Available online at www.sciencedirect.com
ScienceDirect
Reliability Engineering and System Safety 93 (2008) 657–674
RELIABILITY ENGINEERING & SYSTEM SAFETY
www.elsevier.com/locate/ress

TU Delft expert judgment data base

Roger M. Cooke[a,*], Louis L.H.J. Goossens[b]

[a]Resources for the Future and Department of Mathematics, Delft University of Technology, Mekelweg 4, Delft, The Netherlands
[b]Department of Safety Science, Delft University of Technology, TU Delft, The Netherlands

Available online 15 March 2007

Abstract

We review the applications of structured expert judgment uncertainty quantification using the "classical model" developed at the Delft University of Technology over the last 17 years [Cooke RM. Experts in uncertainty. Oxford: Oxford University Press; 1991; Expert judgment study on atmospheric dispersion and deposition. Report Faculty of Technical Mathematics and Informatics No.01-81, Delft University of Technology; 1991]. These involve 45 expert panels, performed under contract with problem owners who reviewed and approved the results. With a few exceptions, all these applications involved the use of seed variables; that is, variables from the experts' area of expertise for which the true values are available post hoc. Seed variables are used to (1) measure expert performance, (2) enable performance-based weighted combination of experts' distributions, and (3) evaluate and hopefully validate the resulting combination or "decision maker". This article reviews the classical model for structured expert judgment and the performance measures, reviews applications, comparing performance-based decision makers with "equal weight" decision makers, and collects some lessons learned.

Keywords: Expert judgment; Rational consensus; Calibration; Information; Subjective probability

0. Introduction

The pros and cons of different weighting schemes remain a subject of research. The European Union contracted the TU Delft to review its applications both within EU projects, and elsewhere, in which experts assessed variables in their field of expertise for which the true values are known, in addition to variables of interest [1,2]. These are called *seed*, or *calibration*, variables. Since then, the TU Delft expert judgment data base has nearly doubled. We now have studies involving over 67,000 experts' subjective probability distributions. The main sectors and summary information are given in Table 1.

The authors believe that this data base represents a unique source from which much can be learned regarding the application of structured expert judgment in quantitative decision support. The entire data, appropriately anonymized, may be obtained from the first author. It is hoped that others will use this data to further develop methods for using structured expert judgment.

We assume that uncertainty is represented as subjective probability and concerns results of possible observations. For a discussion of foundational issues, the reader is referred to [3]. Section 1 discusses goals of a structured expert judgment study; Section 2 provides an explanation of the concepts and methods underlying the Delft expert judgment method. Section 3 gives an updated summary of the results, comparing equal weighting with performance-based weighting and with the best expert. Section 4 discusses seed variables and robustness, and Section 5 is devoted to lessons learned and anecdotal information, common pitfalls, and misconceptions. A concluding section identifies possible topics for future research. Another article in this issue compares the performance of social network weighted combinations, based on citations, and likelihood weighted combinations. One recent study from the Harvard Kuwait project is discussed in detail in another article in this issue.

1. Structured expert judgment

Expert judgment is sought when substantial scientific uncertainty impacts on a decision process. Because there is

*Corresponding author. Tel.: +31 15 2782548; fax: +31 15 2787255.
E-mail address: cooke@rff.org (R.M. Cooke).

0951-8320/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ress.2007.03.005

Table 1
Summary of applications per sector

Sector	No. of experts	No. of variables	No. of elicitations
Nuclear applications	98	2203	20,461
Chemical ind. & gas industry	56	403	4491
Groundwater/water pollution/dike ring/barriers	49	212	3714
Aerospace sector/space debris/aviation	51	161	1149
Occupational sector: ladders/buildings (thermal physics)	13	70	800
Health: bovine/chicken (*Campylobacter*)/SARS	46	240	2979
Banking: options/rent/operational risk	24	119	4328
Volcanoes/dams	231	673	29,079
Rest group	19	56	762
In total	521	3688	67,001

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. 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, and
- 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. [4, 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 [5]. The philosophical underpinnings of this approach are elaborated in [6]. Expert agreement on the representation of the overall community is the weakest, and most accessible, type of consensus to which a study may aspire. Agreement on a "distribution to represent a group", agreement on a distribution and agreement on a number are the other types of consensus, in decreasing accessibility.

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 way the different groups are included equally in the resulting representation of uncertainty. This was the reasoning behind the selection of expert panels in the EU USNRC accident consequence studies with equal weighting [7].

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 [8,9] 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—this is not rational dissent and does not threaten 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 4. 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 [10]. 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 overconfident 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 [8,9] 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, and
- the experts' distributions over the seed variables and the variables of interest.

Issues that arise in building such a model are discussed in [8,9]. 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.

2. The classical model

The above principles have been operationalized in the so-called "classical model", a performance-based linear pooling or weighted averaging model [8,9,11]. The weights 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.

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 $10\%, 20\%, \ldots, 90\%$. 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.

2.1. 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 4 inter-quantile intervals with probability vector

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

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

$$s_1(e) = \#\{i | x_i \leqslant 5\% \ \ quantile\}/N,$$

$$s_2(e) = \#\{i | 5\% \ \ quantile < x_i \leqslant 50\% \ \ quantile\}/N,$$

$$s_3(e) = \#\{i | 50\% \ \ quantile < x_i \leqslant 95\% \ \ quantile\}/N,$$

$$s_4(e) = \#\{i | 95\% \ \ quantile < x_i\}/N,$$

$$s(e) = (s_1, \ldots, s_4).$$

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

$$2NI(s(e)|p) = 2N \sum_{i=1,\ldots,4} s_i \ln(s_i/p_i) \quad (1)$$

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) [8,9].

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

$$\text{calibration score}(e) = p - \text{value} = \text{prob}\{2NI(s(e)|p) \geqslant r | H_e\},$$

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 support 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.

2.2. 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 (other background measures are discussed in [12].

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

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

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 $1,\ldots,N$ is

$$\begin{aligned}&\text{information score}(e)\\&\quad=\text{average relative information wrt background}\\&\quad=(1/N)\sum_{i=1,\ldots,N} I(f_{e,i}|g_i),\end{aligned}$$

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 concentratedness could be contemplated. The above information score is chosen because it is

- familiar,
- tail insensitive,
- scale invariant, and
- slow.

The last 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.

2.3. 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 pros and cons of the linear pool see [8,9,13,14]. 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. An example of this is given in Section 4.

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:

$$\begin{aligned}w_\alpha(e)=1_\alpha(&\text{calibration score})\times\text{calibration score}(e)\\&\times\text{information score}(e),\end{aligned}\tag{2}$$

where $1_\alpha(x)=0$ if $x<\alpha$ and $1_\alpha(x)=1$ otherwise. Cooke [8,9] 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 [8,9].

The scoring rule constraint requires the term 1_α(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 $\mathrm{DM}_\alpha(i)$ be the result of linear pooling for item i with weights proportional to (2):

$$\mathrm{DM}_\alpha(i)=\sum_{e=1,\ldots,E} w_\alpha(e)f_{e,i}\Big/\sum_{e=1,\ldots,E} w_\alpha(e).\tag{3}$$

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

$$\text{calibration score}(\mathrm{DM}_\alpha)\times\text{information score}(\mathrm{DM}_\alpha).\tag{4}$$

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:

$$\begin{aligned}w_\alpha(e,i)=1_\alpha(&\text{calibration score})\\&\times\text{calibration score}(e)\times I(f_{e,i}|g_i).\end{aligned}\tag{5}$$

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

$$\mathrm{IDM}_\alpha(i)=\sum_{e=1,\ldots,E} w_\alpha(e,i)f_{e,i}\Big/\sum_{e=1,\ldots,E} w_\alpha(e,i).\tag{6}$$

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

$$\text{calibration score}(\mathrm{IDM}_\alpha) \times \text{information score}(\mathrm{IDM}_\alpha). \quad (7)$$

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 decision maker would perform better than the result of simple averaging, called the *equal weight DM*, 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 outperform equal weighting or outperform 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.

3. Applications of the classical model

Forty-five 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. For studies by Ter Haar [15], the data have not been retrieved.

These are all studies performed under contract for a problem owner and reviewed and accepted by the contracting party. In most cases these have been published. Table 2 lists these studies, references publications, and gives summary information. The number of variables and number of seed variables are shown, as is the number of effective seed variables. In general the effective number of seeds is equal to the least number of seeds assessed by some expert. In this way each expert is scored with a test of the same power. In the gas panel, the panel and the seed variables were split post hoc into corrosion and environmental panels.

The combined scores of equal weight DM, performance-based DM, and best expert are compared pair wise in Fig. 1. Fig. 2 compares the calibration (p-values) and information scores of the equal weight DM, the performance-based DM, and the best expert.

In 15 of 45 cases the performance-based DM was the best expert, that is, one expert received weight one. In 27 cases the combined score of the performance-based DM was strictly better than both the equal weight DM and the best expert. In one case (13) the equal weight DM performed best, and in two cases (10, 22) the best expert outperformed both equal weights and performance-based weights.

The equal weight DM 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 equal weight DM 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 equal weight DM this happened in 7 cases (15%).

[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.

Table 2
Expert judgment studies

Case	Name/Reference	No. of experts	No. of variables/No. of seeds	No. of effective seeds	Perf. measure	Perform weights	Equal weights	Best expert
1	Dsm-1	10	14/8	8	Calibr'n	0.66	0.53	0.54
Flange leak	[10,14]				Inform'n	1.371	0.8064	1.549
					Combi'n	0.905	0.4274	0.836
2	Dsm-2	8	39/12	11	Calibr'n	0.84	0.5	0.005
Crane risk	[16]				Inform'n	1.367	0.69	2.458
					Combi'n	1.148	0.345	0.012
3	Estec-1	4	48/13	13	Calibr'n	0.43	0.43	0.14
Propulsion	[10,14]				Inform'n	1.72	1.421	2.952
					Combi'n	0.7398	0.611	0.413
4	Estec-2	7	58/26	18	Calibr'n	0.78	0.9	0.0001
Space debris	[17,48]				Inform'n	0.32	0.15	2.29
					Combi'n	0.25	0.14	0.0002
5	Estec-3	6	22/12	12	Calibr'n	0.27	0.12	0.005
Composite materials	[4]				Inform'n	1.442	0.929	2.549
					Combi'n	0.39	0.111	0.013
6	AOT(daily)	9	38/38	6	Calibr'n	0.95	0.95	0.95
Option trading	[18,51]				Inform'n	0.5043	0.2156	0.5043
					Combi'n	0.4791	0.2048	0.4791
7	AOT(risk)	5	11/11	11.00	Calibr'n	0.8287	0.324	0.8287
Risk management	[18,51]				Inform'n	1.212	0.7449	1.212
					Combi'n	1.003	0.2413	1.003
8	Grond5	7	38/10	10	Calibr'n	0.7	0.05	0.4
Groundwater transport	[19,42]				Inform'n	3.008	3.16	3.966
					Combi'n	2.106	0.158	1.586
9	Tuddispr	11	58/36	36	Calibr'n	0.68	0.71	0.36
Dispersion panel TUD	[8,9]				Inform'n	0.827	0.715	1.532
					Combi'n	0.562	0.508	0.552
10	Tnodispr	7	58/36	36	Calibr'n	0.69	0.32	0.53
Dispersion panel TNO	[9]				Inform'n	0.875	0.751	1.698
					Combi'n	0.604	0.24	0.9002
11	Tuddepos	4	56/24	22	Calibr'n	0.45	0.34	0.45
Dry deposition	[8,9]				Inform'n	1.647	1.222	1.647
					Combi'n	0.741	0.415	0.741

Table 2 (*continued*)

Case	Name/Reference	No. of experts	No. of variables/No. of seeds	No. of effective seeds	Perf. measure	Perform weights	Equal weights	Best expert
12 Acrylo-nitrile	Acnexpts [2,11,20,46]	7	43/10	10	Calibr'n Inform'n Combi'n	0.24 3.186 0.764	0.28 1.511 0.423	0.24 3.186 0.764
13 Ammonia panel	Nh3expts [2,11,20,46]	6	31/10	10	Calibr'n Inform'n Combi'n	0.11 1.672 0.184	0.28 1.075 0.301	0.06 2.627 0.158
14 Sulphur tri oxide	So3expts [2,11,20,46]	4	28/7	7	Calibr'n Inform'n Combi'n	0.14 3.904 0.547	0.14 2.098 0.294	0.02 4.345 0.087
15 Water pollution	Waterpol [21]	11	21/11	10	Calibr'n Inform'n Combi'n	0.35 1.875 0.6563	0.35 1.385 0.4847	0.16 2.06 0.3296
16 Dispersion panel	Eunrcdis [22–24]	8	77/23	23	Calibr'n Inform'n Combi'n	0.9 1.087 0.9785	0.15 0.862 0.129	0.13 1.242 0.161
17 Dry deposition	Eunrcdd [22–24]	8	87/14	14	Calibr'n Inform'n Combi'n	0.52 1.339 0.697	0.001 1.184 0.001	0.52 1.339 0.697
18 Rad. transp. in animals	Eunrca_s [22,23,25]	7	80/8	6	Calibr'n Inform'n Combi'n	0.75 2.697 2.023	0.55 1.778 0.978	0.75 2.697 2.023
19 Wet deposition	Euncrwd [22–24]	7	50/19	19	Calibr'n Inform'n Combi'n	0.25 0.451 0.113	0.001 0.726 0.00073	0.01 0.593 0.0059
20 Rad. internal dose	Eunrcint [22,23,26]	8	332/55	28	Calibr'n Inform'n Combi'n	0.85 0.796 0.677	0.11 0.5598 0.062	0.73 0.822 0.6001
21 Rad. early health effects	Eunrcear [22,23,27,47]	9	489/15	15	Calibr'n Inform'n Combi'n	0.23 0.2156 0.0496	0.07 0.1647 0.01153	0.0001 1.375 0.00014
22 Rad. trans. soil	Euncrsoi [22,23,25]	4	244/31	31	Calibr'n Inform'n Combi'n	0.0001 1.024 0.0001	0.0001 0.973 9.7E−05	0.0001 2.376 0.0002
23 Environm. panel	Gas95 [28]	15	106/28	17	Calibr'n Inform'n Combi'n	0.93 1.628 1.514	0.11 1.274 0.14	0.06 2.411 0.145
24 Corrosion panel	Gas95 [28]	12	58/11	11	Calibr'n Inform'n Combi'n	0.16 2.762 0.4419	0.06 1.304 0.078	0.16 2.762 0.4419

25 Moveable barriers floodrisk	Mvblbarr [29]		52/14	14	Calibr'n Inform'n Combi'n	0.43 1.243 0.535	0.22 0.57 0.125	0.04 1.711 0.068
26 Real estate risk	Realestr [30]	5	45/31	31	Calibr'n Inform'n Combi'n	0.82 0.7648 0.6296	0.005 0.1735 0.0009	0.82 0.7678 0.6296
27 River channel	Rivrchnl [31,52]	6	14/8	8	Calibr'n Inform'n Combi'n	0.53 0.843 0.447	0.64 0.289 0.185	0.53 0.843 0.447
28 Montserrat volcano	Mont1 [32,33]	11	13/8	8	Calibr'n Inform'n Combi'n	0.66 1.906 1.258	0.53 0.8217 0.4355	0.66 1.906 1.258
29 Thermal phys. blds	Thrmbld [3,44]	6	48/48	10	Calibr'n Inform'n Combi'n	0.3628 0.5527 0.2005	0.02485 0.1424 0.00354	0.3628 0.5527 0.2005
30 Dike ring failure	Dikring [13,34,43,45]	17	87/47	47	Calibr'n Inform'n Combi'n	0.4 0.614 0.2456	0.05 0.7537 0.03768	0.3 0.6462 0.1938
31 Campylobacter NL	Carma [15]	12	98/10	10	Calibr'n Inform'n Combi'n	0.828 1.48 1.226	0.4735 0.2038 0.09648	0.828 1.48 1.226
32 Campy Greece	CARME-Greece [35,49]	6	98/10	10	Calibr'n Inform'n Combi'n	0.4925 0.8611 0.4241	0.5503 0.3428 0.1886	0.4925 0.8611 0.4241
33 Oper. risk	Opriskbank [36]	10	36/16	16	b Inform'n Combi'n	0.4301 0.7827 0.3263	0.338 0.3219 0.1088	0.1473 0.903 0.133
34 Infosec	infosec [37]	13	32/10	10	Calibr'n Inform'n Combi'n	0.7071 1.721 1.217	0.7971 1.012 0.7159	0.3135 2.232 0.6999
35 PM25	PM25	6	24/12	12	Calibr'n Inform'n Combi'n	0.578 0.807 0.466	0.6451 0.542 0.3497	0.1195 1.486 0.1776
36 Falls ladders	Ladders	7	22/10	10	Calibr'n Inform'n Combi'n	0.2441 0.975 0.238	0.3005 0.4638 0.1394	0.00131 1.801 0.00236
37 Dams	Dams [38]	11	74/11	11	Calibr'n Inform'n Combi'n	0.615 1.248 0.7677	0.492 0.6446 0.3171	0.01088 2.359 0.02566
38 MVO seeds Montserrat follup	MVOseeds [33,39,50]	77	5/5	5	Calibr'n Inform'n Combi'n	0.6084 3.116 1.896	0.3946 1.147 0.4525	0.6084 3.116 1.896

Table 2 (*continued*)

Case	Name/Reference	No. of experts	No. of variables/No. of seeds	No. of effective seeds	Perf. measure	Perform weights	Equal weights	Best expert
39 Pilots	Pilots [32]	31	63/10	10	Calibr'n Inform'n Combi'n	0.4735 0.6903 0.3269	0.5503 0.5946 0.2777	0.1917 1.403 0.2689
40 Sete Cidades	Sete cidades	19	27/10	10	Calibr'n Inform'n Combi'n	0.7901 2.709 2.141	0.1065 0.8409 0.1713	0.4281 2.474 1.059
41 TeideMay_05	TeideMay_05	17	23/10	10	Calibr'n Inform'n Combi'n	0.7069 2.178 1.54	0.1135 1.681 0.1907	0.04706 3.322 0.1563
42 Vesuvio	VesuvioPisa21Mar05	14	79/10	10	Calibr'n Inform'n Combi'n	0.6827 2.43 1.659	0.4735 1.485 0.7029	0.4706 3.622 0.1705
43 Volcrisk	Volcrisk	45	30/10	10	Calibr'n Inform'n Combi'n	0.8283 0.7738 0.641	0.1135 0.5571 0.06322	0.8283 0.7738 0.641
44 Sars	Sars	9	20/10	10	Calibr'n Inform'n Combi'n	0.6827 1.34 0.9149	0.4735 0.6017 0.2849	0.06083 2.31 0.1405
45 Guadeloupe	Guadeloupe	9	57/10	10	Calibr'n Inform'n Combi'n	0.4925 2.158 1.063	0.4735 1.176 0.5567	0.0008 3.649 0.00029

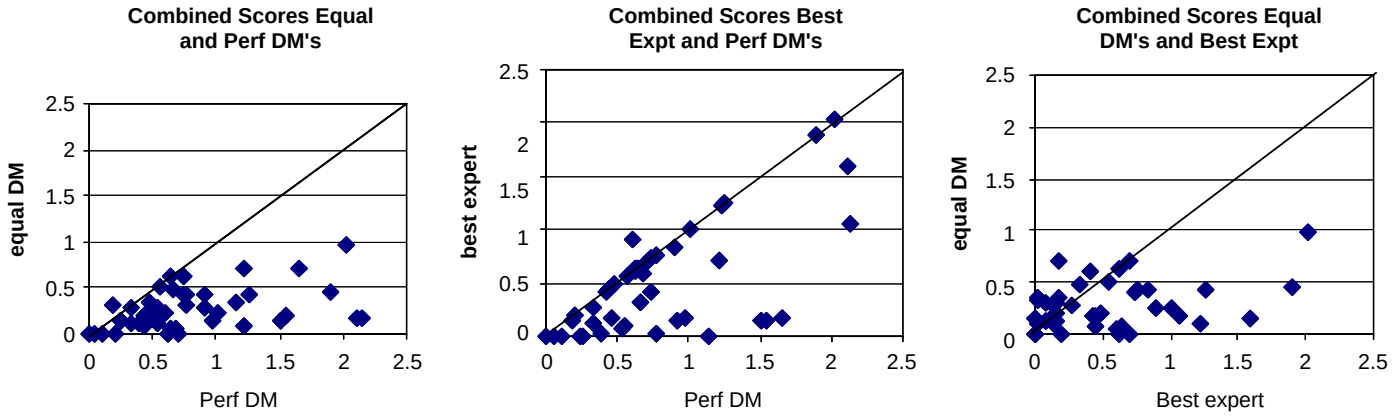

Fig. 1. Combined scores of equal weight DM, performance-based DM, and best expert.

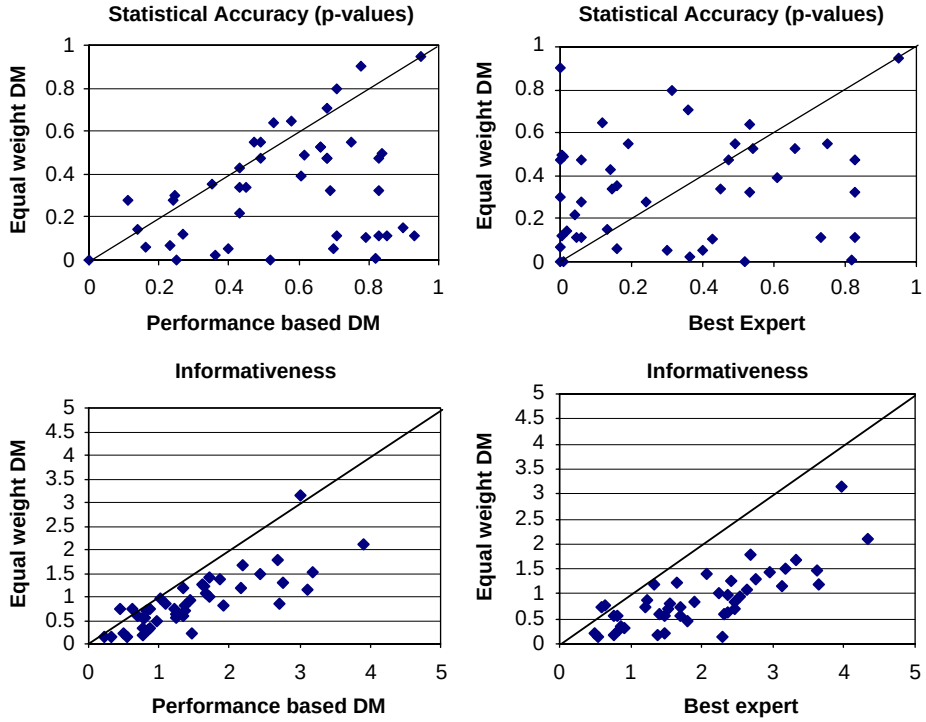

Fig. 2. Calibration (*p*-values) and information scores of equal weight DM, performance-based DM, and best expert.

The study on radiological transport in soil (22) was unusual in that all the experts and all decision makers performed badly. Both the seed variables and the experts were identified by the National Radiological Protection Board, and re-analysis of the seed variables and expert data did not yield any satisfactory explanation for the poor performance. We concluded that this was simply due to the small number of experts and bad luck.

The motivation for performance-based weighting above equal weighting speaks for itself from there data. Sometimes the difference is marginal but sometimes it is quite significant. Most often the equal weight DM is slightly less well-calibrated and significantly less informative, but sometimes the calibration of the equal weight DM is quite poor (17, 26). 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 point is that the performance measures are objective and fully transparent. It is impossible to tweak these measures for extra-scientific expediency.

4. 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 [40]. 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". It is hoped that the TU Delft data base can contribute to a further analysis of these issues.

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 a few studies to illustrate good and poor choices of seed variables and, where possible, to compare with variables of interest.

4.1. Real estate risk

In this study 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 in Fig. 3.

The robustness analyses in this case are also revealing. First we examine the five experts' (3 portfolio managers and 2 risk analysts) and DM's scores, and the relative

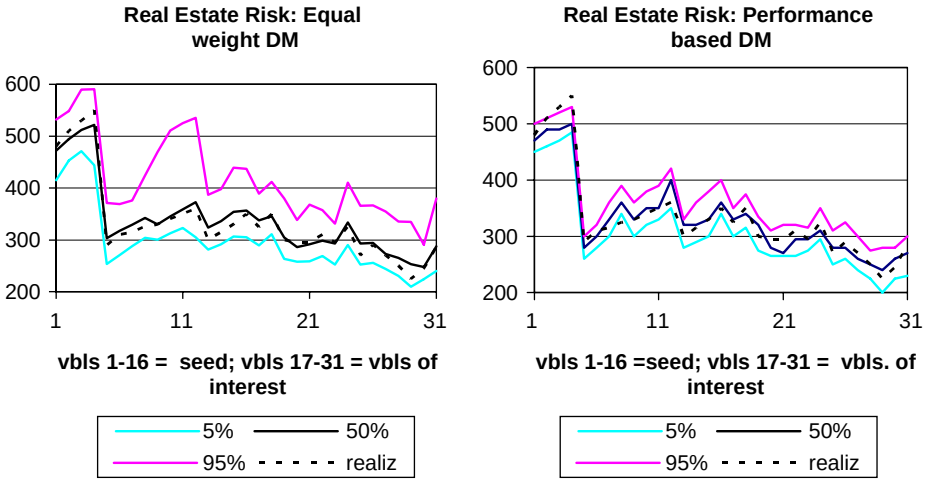

Fig. 3. Real estate risk, equal weight, and performance-based decision makers.

Table 3
Real estate risk, relative information of the five experts to the equal weight combination for all variables and for variables with realizations

Id	Calibr.	Mean rel. inf total	Mean rel. inf seed vbls	Numb real	Unnormalized weight	Rel. inf to eq. wgt DM	
						All vbls	Seed vbls
Portfol1	0.3303	0.7932	0.8572	16	0.2832	0.5004	0.6241
Portfol2	0.1473	1.02	0.9554	16	0	0.7764	0.6545
Portfol3	0.02012	0.2492	0.1556	16	0	0.3633	0.2931
Riskan1	6.06E−05	1.334	1.536	16	0	0.9575	1.21
Riskan2	0.004167	0.5848	0.6126	16	0	0.4579	0.4402
Perf DM	0.3303	0.7932	0.8572	16	0.2832		
Equal DM	0.05608	0.1853	0.179	16	0.01004		

Table 4
Real estate risk; robustness analysis on seed variables

Excluded item	Rel. info/b total	Rel. info/b seeds	Calibr.	Rel. info/orig DM total	Rel. info/orig DM seeds
Q1Rent Amster.	0.5875	0.6234	0.3578	0.3539	0.37
Q2Rent Amster.	0.5974	0.6341	0.3578	0.4402	0.4421
Q3Rent Amster.	0.7921	0.8583	0.5435	0	0
Q4Rent Amster.	0.7859	0.8401	0.5435	0	0
Q1Rent Rotter.	0.5871	0.6047	0.3578	0.4438	0.4565
Q2Rent Rotter.	0.5857	0.6004	0.3578	0.4491	0.4708
Q3Rent Rotter.	0.8009	0.8841	0.387	0	0
Q4Rent Rotter.	0.5872	0.6222	0.3578	0.3505	0.3575
Q1Rent Denhaag	0.7886	0.8478	0.387	0	0
Q2Rent Denhaag	0.7861	0.8406	0.387	0	0
Q3Rent Denhaag	0.784	0.8345	0.387	0	0
Q4Rent DenHaag	0.7845	0.8358	0.387	0	0
Q1Rent Utrecht	0.6034	0.6396	0.288	0.4589	0.4353
Q2Rent Utrecht	0.6069	0.6517	0.288	0.4663	0.4644
Q3Rent Utrecht	0.6013	0.6356	0.288	0.4656	0.464
Q4Rent Utrecht	0.794	0.8638	0.387	0	0
Original Perf DM	0.7932	0.8572	0.3303		

Table 5
Real estate risk; robustness analysis on experts

Excluded expert	Rel. info/b total	Rel. info/b seeds	Calibr.	Rel. info/orig DM total	Rel. info/orig DM seeds
Portfol1	1.006	0.9484	0.1473	1.144	1.058
Portfol2	0.637	0.6899	0.7377	0.2916	0.3328
Portfol3	0.5297	0.4825	0.3303	0	0
Riskan1	0.7921	0.8572	0.3303	0	0
Riskan2	0.7079	0.8195	0.3303	0	0
Original Perf DM	0.7932	0.8572	0.3303	0	0

information of each of the experts to the equal weight combination of their distributions (Table 3). This gives a benchmark for how well the experts agree among themselves. The experts' densities are constructed relative to a background measure, so these comparisons also depend on the background measure. The relatively weak calibration performance of the equal weight DM is due to the fact that only 4 of the 16 seed variables were above the median assessment.[2] At the same time, the equal DM's medians are actually a bit closer to the realizations. Distance between median and realization is an example of a scoring variable which is *not* taken into account by the performance-based DM.[3] Note also that the pattern of informativeness on seed variables is comparable to that on all variables; portfolio manager 3 is least informative and risk analyst 1 is most informative. Note also that low informativeness does not translate automatically into better calibration.

Next we remove the 16 seed variables one at a time and re-compute the performance-based DM (Table 4).

The scores do not change much, but the relative information of the "perturbed DM" with respect to the original DM is rather large for 8 of the variables, comparable to the differences between the experts themselves. The explanation can be found by examining the robustness on experts (Table 5).

If we remove portfolio manager 1, the effect on the DM is large, comparable to the largest relative information between a single expert and the equal weight combination. This is not surprising as portfolio manager 1 coincides with the performance-based DM. Interestingly, we get a significant change by removing portfolio manager 2. This is because the combination of portfolio managers 1 and 3 would give a higher score than portfolio manager 1 alone, or 1 and 2 alone. We should have to give portfolio manager 2 weight zero and portfolio manager 3 positive weight, even though the latter's calibration score is worse than that of the former. The proper scoring rule constraint prevents this from happening. This

[2]The values cited in Table 3 are based on 31 seed variables, using also the variables of interest which became available a year later.

[3]The reason is that distance is scale dependent. In this case the scales of all variables are the same, so such a scoring variable could be used. Of course such a rule may not be proper.

underscores the difference noted in Section 2 between optimization under the proper scoring rule constraint, and unconstrained optimization. In the latter case a better calibrated expert can have less weight than a poorly calibrated expert. The non-robustness in Table 4 is caused by the fact that the removal of some seed variables cause the calibration of portfolio manager 2 to dip below that of portfolio manager 3.

4.2. AEX

In this case, the seed variables *were* the variables of interest, namely the opening price of the Amsterdam Stock Exchange, as estimated at closing the previous day. Note that some of the experts anticipated a large drop on the day corresponding to variable 20. This was not reflected in the performance-based DM, nor in the realization. Other than that, the pattern across seed variables does not look erratic. In spite of the excellent performance of the experts in this case, they were not able to predict the opening price better than the "historical average predictor". In other words, any information the experts might have had at closing time was already reflected in the closing price (Fig. 4).

4.3. Dry deposition

The seed variables were measured deposition velocities, though not configured according to the requirements of the study (per species, windspeed, particle diameter, and surface) (Fig. 5).

Here again, the poor statistical performance of the equal weight DM is due to the fact that all but one of the 14 seed variables fall above the median.

4.4. Dike ring

The seed variables were ratio of predicted versus measured water levels at different, at water levels around 2 m above the baseline. Variables of interest were the same, but at water levels above 3.5 m above the baseline. In this case we had several realizations of this ratio from each of several measuring stations. That explains the step pattern of the quantiles; these are actually the same assessment with several realizations (Fig. 6).

Although all 47 seed variables were used in the analysis, for purposes of comparing expert performance with that in other studies, the effective number of seeds was reduced to 10. This accounts for dependence in the experts'

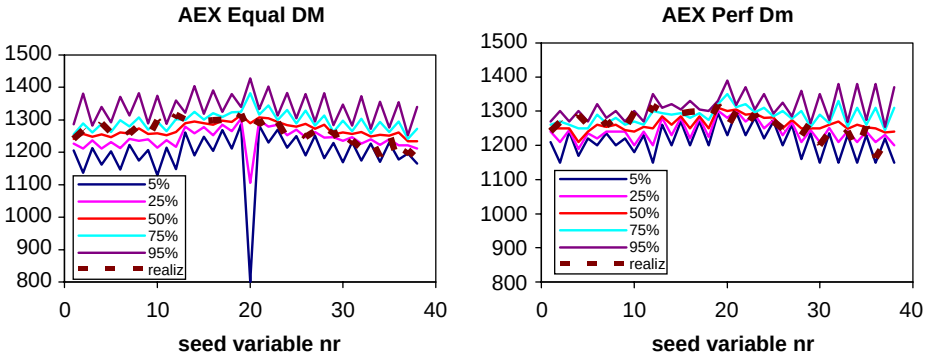

Fig. 4. AEX, equal weight, and performance-based decision makers.

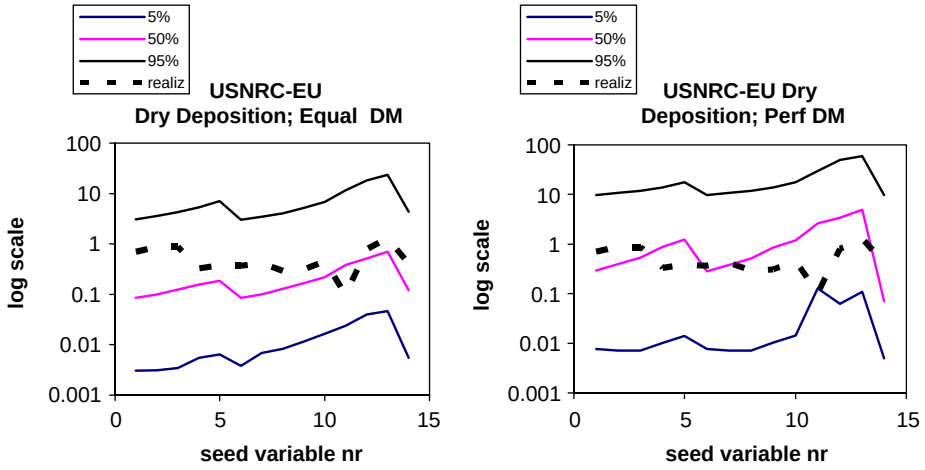

Fig. 5. Dry deposition equal weight, and performance-based decision makers.

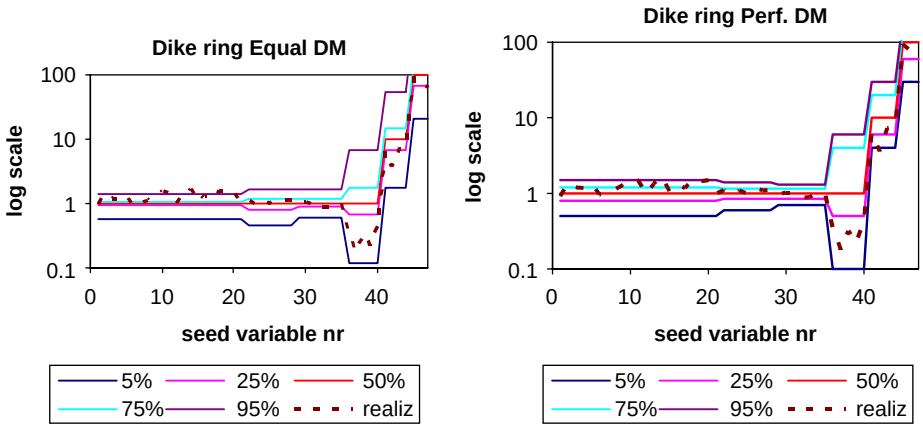

Fig. 6. Dike ring, equal weight, and performance-based decision makers.

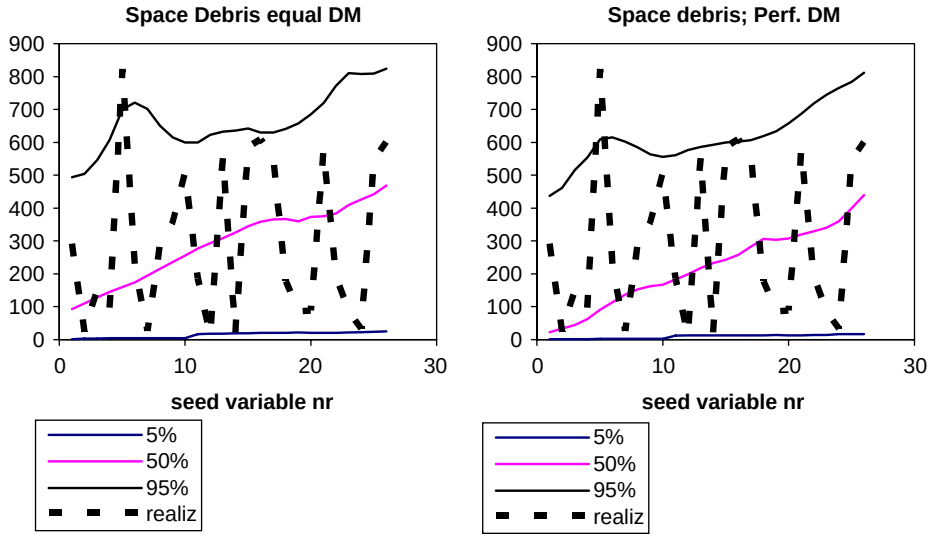

Fig. 7. Space debris, equal weight and performance-based decision makers.

assessments and corresponds to the number most often used for such comparisons.

4.5. Space debris

The seed variables were numbers of tracked space debris particles injected into orbit between the years 1961 and 1986. Variables of interest characterized the debris flux for 10 years into the future. It turned out that the experts did not possess year-by-year knowledge of the debris particles, and gave generic assessments assuming that the number was growing, where in fact the number appears to be quite random. This is a case in which the choice of seed variables was unsuccessful; the experts did not really have relevant knowledge to apply to the task (Fig. 7).[4]

5. Lessons learned from elicitations

A detailed description of the design of an expert judgment study is given in [34]. Suffice to say here that a typical study involves a dry run with one expert to finalize the elicitation questions. This is followed by a plenary meeting of all experts in which the issues are discussed, the study design is explained, and a short elicitation exercise is done. This involves a small number of seed variables, typically 5. Experts are shown how the scoring and combining works. Afterwards, the experts are elicited individually. An elicitation session should not exceed half day. Fatigue sets in after 2 h.

When experts are dispersed it may be difficult and expensive to bring them together. In such cases the training is given to each expert in abbreviated form. The EU-

[4]In this early study, the effective number of seed variables was chosen to optimize the DM's performance, a procedure which is no longer followed. The DOS version of the software used a table of the chi-square distribution and had problems with very low calibration scores. These problems came to the fore when the number of seed variables is high, as in this case.

(*footnote continued*)

USNRC studies made the most intensive investment in training. In general, it is not advisable to configure the exercise such that the presence of *all* experts at one time and place is essential to the study, as this makes the study vulnerable to last minute disruptions.

The following are some practical guidelines for responding to typical comments:

From an expert: I don't know that

Response: No one knows, if someone knew we would not need to do an expert judgment exercise. We are tying to capture your uncertainty about this variable. If you are very uncertain then you should choose very wide confidence bounds.

From an expert: I can't assess that unless you give me more information.

Response: The information given corresponds with the assumptions of the study. We are trying to get your uncertainty conditional on the assumptions of the study. If you prefer to think of uncertainty conditional on other factors, then you must try to unconditionalize and fold the uncertainty over these other factors into your assessment.

From an expert: I am not the best expert for that.

Response: We don't know who are the best experts. Sometimes the people with the most detailed knowledge are not the best at quantifying their uncertainty.

From an expert: Does that answer look OK?

Response: You are the expert, not me.

From the problem owner: So you are going to score these experts like school children?

Response: If this is not a serious matter for you, then forget it. If it is serious, then we must take the quantification of uncertainty seriously. Without scoring we can never validate our experts or the combination of their assessments.

From the problem owner: The experts will never stand for it.

Response: We've done it many times, the experts actually like it.

From the problem owner: Expert number 4 gave crazy assessments, who was that guy?

Response: You are paying for the study, you own the data, and if you really want to know I will tell you. But you don't need to know, and knowing will not make things easier for you. Reflect first whether you really want to know this.

From the problem owner: How can I give an expert weight zero?

Response: Zero weight does not mean zero value. It simply means that this expert's knowledge was already contributed by other experts and adding this expert would only add a bit of noise. The value of unweighted experts is seen in the robustness of our answers against loss of experts. Everyone understands this when it is properly explained.

From the problem owner: How can I give weight one to a single expert?

Response: By giving all the others weight zero, see previous response.

From the problem owner: I prefer to use the equal weight combination.

Response: So long as the calibration of the equal weight combination is acceptable, there is no scientific objection to doing this. Our job as analyst is to indicate the best combination, according to the performance criteria, and to say what other combinations are scientifically acceptable.

6. Conclusion

Given the body of experience with structured expert judgment, the scientific approach to uncertainty quantification is well established. This does not mean the discussion on expert judgment method is closed.

First of all, we may note that a full expert judgment study is not cheap. Most of the studies mentioned above involved 1–3 man months. This cost could be reduced somewhat if we did not need to develop seed variables. However, simply using equal weights does not seem to be a convincing alternative. Other methods of measuring and verifying performance would be welcome, especially if they are less resource-intensive.

The classical model is based on the two performance measures, calibration and information in conjunction with the theory of proper scoring rules. It satisfies necessary conditions for rational consensus, but is not *derived* from those conditions. Other weighting schemes could surely be developed with do as well or better in this regard, and other performance measures could be proposed and explored.

Once we acknowledge that our models must by quantified with uncertainty distributions, rather than 'nominal values' of undetermined pedigree, many new challenges confront modelers, analysts, and decision makers.

Experts can quantify their uncertainty about potentially observable phenomena with which they have some familiarity. The requirements of the study at hand may go beyond that. For example, in quantifying the uncertainty of models for transport of radiation through soils, plants, and animals, it emerged that the institutes which built and maintained these models could not supply any experts who were able to quantify uncertainty on the transfer coefficients in these models. Experts could quantify uncertainty with regard to quantities which can be expressed as functions of the transport models themselves. Processing data of this sort required development of sophisticated techniques of probabilistic inversion [19,41].

Perhaps the greatest outstanding problems concern the elicitation, representation, and computation with dependence. Everyone knows that the ubiquitous assumption of independence in uncertainty analysis is usually wrong, and sometimes seriously wrong. This is a subject that must receive more attention in the future [37].

Acknowledgements

The authors gratefully acknowledge the contributions of many people who cooperated in developing this data base. Willy Aspinall and Tim Bedford are independently responsible for a quarter of the studies.

References

[1] Goossens LHJ, Cooke RM, Kraan BCP. 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; 1995.

[2] Goossens LHJ, Cooke RM, Kraan BCP. Evaluation of weighting schemes for expert judgment studies. In: Mosleh, Bari, editors. Proceedintgs PSAM4. New York: Springer; 1998. p. 1937–42.

[3] Cooke RM. The anatomy of the Squizzle—the role of operational definitions in science. Reliab Eng Syst Safety 2004;85:313–9.

[4] NUREG/CR-6372. Recommendations for probabilistic seismic hazard analysis: guidance on uncertainty and use of experts. US Nuclear Regulatory Commission; 1997.

[5] Winkler RL, Wallsten TS, Whitfield RG, Richmond HM, Hayes SR, Rosenbaum AS. An assessment of the risk of chronic lung injury attributable to long-term ozon exposure. Oper Res 1995;43(1):19–27.

[6] Budnitz RJ, Apostolakis G, Boore DM, Cluff LS, Coppersmith KJ, Cornel CA, Morris PA. Use of technical expert panels: applications to probabilistic seismic hazard analysis. Risk Anal 1998;18(4):463–9.

[7] Goossens LHJ, Harper FT. Joint EC/USNRC expert judgement driven radiological protection uncertainty analysis. J Radiol Protect 1998;18(4):249–64.

[8] Cooke RM. Experts in uncertainty. Oxford: Oxford University Press; 1991.

[9] Cooke RM. Expert judgment study on atmospheric dispersion and deposition. Report Faculty of Technical Mathematics and Informatics No.01-81, Delft University of Technology; 1991.

[10] Cooke RM, Mendel M, Thijs W. Calibration and information in expert resolution. Automatica 1988;24(1):887–94.

[11] Goossens LHJ, Cooke RM, van Steen J. Final report to the Dutch Ministery of Housing, Physical Planning and Environment: on the use of expert judgment in risk and safety studies, vols. I–5. Delft; 1989.

[12] Yunusov AR, Cooke RM, Krymsky VG. Rexcalibr-integrated system for processing expert judgement: In: Goossens LHJ, editor. Proceedings of the 9th annual conference risk analysis: Blz. 587–589: facing the new millennium, Rotterdam, The Netherlands, Delft University Press; October 10–13, 1999. [ISBN: 90-407-1954-3].

[13] French S. Group consensus probability distributions: a critical survey. In: Bernardo JM, De Groot MH, Lindley DV, Smith AFM, editors. Bayesian statistics. Amsterdam: Elsevier, North Holland; 1985. p. 182–201.

[14] Genest C, Zidek J. Combining probability distributions: a critique and an annotated bibliography. Statist Sci 1986;1(1):114–1490.

[15] Ter Haar TR, Retief JV, Dunaiski PE. Towards a more rational approach of the serviceability limit states design of industrial steel structures paper no. 283. In: 2nd World conference on steel in construction, San Sebastian, Spain; 1998.

[16] Akkermans DE. Crane failure estimates at DSM. Expert judgment in risk and reliability analysis; experience and perspective. In: ESRRDA conference, Brussels, October 11, 1989.

[17] Lopez de la Cruz J. Applications of probability models and expert judgement analysis in information security. Master's thesis, TU Delft; 2004.

[18] Van Elst NP. Betrouwbaarheid beweegbare waterkeringen [Reliability of movable water barriers]. Delft University Press, WBBM report Series 35; 1997.

[19] Chou D, Kurowicka D, Cooke RM. Techniques for generic probabilistic inversion. Comput Statist Data Anal 2006, to appear.

[20] Goossens LHJ. Water pollution. TU Delft for Dutch Min. of Environment, VROM; 1994.

[21] Goossens LHJ, Cooke RM, Woudenberg F, van der Torn P. Probit functions and expert judgment. Report prepared for the Ministry of Housing, Physical Planning and Environment, the Netherlands; Delft University of Technology, Safety Science Group and Department of Mathematics, and Municipal Health Service, Rotterdam, Section Environmental Health; October 1992.

[22] Cooke RM, Jager E. Failure frequency of underground gas pipelines: methods for assessment with structured expert judgment. Risk Analysis 1998;18(4):511–27.

[23] Goossens LHJ, Cooke RM, Woudenberg F, van der Torn P. Expert judgement and lethal toxicity of inhaled chemicals. J Risk Res 1998;1(2):117–33.

[24] Goossens LHJ, Harrison JD, Harper FT, Kraan BCP, Cooke RM, Hora SC. Probabilistic accident consequence uncertainty analysis: internal dosimetry uncertainty assessment, vols. 1 and 2. Prepared for US Nuclear Regulatory Commission and Commission of European Communities, NUREG/CR-6571, EUR 16773, Washington, USA, and Brussels-Luxembourg; 1998.

[25] Brown J, Goossens LHJ, Harper FT, Haskin EH, Kraan BCP, Abbott ML, Cooke RM, Young ML, Jones JA, Hora SC, Rood A. Probabilistic accident consequence uncertainty analysis: food chain uncertainty assessment, vols. 1 and 2. Prepared for US Nuclear Regulatory Commission and Commission of European Communities, NUREG/CR-6523, EUR 16771, Washington, USA, and Brussels-Luxembourg; 1997.

[26] Goossens LHJ, Boardman J, Harper FT, Kraan BCP, Young ML, Cooke RM, Hora SC, Jones JA. Probabilistic accident consequence uncertainty analysis: uncertainty assessment for deposited material and external doses, vols. 1 and 2. Prepared for US Nuclear Regulatory Commission and Commission of European Communities, NUREG/CR-6526, EUR 16772, Washington, USA, and Brussels-Luxembourg; 1997.

[27] Harper FT, Goossens LHJ, Cooke RM, Hora SC, Young ML, Päsler-Sauer J, Miller LA, Kraan BCP, Lui C, McKay MD, Helton JC, Jones JA. Joint USNRC/CEC consequence uncertainty study: summary of objectives, approach, application, and results for the dispersion and deposition uncertainty assessment, vols. I–III. NUREG/CR-6244, EUR 15855, SAND94-1453, Washington, USA, and Brussels-Luxembourg; 1995.

[28] Cooke RM. Uncertainty in dispersion and deposition in accident consequence modeling assessed with performance-based expert judgment. Reliab Eng Syst Safety 1994;45:35–46.

[29] Van der Fels-Klerx HJ, Cooke RM, Nauta MJ, Goossens LHJ, Havelaar AH. A structured expert judgement study for a model of Campylobacter transmission during broiler chicken processing. Risk Anal 2005, to appear.

[30] Offerman J. Safety analysis of the carbon fibre reinforced composite material of the Hermes cold structure. TU-Delft/ESTEC May, Noordwijk, the Netherlands; 1990.

[31] Willems, A. Het gebruik van kwantitatieve technieken in risicoanalyses van grootschalige infrastructuurprojecten [The use of quantitative techniques in risk analysis of large infrastructural projects, in Dutch]. Min. van Verkeer en Waterstaat, DG rijkswaterstaat, Bouwdienst, Tu Delft Masters thesis, Delft; 1998.

[32] Aspinall W. Expert judgment case studies. Cambridge program for industry, risk management and dependence modeling. Cambridge: Cambridge University Press; 1996.

[33] Aspinall W, Cooke RM. Expert judgement and the Montserrat Volcano eruption. In: Mosleh A, Bari RA, editors. Proceedings of the 4th international conference on probabilistic safety assessment and management PSAM4, vol. 3, New York City, USA, September 13–18, 1998. p. 2113–8.

[34] Cooke RM, Goossens LJH. 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 Radiat Protect Dosimetry 2000;90(3):303–11.
[35] Qing X. Risk analysis for real estate investment. PhD thesis, Dept. of Architecture, TU Delft; 2002.
[36] Bakker M. Quantifying operational risks within banks according to Basel II. Masters thesis, Delft University of Technology, Dept. of Mathematics; 2004.
[37] Kurowicka D, Cooke RM. Uncertainty analysis with high dimensional dependence. New York: Wiley; 2006.
[38] Brown AJ, Aspinall WP. Use of expert opinion elicitation to quantify the internal erosion process in dams. In: Proceedings of the 13th Biennial British Dams Society conference, University of Kent, Canterbury, June 22–26, 2004. 16p.
[39] Aspinall WP, Loughlin SC, Michael FV, Miller AD, Norton GE, Rowley KC, Sparks RSJ, Young SR. The Montserrat Volcano Observatory: its evolution, organization, role and activities. In: Druitt TH, Kokelaar BP, editors. The eruption of Soufrière Hills Volcano, Montserrat, from 1995 to 1999. London: Geological Society; 2002 [Memoir].
[40] Lin, Bien, this volume.
[41] Kraan B, Bedford T. Probabilistic inversion of expert judgments in the quantification of model uncertainty. Manage Sci 2005;51(6):995–1006.
[42] Claessens M. An application of expert opinion in ground water transport, TU Delft, DSM Report R 90 8840; 1990 [in Dutch].
[43] Cooke RM, Slijkhuis KA. Expert judgment in the uncertainty analysis of dike ring failure frequency. In: Blischke WR, Prabhakar Murthy DN, editors. Case studies in reliability and maintenance. New York: Wiley; 2003. p. 331–52 [ISBN: 0-471-41373-9].
[44] De Wit MS. Uncertainty in predictions of thermal comfort in buildings. PhD dissertion, Department of Civil Engineering, Delft University of Technology, Delft; 2001.
[45] Frijters M, Cooke R, Slijkuis K, van Noortwijk J. Expert judgment uncertainty analysis for inundation probability. Bouwdienst, Rijkswaterstaat, Utrecht: Ministry of Water Management; 1999 [in Dutch].
[46] Goossens LHJ, Cooke RM, Woudenberg F, van der Torn P. Probit relations of hazardous substances through formal expert judgement. Loss prevention and safety promotion in the process industries, vol. II. Amsterdam: Elsevier Science B.V.; 1995. p. 173–82.
[47] Haskin FE, Goossens LHJ, Harper FT, Grupa J, Kraan BCP, Cooke RM, Hora SC. Probabilistic accident consequence uncertainty analysis: early health uncertainty assessment, vols. 1 and 2. Prepared for US Nuclear Regulatory Commission and Commission of European Communities, NUREG/CR-6545, EUR 16775, Washington, USA, and Brussels-Luxembourg; 1997.
[48] Meima B. Expert opinion and space debris. Technological designer's thesis, Faculty of Technical Mathematics and Informatics, Delft University of Technology, Delft; 1990.
[49] Sarigiannidis G. CARMA-Greece: an expert judgment study and the probabilistic inversion for chicken processing lines. Masters thesis, Delft University of Technology, Dept. of Mathematics; 2004.
[50] Sparks RSJ, Aspinall WP. Volcanic activity: frontiers and challenges in forecasting, prediction and risk assessment. In: Sparks RSJ, Hawkesworth CJ, editors. State of the planet: frontiers and challenges, IUGG/AGU: Geophysical Monograph Series, vol. 150; 2004. 414pp.
[51] Van Overbeek FNA. Financial experts in uncertainty. Masters thesis, Department of Mathematics, Delft University of Technology, Delft; 1999.
[52] Willems A, Janssen M, Verstegen C, Bedford T. Expert quantification of uncertainties in a risk analysis for an infrastructure project. J Risk Res 2005;8(12):3–17.