EE does not rise to the level of good science, in my opinion. Spending lots of resources on peer review is not a substitute for performance validation.

Completely left out:

- Dependence assessment, it was a big deal in the EU-USNRC studies and has only become bigger since.
- Performance measurement, how, why, experience to date
- Probabilistic inversion: experts assess only results of possible – if impractical – observations. Pulling this uncertainty back onto the parameter space of a model was a big deal in the EU-USNRC studies and has only become bigger since.

The state-of-the-art anno 2000 can be retrieved from the EU-USNRC reports, but this is an active area and the baseline has moved; see more recent articles.