Appearing in *Making decisions with scant information - front-end decision-making in major projects, T. Williams, K. Samset and K. Sunnevag*

Obtaining distributions from Groups for Decisions under Uncertainty

Roger M. Cooke
Resources for the Future and
Department of Mathematics,
Delft University of Technology
Jan 31, 2008

Abstract: This paper considers the problem of obtaining group distributions from the standpoint of fundamental decision theory. Decision theory requires a probability distribution over possible states of the world and a value or utility function over possible outcomes of alternative actions. In a group context, this means that we should have a distribution over value functions which characterizes a population of stakeholders, and a method for combining the subjective uncertainty distributions from qualified experts. This paper focuses on both problems. Regarding combining experts' uncertainty distributions, a rich literature and body of experience is available. Techniques for capturing distributions over value functions are under development, and promising techniques are on the horizon.

1. Introduction

Decision theory provides a model for rational decision making under uncertainty. The model involves

- defining the decision space of possible actions
- quantifying uncertainty regarding the true but unknown state of the world
- quantifying the values of possible outcomes of actions

With this input, we can evaluate the expected value of each possible action and choose the action with the highest expectation.

In the realm of personal decision making, the application of this model for rational decision under uncertainty is relatively clear. In dealing with highly structured decisions involving different groups and different stakeholders, we confront two profound problems in applying the rational decision model:

1) a 'group uncertainty distribution' over possible states of the world must be defined
2) a 'group valuation' over outcomes must be defined.

Defining a group uncertainty distribution is the province of structured expert judgment. In many structured decision contexts, people are prepared to nominate a set or sets of experts, whose judgments of uncertainty would form the basis for uncertainty quantification. How exactly this should be done is subject to discussion, but *that* it should be done is relatively unproblematic (Budnitz et al, 1998, Cooke 1991A Winkler et al.1995, French 1985, Genest

and Zidek, 1986, O'Hagan 2006). The reason for this is that assessments of uncertainty tend to converge as more observations are preformed, and experts are considered to have more knowledge of 'what will happen' and thereby are better able to anticipate future observations.

With regard to valuation of outcomes the situation is fundamentally different. Values of diverse stakeholders may be conflicting. Moreover, there is nothing corresponding to 'updating valuations based on observations'. There is no such thing as a community of experts who can advise the various stakeholders on where their values should lie, or what is really best for them. The modeling of valuations for structured group decision problems should aim at finding a distribution over value functions which describes the distribution of values in the population of stakeholders.

Whereas methods for structured expert judgment have been developed over the last 20 years and have been applied extensively, the situation with regard to quantifying distributions over stakeholder valuations is relatively new. There are techniques of 'random utility models' developed in the economics literature for "discrete choice" problems, but existing methods make very restrictive assumptions and do not directly aim at estimating a distribution over utility or value functions.

This paper reviews work on structured expert judgment and indicate a new approach to quantifying stakeholder values based on a technique called probabilistic inversion. The latter technique has been applied only a few times, but would seem to hold some promise.

2. Stakeholder Preference

For a discussion of stakeholder preference modeling with probabilistic inversion, see (Neslo et al, 2008, Train 2003, Andersen at al.1996 and references therein). Only a summary account will be given here.

Suppose we wish to model stakeholder preferences for a set of N alternatives. Suppose that we can scale these utilities such that they have the same "0" and "1" where all other utilities fall between these bounds. We may then express the distribution of stakeholder utilities as a distribution over $[0, 1]^N$. Alternatively, we may score the alternatives on a set of K criteria, and seek a distribution over the criteria weights such that distribution of the weighted sum of criteria scores reflects the distribution of stakeholder preferences.

Various "discrete choice" methods or "random utility" models have been developed for this purpose. Probabilistic inversion has the advantage that no assumption is made regarding the dependences or interactions between the utilities of the various alternatives; rather the interaction structure is inferred from the stakeholder preference data with probabilistic inversion.

This is illustrated with a recent study on the valuation of threats to marine coastal ecosystems (Neslo et al 2008). 64 stakeholders were presented with 30 threat scenarios and asked to rank the top five. This produced a set of probabilities that each of the 30 scenarios could be ranked in position 1,…5. The exercise was designed such that some rankings were inconsistent with

the multi-criteria model in the sense that some scenarios were dominated by others on each criterion. The relatively low probabilities for inconsistent rankings gave a rough validation of the multi criteria model.

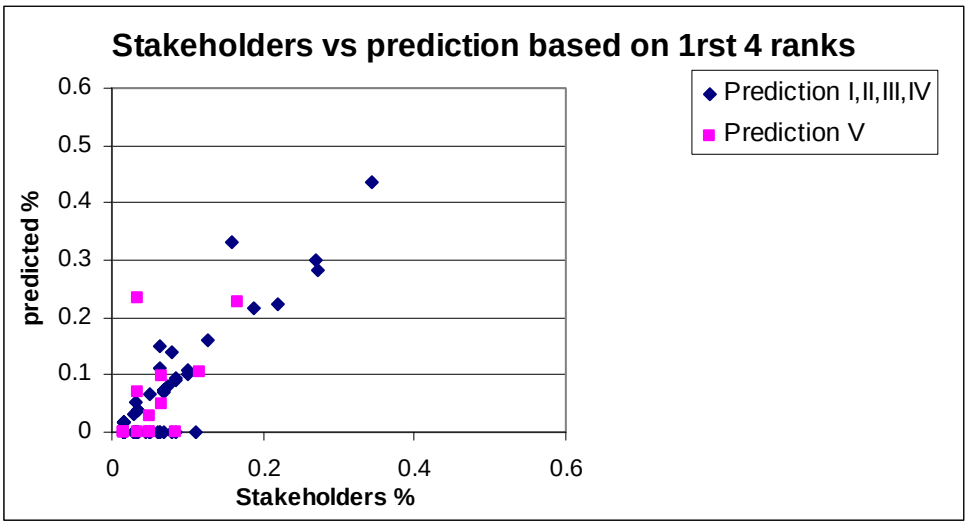
To find a distribution over criteria weights reflecting the discrete choice preference data, we start with a non-informative distribution over the criteria weights, and adapt this distribution such that:

- the probability of drawing a vector of weights which ranks scenario *n* in position *j* is as close as possible to the percentage of stakeholders ranking scenario *n* in position *j; 1* $\leq n \leq 30$*; 1* $\leq j \leq 4$.
- the resulting distribution is minimally informative with respect to the initial non-informative distribution.

The model was fit on the first 4 rankings, and used to predict the 5th rankings. Figure 1 compares the predicted and observed percentages of rankings. The model is first used to "retrodict" or "recover" the first four rankings, These are the data actually used to fit the model, so this comparison is a check of model fit rather than model prediction. Using the model, we can predict the percentages of experts ranking the various scenarios in the 5th position. These percentages were not used in fitting the model and test the ability of the model to predict preferences of the population of stakeholders. Of course, we should hope that the predictions and retrodictions show similar agreement with the observed rankings.

The retrodictions are shown as diamonds and the predictions are shown as squares. The percentages along the horizontal axis correspond to rankings that were inconsistent with the multi-criteria model. Figure 1 shows that the weight distribution predicts stakeholder preferences reasonably well; except for the inconsistent rankings, the predicted percentages of stakeholders assigning a given rank to a given scenario agree reasonably with the observed percentages. Moreover the pattern of agreement for retrodictions and predictions is similar. This justifies us in using this model to predict other unobserved rankings of the population. Thus, if new scenarios need to be evaluated, we need not re-convene the 64 stakeholders and repeat whole exercise; instead we can use the model to assess the new scenarios together with the original 30 scenarios.

Figure 1: Predicted and observed percentages of stakeholder rankings in marine coastal ecosystem study



3. Structured expert judgment

Expert judgment is sought when substantial scientific uncertainty impacts on a decision process. Because there is uncertainty, the experts themselves are not certain and hence will typically not agree. Informally soliciting expert advice is not new. *Structured* expert judgment refers to an attempt to subject this process to transparent methodological rules, with the goal of treating expert judgments as scientific data in a formal decision process. Standard sources include Cooke (1991), European Procedures Guide on Expert Judgment (Cooke and Goossens, 2000) and O'Hagan (2006). A recent special issue of *Reliability Engineering and System Safety* (2008) covers standard as well as emerging techniques. This section is based in part on (Cooke and Goossens 2008).

The process by which experts come to agree is the scientific method itself. Structured expert judgment cannot pre-empt this role and therefore cannot have expert agreement as its goal. We may broadly distinguish three different goals to which a structured judgment method may aspire:

- Census
- Political consensus
- Rational consensus

A study aiming at *census* will simply try to survey the distribution of views across an expert community. An illustration of this goal is found in the *Nuclear Regulatory Commission's Recommendations for Probabilistic Seismic Hazard Analysis: Guidance on Uncertainty and Use of Experts:*

> "To represent the overall community, if we wish to treat the outlier's opinion as equally credible to the other panelists, we might properly

> assign a weight (in a panel of 5 experts) of 1/100 to his or her position, not 1/5" (NUREG/CR-6372, p.36)

The goal of "representing the overall community" may in this view lead to a differential weighting of experts' views according to how representative they are of other experts. A similar goal is articulated in (Winkler et al 1995). The philosophical underpinnings of this approach are elaborated in (Budnitz et al 1998). Expert agreement on the representation of the overall community is the weakest, and most accessible, type of consensus to which a study may aspire. Other types of consensus are

- agreement on a distribution to represent a group,
- agreement on a distribution and
- agreement on a number

Political consensus refers to a process in which experts are assigned weights according to the interests or stakeholders they represent. In practice, an equal number of experts from different stakeholder groups would be placed in an expert panel and given equal weight in this panel. In this ways the different groups are included equally in the resulting representation of uncertainty.

Rational consensus refers to a group decision process. The group agrees on a method according to which a representation of uncertainty will be generated for the purposes for which the panel was convened, without knowing the result of this method. It is not required that each individual member adopt this result as his/her personal degree of belief. This is a form of agreement on a distribution to represent a group. To be rational this method must comply with necessary conditions devolving from the general scientific method. Cooke (1991) formulates necessary conditions or principles which any method warranting the predicate "scientific" should satisfy:

- **Scrutability/accountability:** All data, including experts' names and assessments, and all processing tools are open to peer review and results must be reproducible by competent reviewers.
- **Empirical control:** Quantitative expert assessments are subjected to empirical quality controls.
- **Neutrality:** The method for combining/evaluating expert opinion should encourage experts to state their true opinions, and must not bias results.
- **Fairness:** Experts are not pre-judged, prior to processing the results of their assessments.

Thus, a method is proposed which satisfies these conditions and to which the parties pre-commit. The method is applied and after the result of the method is obtained, parties wishing to withdraw from the consensus incur a burden of proof. They must demonstrate that some heretofore unmentioned necessary condition for rational consensus has been violated. Absent that, their dissent is not "rational". Of course any party may withdraw from the consensus because the result is hostile to his or her interests. Since such withdrawl is based on interest rather than arguments, this is not rational dissent and does not endanger rational consensus.

The requirement of empirical control will strike some as peculiar in this context. How can there be empirical control with regard to expert subjective probabilities? To answer this question we must reflect on the question 'when is a problem an expert judgment problem?' We would not have recourse to expert judgment to determine the speed of light in a vacuum. This is physically measurable and has been measured to everyone's satisfaction. Any experts we queried would give the same answer. Neither do we consult expert judgment to determine the proclivities of a god. There are no experts in the operative sense of the word for this issue. A problem is susceptible for expert judgment only if there is relevant scientific expertise. This entails that there are theories *and* measurements relevant to the issues at hand, but that the quantities of interest themselves cannot be measured in practice. For example, toxicity of a substance for humans is measurable in principle, but is not measured for obvious reasons. However, there are toxicity measurements for other species which might be relevant to the question of toxicity in humans. Other examples are given in section 6.

If a problem is an expert judgment problem, then necessarily there will be relevant experiments or measurements. Questions regarding such experiments can be used to implement empirical control. Studies indicate that performance on so-called almanac questions does not predict performance on variables in an expert's field of expertise. (Cooke, Mendel and Thijs., 1988). The key question regarding seed variables is this: Is performance on seed variables judged relevant for performance on the variables of interest? For example, should an expert who gave very over-confident off-mark assessments on the variables for which we knew the true values be equally influential on the variables of interest as an expert who gave highly informative and statistically accurate assessments? That is indeed the choice that often confronts a problem owner after the results of an expert judgment study are in. If seed variables in this sense cannot be found, then rational consensus is not a feasible goal and the analyst should fall back on one of the other goals.

The above definition of “rational consensus" for group decision processes is evidently on a very high level of generality. Much work has gone into translating this into a workable procedure which gives good results in practice. This workable procedure is embodied in the "classical model" of (Cooke 1991) described in the following section.

Before going into details it is appropriate to say something about Bayesian approaches. Since expert uncertainty concerns experts' subjective probabilities many people believe that expert judgment should be approached from the Bayesian paradigm. This paradigm, recall, is based on the representation of preference of a rational individual in terms of maximal expected utility. If a Bayesian is given experts' assessments on variables of interest and on relevant seed variables, then (s)he may update his/her prior on the variables of interest by conditionalizing on the given information. This requires that the Bayesian formulates his/her joint distribution over

- the variables of interest
- the seed variables
- the experts' distributions over the seed variables and the variables of interest.

Issues that arise in building such a model are discussed in Cooke (1991). Suffice to say here that a group or rational individuals is not itself a rational individual, and group decision problems are notoriously resistant to the Bayesian paradigm.

4. The classical model

The above principles have been operationalized in the so called "classical model", a performance based linear pooling or weighted averaging model (Goossens Cooke, and Kraan 1998, Cooke 1991). This model has been applied in 45 contracted studies, involving upwards of 67,000 individual elicitations. An overview of the applications is presented in Table 1.

Weights for a performance based combination of expert distributions are derived from experts' calibration and information scores, as measured on seed variables. Seed variables serve a threefold purpose:

(i) to quantify experts' performance as subjective probability assessors,
(ii) to enable performance-optimized combinations of expert distributions, and
(iii) to evaluate and hopefully validate the combination of expert judgments.

The name "classical model" derives from an analogy between calibration measurement and classical statistical hypothesis testing. It contrasts with various Bayesian models in that it does not assume prior information.

Table 1 Summary of applications per sector

Sector	# of experts	# of variables	# of elicitations
Nuclear applications	98	2,203	20,461
Chemical ind. & gas industry	56	403	4,491
Groundwater / water pollution / dike ring / barriers	49	212	3,714
Aerospace sector / space debris /aviation	51	161	1,149
Occupational sector: ladders / buildings (thermal physics)	13	70	800
Health: bovine / chicken (*Campylobacter*) / SARS	46	240	2,979
Banking: options / rent / operational risk	24	119	4,328
Volcanoes / dams	231	673	29079
Rest group	19	56	762
In total	***521***	***3688***	***67001***

The performance based weights use two quantitative measures of performance, *calibration* and *information*. Loosely, calibration measures the statistical likelihood that a set of experimental results correspond, in a statistical sense, with the expert's assessments. Information measures the degree to which a distribution is concentrated.

These measures can be implemented for both discrete and quantile elicitation formats. In the discrete format, experts are presented with uncertain events and perform their elicitation by assigning each event to one of several pre-defined probability bins, typically 10%, 20%,…90%. In the quantile format, experts are presented an uncertain quantity taking values in a continuous range, and they give pre-defined quantiles, or percentiles, of the subjective uncertainty distribution, typically 5%, 50% and 95%. The quantile format has distinct advantages over the discrete format, and all the studies reported below use this format. In five studies the 25% and 75% quantiles were also elicited. To simplify the exposition we assume that the 5%, 50% and 95% values were elicited.

Calibration
For each quantity, each expert divides the range into 4 inter-quantile intervals for which his/her probabilities are known, namely $p_1 = 0.05$: less than or equal to the 5% value, $p_2 = 0.45$: greater than the 5% value and less than or equal to the 50% value, etc.

If N quantities are assessed, each expert may be regarded as a statistical hypothesis, namely that each realization falls in one of the four inter-quantile intervals with probability vector

$$p = (0.05, 0.45, 0.45, 0.05).$$

Suppose we have realizations $x_1,...x_N$ of these quantities. We may then form the sample distribution of the expert's inter quantile intervals as:

$$s_1(e) = \#\{ i \mid x_i \leq 5\% \text{ quantile}\}/N$$
$$s_2(e) = \#\{ i \mid 5\% \text{ quantile} < x_i \leq 50\% \text{ quantile}\}/N$$
$$s_3(e) = \#\{ i \mid 50\% \text{ quantile} < x_i \leq 95\% \text{ quantile}\}/N$$
$$s_4(e) = \#\{ i \mid 95\% \text{ quantile} < x_i \}/N$$
$$s(e) = (s_1,...s_4)$$

Note that the sample distribution depends on the expert *e*. If the realizations are indeed drawn independently from a distribution with quantiles as stated by the expert then the quantity

$$2NI(s(e) \mid p) = 2N \sum_{i=1..4} s_i \ln(s_i / p_i) \qquad (1)$$

is asymptotically distributed as a chi-square variable with 3 degrees of freedom. This is the so-called likelihood ratio statistic, and I(s | p) is the relative information of distribution *s* with respect to *p*. If we extract the leading term of the logarithm we obtain the familiar chi-square test statistic for goodness of fit. There are advantages in using the form in (1) (Cooke 1991).

If after a few realizations the expert were to see that all realization fell outside his 90% central confidence intervals, he might conclude that these intervals were too narrow and might broaden them on subsequent assessments. This means that for this expert the uncertainty distributions are not independent, and he learns from the realizations. Expert learning is not a goal of an expert judgment study and his joint distribution is not elicited. Rather, the decision maker wants experts who do not need to learn from the elicitation. Hence the decision maker scores expert *e* as the statistical likelihood of the hypothesis

H_e: "the inter quantile interval containing the true value for each variable is drawn independently from probability vector p."

A simple test for this hypothesis uses the test statistic (1), and the likelihood, or p-value, or calibration score of this hypothesis, is:

$$Calibration\ score(e) = p\text{-}value = Prob\{\ 2NI(s(e) \mid p) \geq\ r \mid H_e\}$$

where r is the value of (1) based on the observed values $x_1,\ldots x_N$. It is the probability under hypothesis H_e that a deviation at least as great as r should be observed on N realizations if H_e were true. Calibration scores are absolute and can be compared across studies. However, before doing so, it is appropriate to equalize the power of the different hypothesis tests by equalizing the effective number of realizations. To compare scores on two data sets with N and N' *realizations*, we simply use the minimum of N and N' in (1), without changing the sample distribution s. In some cases involving multiple realizations of one and the same assessment, the effective number of seed variables is based on the number of assessments and not the number of realizations.

Although the calibration score uses the language of simple hypothesis testing, it must be emphasized that we are not rejecting expert-hypotheses; rather we are using this language to measure the degree to which the data supports the hypothesis that the expert's probabilities are accurate. Low scores, near zero, mean that it is unlikely that the expert's probabilities are correct.

Information

The second scoring variable is information. Loosely, the information in a distribution is the degree to which the distribution is concentrated. Information cannot be measured absolutely, but only with respect to a background measure. Being concentrated or "spread out" is measured relative to some other distribution. Commonly, the uniform and log-uniform background measures are used.

Measuring information requires associating a density with each quantile assessment of each expert. To do this, we use the unique density that complies with the experts' quantiles and is minimally informative with respect to the background measure. This density can easily be found with the method of Lagrange multipliers. For a uniform background measure, the density is constant between the assessed quantiles, and is such that the total mass between the quantiles agrees with p. The background measure is not elicited from experts as indeed it must be the same for all experts; instead it is chosen by the analyst.

The uniform and log-uniform background measures require an *intrinsic range* on which these measures are concentrated. The classical model implements the so-called k% overshoot rule: for each item we consider the smallest interval $I = [L, U]$ containing all the assessed quantiles of all experts and the realization, if known. This interval is extended to

$$I^* = [L^*, U^*];\ L^* = L - k(U\text{-}L)/100;\ \ U^* = U + k(U\text{-}L)/100.$$

The value of k is chosen by the analyst. A large value of *k* tends to make all experts look quite informative, and tends to suppress the relative differences in information scores. The information score of expert *e* on assessments for uncertain quantities 1…N is

$$\textit{Information Score}(e) = \textit{Average Relative information wrt Background} = (1/N) \sum_{i=1..N} I(f_{e,i} \mid g_i)$$

where g_i is the background density for variable i and $f_{e,i}$ is expert e's density for item i. This is proportional to the relative information of the expert's joint distribution given the background, under the assumption that the variables are independent. As with calibration, the assumption of independence here reflects a desideratum of the decision maker and not an elicited feature of the expert's joint distribution. The information score does not depend on the realizations. An expert can give himself a high information score by choosing his quantiles very close together.

Evidently, the information score of *e* depends on the intrinsic range and on the assessments of the other experts. Hence, information scores cannot be compared across studies.

Of course, other measures of concentrated-ness could be contemplated. The above information score is chosen because it is

- familiar
- tail insensitive
- scale invariant
- slow

The latter property means that relative information is a slow function; large changes in the expert assessments produce only modest changes in the information score. This contrasts with the likelihood function in the calibration score, which is a very fast function. This causes the product of calibration and information to be driven by the calibration score.

Decision maker

A combination of expert assessments is called a "decision maker" (DM). All decision makers discussed here are examples of linear pooling. For a discussion of pro's and con's of the linear pool see (French, 1985, Genest and Zidek, 1986, Cooke 1991). The classical model is essentially a method for deriving weights in a linear pool. "Good expertise" corresponds to good calibration (high statistical likelihood, high p-value) and high information. We want weights which reward good expertise and which pass these virtues on to the decision maker.

The reward aspect of weights is very important. We could simply solve the following optimization problem: find a set of weights such that the linear pool under these weights maximizes the product of calibration and information. Solving this problem on real data, we have found that the weights do not generally reflect the performance of the individual experts.

As we do not want an expert's influence on the decision maker to appear haphazard, and we do not want to encourage experts to game the system by tilting their assessments to achieve a

desired outcome, we must impose a strictly scoring rule constraint on the weighing scheme. Roughly, this means that an expert achieves his maximal expected weight by and only by stating assessments in conformity with his/her true beliefs.

Consider the following score for expert *e*:

$$w_\alpha(e) = 1_\alpha(\text{calibration score}) \times \text{calibration score}(e) \times \text{information score}(e) \quad (2)$$

where $1_\alpha(x) = 0$ if $x < \alpha$ and $1_\alpha(x) = 1$ otherwise. Cooke (1991) shows that (2) is an asymptotically strictly proper scoring rule for average probabilities. This means the following: suppose an expert has given his quantile assessments for a large number of variables and subsequently learns that his judgments will be scored and combined according the classical model. If (s)he were then given the opportunity to change the quantile values (e.g. the numbers 5%, 50% or 95%) in order to maximize the expected weight, the expert would choose values corresponding to his/her true beliefs. Note that this type of scoring rule scores a set of assessments on the basis of a set of realizations. Scoring rules for individual variables were found unsuitable for purposes of weighting, for which discussion we refer to (Cooke 1991).

The scoring rule constraint requires the term $1_\alpha(\text{calibration score})$, but does not say what value of α we should choose. Therefore, we choose α so as to maximize the combined score of the resulting decision maker. Let $DM_\alpha(i)$ be the result of linear pooling for item *i* with weights proportional to (2):

$$DM_\alpha(i) = \sum_{e=1,..E} w_\alpha(e)\, f_{e,i} \;/\; \sum_{e=1,..E} w_\alpha(e) \quad (3)$$

The *global weight DM* is DM_{α^*} where α* maximizes

$$\text{calibration score}(DM_\alpha) \times \text{information score}(DM_\alpha). \quad (4)$$

This weight is termed global because the information score is based on all the assessed seed items

A variation on this scheme allows a different set of weights to be used for each time. This is accomplished by using information scores for each item rather than the average information score:

$$w_\alpha(e,i) = 1_\alpha(\text{calibration score}) \times \text{calibration score}(e) \times I(f_{e,i} \mid g_i) \quad (5)$$

For each α we define the Item weight DM_α for item *i* as

$$IDM_\alpha(i) = \sum_{e=1,..E} w_\alpha(e,i)\, f_{e,i} \;/\; \sum_{e=1,..E} w_\alpha(e,i) \quad (6)$$

The *item weight DM* is IDM_{α^*} where α* maximizes

$$\textit{calibration score}(IDM_a) \times \textit{information score}(IDM_a). \quad (7)$$

Item weights are potentially more attractive as they allow an expert to up- or down- weight him/herself for individual items according to how much (s)he feels (s)he knows about that item. "knowing less" means choosing quantiles further apart and lowering the information score for that item. Of course, good performance of item weights requires that experts can perform this up-down weighting successfully. Anecdotal evidence suggests that item weights improve over global weights as the experts receive more training in probabilistic assessment. Both item and global weights can be pithily described as optimal weights under a strictly proper scoring rule constraint. In both global and item weights calibration dominates over information, information serves to modulate between more or less equally well calibrated experts.

Since any combination of expert distributions yields assessments for the seed variables, any combination can be evaluated on the seed variables. In particular, we can compute the calibration and the information of any proposed decision maker. We should hope that the *performance weighted decision maker (PWDM)* would perform better than the result of simple averaging, called the Equal weight Decision Maker *EWDM*, and we should also hope that the proposed DM is not worse than the best expert in the panel.

In the classical model calibration and information are combined to yield an overall or combined score with the following properties:

1. Individual expert assessments, realizations and scores are published. This enables any reviewer to check the application of the method, in compliance with the principle of **accountability / scrutability.**
2. Performance is measured and hopefully validated, in compliance with the principle of **empirical control**. An expert's weight is determined by performance.
3. The score is a long run proper scoring rule for average probabilities, in compliance with the principle of **neutrality**.
4. Experts are treated equally, prior to the performance measurement, in compliance with the principle of **fairness**.

Expert names and qualifications are part of the published documentation of every expert judgment study in the data base; however, they are not associated with assessments in the open literature. The experts reasoning is always recorded and sometimes published as expert rationales.

There is no mathematical theorem that either item weights or global weights out-perform equal weighting or out-perform the best expert. It is not difficult to construct artificial examples where this is not the case. Performance of these weighting schemes is a matter of experience. In practice, global weights are used unless item weights perform markedly better. Of course there may be other ways of defining weights that perform better, and indeed there might be better performance measures. Good performance on one individual data set is not convincing. What is convincing is good performance on a large diverse data set, such as the

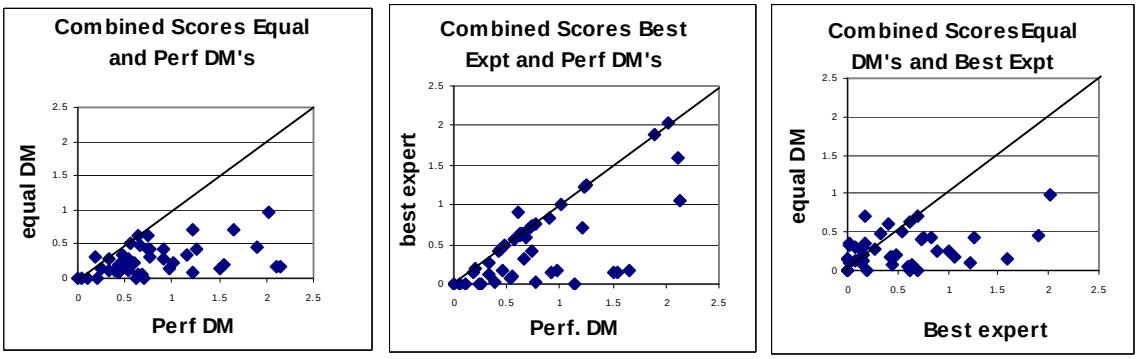
TU Delft expert judgment data base. In practice a method should be easy to apply, easy to explain, should do better than equal weighting and should never do something ridiculous.

5. Applications of the classical model

45 expert panels involving seed variables have been performed to date[1]. Because most of these studies were performed by or in collaboration with the TU Delft, it is possible to retrieve relevant details of these studies, and to compare performance of performance based and equal weight combination schemes.

The combined scores of EWDM, PWDM and Best Expert are compared pair wise in Figure 2. Figure 3 compares the calibration (p-values) and information scores of the EWDM, the PWDM and the best expert.

Figure 2. Combined scores of EWDM, PWDM and Best Expert



In 15 of 45 cases the PWDM was the best expert, that is, one expert received weight one. In 27 cases the combined score of the PWDM was strictly better than both the EWDM and the best expert. In one case the EWDM performed best, and in two cases the best expert out-performed both equal weights and performance based weights.

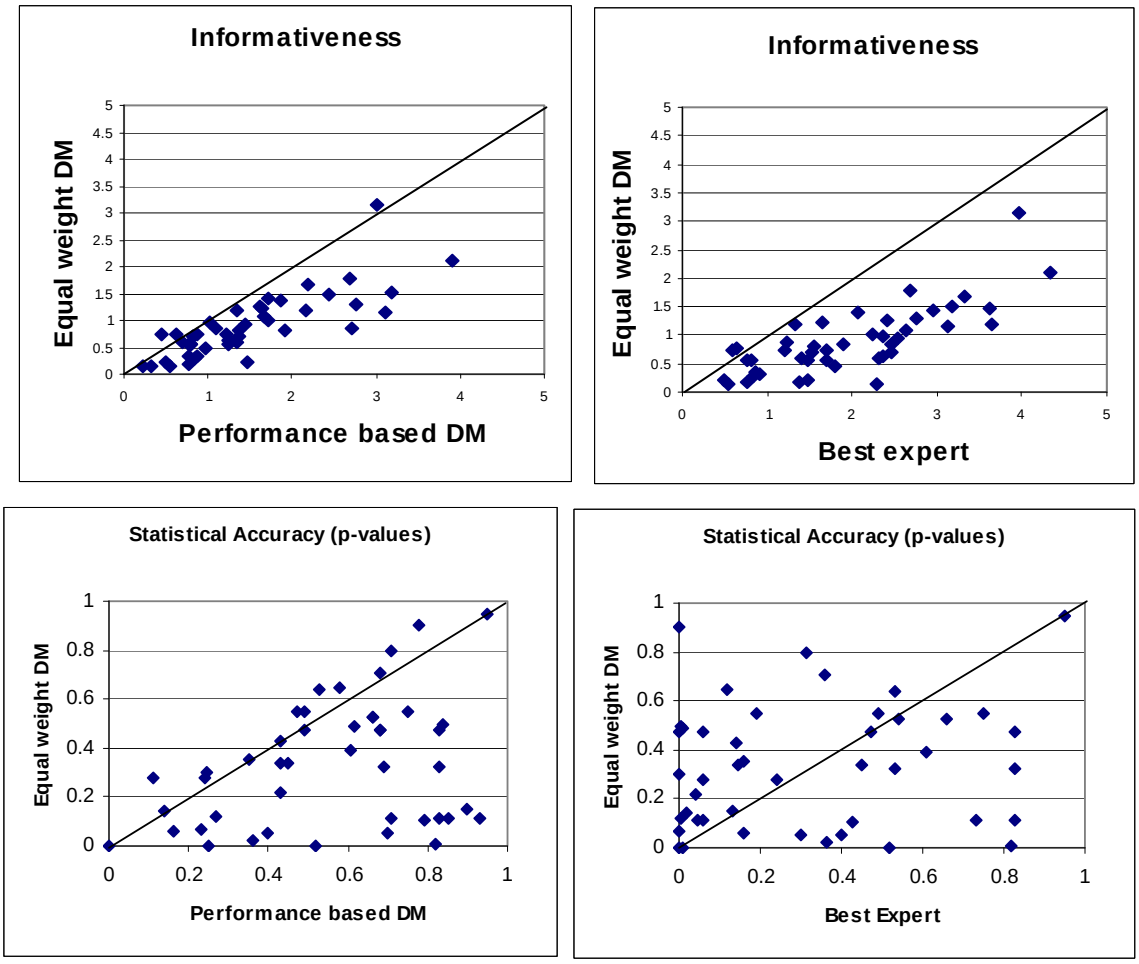
The EWDM is better calibrated than the best expert in 25 of the 45 cases, but in only 2 cases more informative. In 18 cases the combined score of the EWDM is better than that of the best expert. In 12 of the 45 cases the calibration of the best expert is less than or equal to 0.05; for the EWDM this happened in 7 cases (15%).

The motivation for performance based weighting above equal weighting speaks for itself from this data. Sometimes the difference is marginal but sometimes it is quite significant. Most often the EWDM is slightly less well calibrated and significantly less informative, but sometimes the calibration of the EWDM is quite poor. Finally we remark that the experts overwhelmingly have supported the idea of performance measurement. This sometimes comes as a surprise for people from the social sciences, but not for natural scientists. The essential

[1] These results are obtained with the EXCALIBUR software, available from http://delta.am.ewi.tudelft.nl/risk/. The windows version upgraded chi square and information computational routines, and this may cause differences with the older DOS version, particularly with regard to very low calibration scores.

point is that the performance measures are objective and fully transparent. It is impossible to tweak these measures for extra-scientific expediency.

Figure 3 Calibration (p-values) and Information scores of EWDM, PWDM and best expert



6. Seed variables, variables of interest and robustness

A recurring question is the degree to which performance on seed variables predicts performance on the variables of interest. Forecasting techniques always do better on data used to initialize the models than on fresh data. Might that not be the case here as well? Obviously, we have recourse to expert judgment *because* we cannot observe the variables of interest, so this question is likely to be with us for some time. Experts' information scores *can* be computed for the variables of interest and compared with the seed variables (see below). More difficult is the question whether calibration differences in experts and DMs "persist" outside the set of seed variables. Questions related to this are:

1. Are the differences in experts' calibration scores due to chance fluctuations?
2. Is an expert's ability to give informative and well calibrated assessments persistent in time, dependent on training, seniority, or related to other psycho-social variables, etc?

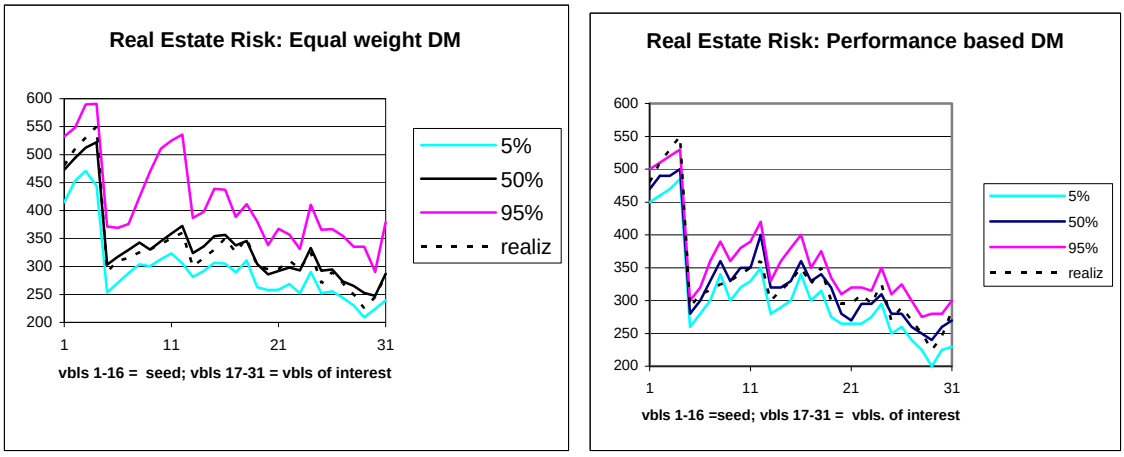
There has been much published and speculated on these questions, and the issue cannot be reviewed, let alone resolved here (see however Lin and Bier 2008). If differences in experts' performance did *not* persist beyond the seed variables, then that would certainly cast a long shadow over performance based combination. If, on the other hand, there are real and reasonably persistent differences in expert performance, then it is not implausible that a performance based combination could systematically do 'better than average'.

Closely related is the question of robustness: to what extent would the results change if different experts or different seed variables had been used. This last question can be addressed, if not laid to rest, by removing seed variables and experts one at a time and re-computing the decision maker. We discuss one example where the variables of interest were later observed, and performance wioth respect to seed variables could be compared.

Real estate risk

In Qing (2002) the seed variables were prime office rent indices for large Dutch cities, published quarterly (variables 1 through 16). The variables of interest were rents of the actual properties managed by the investment firm. After one year, the realized rents were retrieved and compared with the predictions. The results for the equal and performance DM are shown below. Evidently, for both PWDM and EWDM, the performance on seed variables and variables of interest is quite similar. Note that the EWDM has larger 90% confidence bands.

Figure 4: Performance versus equal weight combinations for Real estate risk, seed variables and variables of interest



Out – of - Sample validation?

In his review of (Cooke and Goossens 2008), Clemen (Clemen 2008) raised the important question: does the performance of the performance-weighted DM (PWDM) persist beyond the set of seed variables. Clemen believes that there is no significant difference between the PWDM (PWDM) and the EWDM (EWDM) outside the variables on which PWDM has been constructed.

As noted above PWDM does use optimization to remove a degree of freedom in the definition of the classical model. On every study we routinely perform robustness analysis by removing seed variables (and experts) one at a time and re-computing PWDM It is not uncommon to see the calibration scores of PWDM fluctuate by a factor 2 or 3 on ten seed variables.

Out-of-sample validation involves basing PWDM on an initial set of seed variables, then using this PWDM on *other* variables and comparing performance with EWDM on these other variables. This corresponds to the way PWDM is actually used. We can do this by splitting the set of seed variables into two halves, initializing the model on one half and comparing performance on the other half. Of course, this requires a relatively large number of seed variables. 13 studies with at least 16 seed variables are available. Dividing the seed variables in half gives two validation runs, using the first half to predict the second and conversely. Note that the variables on which the PWDM is initialized in these two runs are disjoint. The item weight PWDM could not be computed without writing new code, so the choice of item versus global weights is denied the PWDM on this exercise.

There are 26 out of sample studies (two for each of the 13 studies). In 20 of the 26 studies the out-of-sample PWDM out-performs EWDM. The probability of seeing 20 or more "successes" on 26 trials if PWDM were no better than EWDM is 0.0012.

Clemen reports results on 14 validation studies that are somewhat more pessimistic (9 "success" on 14 trials). His method involves removing seed variables singly, computing PWDM on the remaining seeds, and using this PWDM to predict the eliminated seed. On a study with 10 seed variables there are thus 10 *different* PWDM's. Each pair of the 10 DM's share 8 common seeds. The criteria for selecting the 14 studies are not specified. It is difficult to see how all these factors would affect the results. Perhaps the following reasoning partially explains Clemen's less optimistic result: With a small number of seeds, removing one seed favors experts who assessed *that* seed badly and hurts experts who assessed *that* seed well, thus tilting the PWDM toward a bad assessment of *that* seed. This happens on *every* seed thus cumulating the adverse effect on PWDM. This does not happen when *one* PWDM predicts the entire out-of-sample set of seeds. In any case, Clemen's method is not the same as picking *one* PWDM and comparing it on new observations with the EWDM.

7. Conclusions

Structured expert judgment has become an applicable tool in quantitative studies when input from data or experiments is lacking. The expert judgment data base provides a resource for evaluating the performance of various expert judgement combination schemes. This has enabled us to demonstrate that performance based expert judgement models are statistically superior to simple averaging. This conclusion is based on extensive experience over a wide range of studies from diverse areas.

Modeling stakeholder preferences is less well articulated, but important ideas emerge from the field of discrete choice. The key issue for further progress is to develop tools for validating proposed models on the basis of observed preference data. Lack of external validation has

plagued many of the multi criteria approaches. It was argued that probabilistic inversion techniques may suggest ways forward in this regard.;

References

1. Anderson, S.P., A. de Palma and J-F Thissen, 1996 Discrete Choice Theory of Product Differentiation, MIT Press, Cambridge.
2. Budnitz, R.J., Apostolakis, G., Boore, D.M., Cluff, L.S., Coppersmith, K.J., Cornel, C.A., and Morris, P.A. (1998) "Use of technical expert panels: applications to probabilistic seismic hazard analysis", *Risk Analysis*, vol 18 no. 4 463-469.
3. Chou, D., Kurowicka, D. and Cooke, R.M. *(2006)* "Techniques for generic probabilistic inversion" appearing in *Comp. Stat and Data Analysis.*
4. Clement, R. T. "Comments" in Special issue on expert judgment Reliability Engineering & System Safety, Available online 12 March 2007.
   Cooke R. M. and Misiewicz, J. (2007) Discrete Choice with Probabilistic Inversion: Application to energy policy choice and wiring failure, *Mathematical Methods in Reliability.*
5. Cooke, R.M. (1991A) Experts in Uncertainty, Oxford University Press, Oxford.
6. Cooke, R.M. Goossens, L.J.H., (2000) *Procedures guide for structured expert judgment* Project report EUR 18820EN, Nuclear science and technology, specific programme Nuclear fission safety 1994-98, Report to: European Commission. Luxembourg, Euratom. Also in *Radiation Protection Dosimetry* Vol. 90 No. 3.2000, 64 7, pp 303-311.
7. Cooke, R.M., ElSaadany, S., Xinzheng Huang, X. (2008) *On the Performance of Social Network and Likelihood Based Expert Weighting Schemes,* Special issue on expert judgment Reliability Engineering & System Safety, 645-756, Available online 12 March 2007.
8. Cooke, R.M., Goossens, L.H.J. (2008) *TU Delft Expert Judgment Data Base,* Special issue on expert judgment Reliability Engineering & System Safety (657-674), Available online 12 March 2007.
9. Cooke, R.M., Mendel, M., Thijs, W. (1988), "Calibration and Information in Expert Resolution". *Automatica*, 24, 1, 8-87-94, 1988.
10. Cooke, R.M.,(ed) . (2008) *Special issue on expert judgment* Reliability Engineering & System Safety, Available online 12 March 2007.
11. Csiszar I. (1975) I-divergence geometry of probability distributions and minimization problems. *Ann. of Probab.*, 3:146-158.
12. Deming, W.E., and Stephan, F.F. (1944). On a least squares adjustment to sample frequency tables when the expected marginal totals are known, Ann Math. Statist.
13. French, S. (1985) "Group consensus probability distributions: a critical survey"in J.M. Bernardo, M.H. De Groot, D.V. Lindley and A.F.M. Smith (eds) *Bayesian Statistics*, Elsevier North Holland, pp 182-201.
14. Genest, C. and Zidek, J. (1986), "Combining probability distributions: a critique and an annotated bibliography" *Statistical Science* vol. 1, no.1 pp 114-1490, 1986.

15. Goossens LHJ, Cooke RM, and Kraan, BCP, (1996) *Evaluation of weighting schemes for expert judgment studies***,** Final report prepared under contract Grant No. Sub 94-FIS-040 for the Commission of the European Communities, Directorate General for Science, Research and Development XII-F-6, Delft University of Technology, Delft, the Netherlands.
16. Goossens LHJ, Cooke RM, and Kraan, BCP, (1998) "Evaluation of weighting schemes for expert judgment studies, *Proceedintgs PSAM4* (Mosleh and Bari eds.) Springer, New Your, 1937-1942.
17. Goossens, L.H.J., and Harper, F.T., (1998) "Joint EC/USNRC expert judgement driven radiological protection uncertainty analysis*", Journal of Radiological Protection*, Vol.18, No.4, 249-264.
18. Goossens, L.H.J., Cooke, R.M. and van Steen, J. (1989) Final Report to the Dutch Ministery of Housing, Physical Planning and Environment: *On The Use of Expert Judgment in Risk and Safety Studies* Vol. I –Vol 5. Delft.
19. Kraan, B. and Bedford, T. (2005) "Probabilistic inversion of expert judgments in the quantification of model uncertainty" *Management Science* vol. 51, no 6, 995-1006.
20. Kraan, B.C.P and Bedford. T.J. 2005 Probabilistic inversion of expert judgements in the quantification of model uncertainty. *Management Science*, 51(6):995-1006.
21. Kraan. B.C.P. *Probabilistic Inversion in Uncertainty Analysis and related topics*. 2002 PhD dissertation, TU Delft, Dept. Mathematics.
22. Kruithof J.. Telefoonverkeersrekening. *De Ingenieur*, 52(8):E15{E25, 1937.
23. Kurowicka, D. and Cooke, R.M. (2006) *Uncertainty Analysis with High Dimensional Dependence*, Wiley, New York, 2006.
24. Lin, Shi-Woei and Bier, V.M. (2008), A Study of Expert Overconfidence, Special issue on expert judgment Reliability Engineering & System SafetyAvailable online 12 March 2007.
    Neslo, R. Micheli, F., Kappel, C.V. Selkoe, IK.A. Halpern, B.S. and Cooke, R.M. "Stakeholder Preferences for Coastal Scenarios", Ressources for the Future 2008.
25. NUREG/CR-6372 (1997) Recommendations for Probabilistic Seismic Hazard Analysis:Guidance on Uncertainty and Use of Experts., US Nuclear Regulatory Commission,
26. O'Hagan, A., Buck, C.E., Daneshkhah, Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J. Oakley, J.E. and Rakow, T. 2006. Uncertain Judgments, Eliciting experts' Probabilities, Wiley.
27. Qing, X. (2002) Risk analysis for real estate investment, PhD thesis, Dept. of Architecture, TU Delft.
28. Rabin Neslo,R. Fiorenza Micheli,F., Carrie V. Kappel, C.V., Kimberly A. Selkoe, K.A. Benjamin S. Halpern, B.S. Roger M. Cooke, R.M. (2008) **Modeling Stakeholder Preferences with Probabilistic Inversion: Application to Prioritizing Marine ecosystem Vulnerabilities, Dept. of Mathematics, TU. Delft,**
29. Train K.E. 2003 "Discrete Choice Methods with Simulation" Cambridge University Press.
30. Winkler, R.L., Wallsten, T.S., Whitfield, R.G. Richmond, H.M. Hayes, S.R. and Rosenbaum, A.S. (1995) "An assessment of the risk of chronic lung injury

attributable to long-term ozon exposure" *Operations Research*, vol. 43, no. 1 19 - 27.