VOL. 93, NO. 5, MAY 2008

ISSN 0951-8320

RELIABILITY ENGINEERING & SYSTEM SAFETY

Special Issue
EXPERT JUDGEMENT

Guest Editors
ROGER M. COOKE

Editor-in-Chief
G. E. Apostolakis

Editor
C. Guedes Soares

Available online at
ScienceDirect
www.sciencedirect.com

Available online at www.sciencedirect.com
ScienceDirect
Reliability Engineering and System Safety 93 (2008) 745–756
RELIABILITY ENGINEERING & SYSTEM SAFETY
www.elsevier.com/locate/ress

On the performance of social network and likelihood-based expert weighting schemes

Roger M. Cooke[a,*], Susie ElSaadany[b], Xinzheng Huang[a]

[a] *Delft Institute of Applied Mathematics, The Netherlands*
[b] *Health Canada, Canada*

Abstract

Using expert judgment data from the TU Delft's expert judgment database, we compare the performance of different weighting schemes, namely equal weighting, performance-based weighting from the classical model [Cooke RM. Experts in uncertainty. Oxford: Oxford University Press; 1991.], social network (SN) weighting and likelihood weighting. The picture that emerges with regard to SN weights is rather mixed. SN theory does not provide an alternative to performance-based combination of expert judgments, since the statistical accuracy of the SN decision maker is sometimes unacceptably low. On the other hand, it does outperform equal weighting in the majority of cases. The results here, though not overwhelmingly positive, do nonetheless motivate further research into social interaction methods for nominating and weighting experts. Indeed, a full expert judgment study with performance measurement requires an investment in time and effort, with a view to securing external validation. If high confidence in a comparable level of validation can be obtained by less intensive methods, this would be very welcome, and would facilitate the application of structured expert judgment in situations where the resources for a full study are not available. Likelihood weights are just as resource intensive as performance-based weights, and the evidence presented here suggests that they are inferior to performance-based weights with regard to those scoring variables which are optimized in performance weights (calibration and information). Perhaps surprisingly, they are also inferior with regard to likelihood. Their use is further discouraged by the fact that they constitute a strongly improper scoring rule.

1. Introduction

Using expert judgment data from the TU Delft's expert judgment database, we compare the performance of different weighting schemes, namely equal weighting, performance-based weighting from the classical model [1], social network (SN) weighting and likelihood weighting.

The classical model and experience with applications to date is described in Cooke and Goossens [2]. Over the range of applications, the classical model outperforms equal weighting and best experts. However, two issues with this model emerge from that discussion, namely:

(i) The classical model is more resource intensive than simple equal weighting; is it possible to capture the advantages of differential expert weighting in a less intensive manner?

(ii) The classical model satisfies necessary conditions for rational consensus, but is not *derived* from first principles, and other weighting schemes may perform as well or better. Can other weighting schemes be implemented and evaluated using the data generated with the classical model?

SN theory was proposed as an expert rating scheme that might address issue (i) above. SN theory has been implemented using weights that are based on experts' citations. Implementing these weights requires panels of experts who publish extensively. Suitable data for comparing SN weights and performance-based weights comes from a large uncertainty analysis, the European Union and US Nuclear Regulatory Commission (EU-USNRC) on accident consequence models for nuclear power plants. This large study involved 10 panels of internationally reputed experts, of which seven involved seed or calibration variables: variables for which the true values are known post hoc. The seed variables form the basis for

*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.2007.03.017

performance-based combinations of expert judgments and also afford the possibility of comparing various combination schemes, or "decision makers" (DMs).

With regard to (ii), several suggestions have been made in recent literature, which may be tested using the classical model's data repository. One of these involves so-called "likelihood weights" [3], in which an expert's likelihood weight is proportional to the probability which s/he assigns to the observed outcomes. While these are not less resource intensive, they devolve from different lines of reasoning and are therefore of interest. The classical model data repository involves expert elicitations involving either five (five studies) or three quantiles (forty studies). The likelihood weights are most amenable for cases where the experts assessed five quantiles, and this motivates restricting the comparison to the five studies in which experts assessed five quantiles.

The classical model is reviewed in some mathematic detail in Cooke and Goossens [2]. For the purposes of this comparison, a very brief synopsis is presented in Section 1. The second section reviews the EU-USNRC data used for this comparison. The third section outlines the application of SN theory to derive expert weights, and the fourth section presents the comparative results. Section 5 discusses likelihood weights and Section 6 presents results with likelihood weights. A final conclusion draws conclusions. An appendix contains more detailed output from each panel showing the individual expert scores and the SN weights.

The overall conclusion of these comparisons is that SN and likelihood weights exhibit a performance in terms of calibration (p-value) and information that is intermediate between the performance-based weights of the classical model and equal weighting. The larger conclusion is that extensive empirical data on expert assessments with observations of assessed quantities is available to test expert combination schemes. In Kallen and Cooke [4] this data was used to test the copula method of combining experts [5]. This data is available to researchers upon request from the first author.

2. Structured expert judgment

The goal of applying structured expert judgment, as understood here, is to enhance rational consensus. Note that this is not the same as maximizing the expected utility of a rational individual. Recalling that a group of rational agents is not itself a rational agent, rational consensus is not concerned with changing the beliefs of individuals but rather with finding a representation of uncertainty to be used in a group decision context.

Necessary conditions for achieving this goal are laid down as methodological principles (see [1]):

- *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.

We claim that these are *necessary* conditions for rational consensus, we do not claim that they are sufficient as well. Hence, a rational subject could accept these and yet reject a method, which implements them. In such a case, however, s/he incurs a burden of proof to formulate additional conditions for rational consensus which the method putatively violates.

The classical model: The above principles have been operationalized in the so-called classical model, a performance-based linear pooling or weighted averaging model. The weights are derived from experts' calibration and information scores, as measured on calibration or seed variables. These are variables from the experts' field whose values become known to the experts post hoc. 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 the classical model calibration and information are combined to yield an overall or combined score with the following properties:

1. Calibration dominates over information, information serves to modulate between more or less equally well calibrated experts.
2. The score is a long run proper scoring rule, that is, an expert achieves his/her maximal expected score, in the long run, by and only by stating his/her true beliefs. Hence, the weighting scheme, regarded as a reward structure, does not bias the experts to give assessments at variance with their real beliefs, in compliance with the principle of neutrality.
3. Calibration is scored as "statistical likelihood with a cut-off". An expert is associated with a statistical hypothesis, and the seed variables enable us to measure the degree to which that hypothesis is supported by observed data. If this likelihood score is below a certain cut-off point, the expert is unweighted. The use of a cut-off is driven by property (2) above. Whereas the theory of proper scoring rules says that there must be such a cut-off, it does not say what value the cut-off should be.

4. The cut-off value for (un)weighting experts is determined by optimizing the calibration and information performance of the combination.

A fundamental assumption of the classical model (as well as Bayesian models) is that the future performance of experts can be judged on the basis of past performance, as reflected in the seed variables. Seed variables enable empirical control of any combination schemes, not just those that optimize performance on seed variables. Therefore, choosing good seed variables is of general interest, see Cooke et al. [6] for background and detail.

3. EU-USNRC expert judgment data

The expert panels in the EU-USNRC study are summarized in Table 1 below. The panel for deposited material did not involve seed variables, mainly due to time and budget constraints. The countermeasure panel was deemed too location specific to support the generation of plausible seed variables. The late health panel involved seed variables that become known with the latest analysis of Hiroshima and Nagasaki survivor data. This data has recently become available, but its analysis has been complicated by an unanticipated change of protocol in the data format and is still ongoing. Hence, there are seven panels for which seed variables are presently available.

Experts were nominated for these panels by a semi formal procedure taking account of:

- scientific publications;
- recommendations of a wide class of experts;
- experience with previous studies.

The expert judgment protocol followed in this application entails that the names of experts are published together with their rationales, but the names are not associated with either rationales or assessments in the open literature. This association is preserved to enable a competent peer review if the problem owner so desires. These names were used in determining the SN weights, but the names are not associated with assessments or scores in this study. References are given where the expert names and rationales can be retrieved.

Table 2 shows the number of variables (questions) elicited from the experts in each panel, and the number of seed variables.

4. Social network theory

The central idea of SN theory is that relations between agents in a network of social interactions are more indicative of importance/influence/value than attributes of individual agents [15]. In the scientific domain, interaction, or connectedness, may be interpreted in many ways, for example:

1. telephone and/or email traffic with colleagues;
2. visits, seminars, publications;
3. co-authorship;
4. scientific citations.

To implement SN theory as a method for determining weights for combining expert judgments, we require an index of interaction, which is meaningful and easily measured. From this point of view, scientific citations possess clear advantages.

Citation is nowadays widely recognized as the primary instrument for estimating the impact of scholarly work and is therefore chosen as our target relation in the experts' network. The weights of the experts are determined by citations between the experts themselves, in the following manner.

Table 1
Expert panels of the EC/USNRC joint project, including countermeasures[a]

Expert panel	Number of experts[b]	Year	Reference
Atmospheric dispersion	8	1993	Harper et al. [7] Cooke et al. [6]
Deposition (wet and dry)	8	1993	Harper et al. [7] Cooke et al. [6]
Behaviour of deposited material and its related doses	10	1995	Goossens et al. [8]
Foodchain on animal transfer and behaviour	7	1995	Brown et al. [9]
Foodchain on plant/soil transfer and processes	4	1995	Brown et al. [9]
Internal dosimetry	6	1996	Goossens et al. [10]
Early health effects	7	1996	Haskin et al. [11]
Late health effects	10	1996	Little et al. [12]
Countermeasures	9	2000	Goossens et al. [13]

[a] The countermeasures panel was not part of the USNRC/CEC Project, but part of the CEC follow-up project on uncertainty analysis of the COSYMA software package.

[b] The general goal of the panels was to have half of the experts coming from Europe and the other half coming from the USA. This has not been achieved in all panels for various reasons.

Table 2
Numbers of questions and seed variables questions of the expert panels of the EC/USNRC joint project, including countermeasures

Expert panel	Number of questions	Number of seeds	Remarks
Atmospheric dispersion	77	23	
Deposition (wet and dry)	87	19	14 for dry depos. 5 for wet depos.
Behaviour of deposited material and its related doses	505	0	No seed questions
Foodchain on animal transfer and behaviour[a]	80	8	
Foodchain on plant/soil transfer and processes	244	31	
Internal dosimetry	332	55	
Early health effects	489	15	
Late health effects	111	8	Post hoc values
Countermeasures	111	0	Country specific

[a]Since the practices of farming with respect to animals is different in Europe and in the USA the questionnaires were adapted for European and American experts (see Table 7).

Citation searches are carried out through Thomson ISI Web of Knowledge [v3.0].

The rules we follow when performing the searches are:

1. The weight of an expert is determined by the number of papers by the other experts in the panel, which cite him. If an expert in one paper cites 2 or more papers from another expert, we consider it as 1 citation. Thus we do not need look into every paper from an expert to find his weights.
2. If two experts co-author a paper and cite a third expert, this paper is counted twice.
3. Self-citation is excluded. In most cases the number of self-citations dominates citation from others in the expert panel.
4. We do not distinguish the order (e.g. first author, second author, etc.) of the author.

Of course there are some problems working with the citation index:

1. Names may be misspelled, or initials may be incomplete.
2. The same names may belong to different scientists, esp. for common names like "J. Brown", "P. Jacob".

One advantage of considering citation only between experts in the panel is that it largely removes these otherwise formidable problems.

One objection to citation-based weights is that it naturally favours older scientists, as they have more published work than scientists at the beginning of their careers. It would be possible to address this by counting only citations from the last N years. Of course, the choice of any particular N may drive the outcome and may be difficult to defend. We might consider a discounting procedure, but this would merely shift the discussion from the choice of N to the choice of a discount rate.

Simply counting the number of times an expert is cited measures his connectedness to the panel *as a whole*, it does not measure interactions between two given experts. Individual interactions between experts might also contain interesting information. A challenge for the future might be to find a way to integrate such information in the derivation of expert weights. The present implementation must be viewed as a first attempt to apply SN theory to the problem of expert combination.

5. Results

The results of scoring the combined experts (decision makers, DMs) in the seven panels with seed variables are shown in Table 3 below. It will be noted that in the soil/plant panel, there was not good performance on any of the DMs. This situation is unique in the annals of expert judgment, and is included here to demonstrate that good performance is not a foregone conclusion. In this case, the conclusion was that the number of experts was too small to achieve a satisfactory performance for the DM. The number beneath the panel name is the number of citations on which the analysis is based.

The performance-based DM (either global or item weights depending on the study) outperforms the others in both statistical accuracy (p-value) and relative information with respect to the background measure (Rel. inf). The SN DM outperforms the equal weight DM on four of the seven panels. In only the early health panel is the SN DM significantly less accurate statistically than the equal weight DM. Figs. 1 and 2 show the same information graphically.

Fig. 3 compares the ranks of the SN weights and the combined performance scores. The soil panel has been excluded owing to the poor performance and small number of experts. We see that in two cases (dispersion, animal) the ranks are in good agreement. In early health they are anti-correlated, and the remaining cases are indeterminate.

For four panels, we investigated the situation when the experts who weighted 0 according to citations are removed from the expert pool. This concerns selection of experts before any elicitation. The result given in Table 4 does not encourage us to conduct elicitation only among those experts with nonzero SN weights. From the case early health effects we see it might be very dangerous to do so.

Table 3
Results for social network weights, performance-based weights, and equal weights

		p-value	Rel. inf	# seeds	Combined score
Early health	SocNet	0.002176	0.2181	15	0.000475
130	Perf	0.3889	0.4345	15	0.169
	Equal	0.09153	0.167	15	0.01528
Internal dose	SocNet	0.07101	0.5997	55	0.04259
180	Perf	0.8318	0.7745	55	0.6442
	Equal	0.1125	0.5164	55	0.05812
Soil/plant	SocNet	3.08E−07	0.2489	31	7.68E−08
78	Perf	4.22E−06	0.3317	31	1.40E−06
	Equal	3.08E−07	0.2117	31	6.53E−08
Animal	SocNet	0.557	0.5123	8	0.2854
202	Perf	0.7565	1.11	8	0.8396
	Equal	0.557	0.3573	8	0.199
Wet deposition	SocNet	0.1245	0.7048	19	0.08913
37	Perf	0.2556	0.4024	19	0.1029
	Equal	0.003239	0.6491	19	0.002103
Dry deposition	SocNet	0.3992	0.1516	14	0.06051
37	Perf	0.659	0.1789	14	0.1179
	Equal	0.00169	0.1629	14	0.000275
Dispersion	SocNet	0.355	0.3483	23	0.1236
62	Perf	0.8592	0.444	23	0.3815
	Equal	0.2593	0.2467	23	0.06397

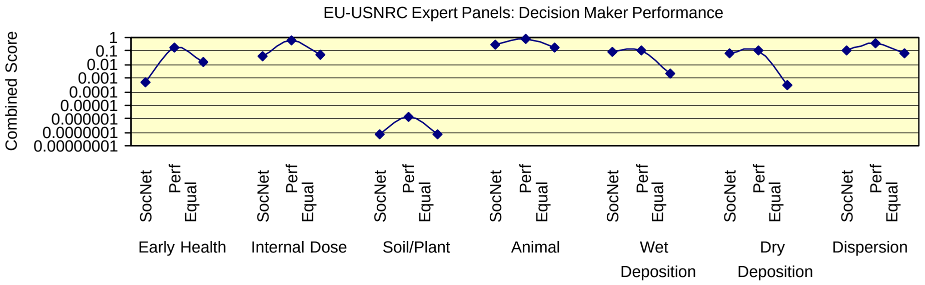

Fig. 1. Combined scores (calibration × information) for social network weights, performance-based weights, and equal weights.

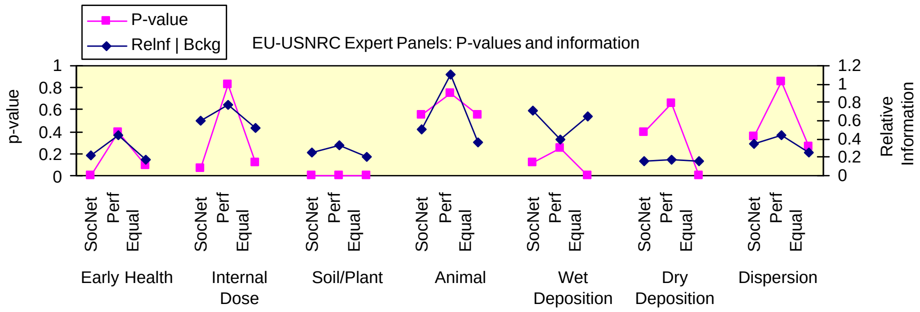

Fig. 2. *p*-values and information for social network weights, performance-based weights, and equal weights.

6. Likelihood weights

A natural suggestion for weighting experts on the basis of observed outcomes is simply to assign a weight proportional to the assessed probability of the observed outcomes. These are termed "likelihood weights". A recent suggestion of likelihood weights for Bayesian belief nets is put forward in Stiber et al. [3]. Unlike SN weights, likelihood weights require seed variables, and in this sense they are no less resource intensive than

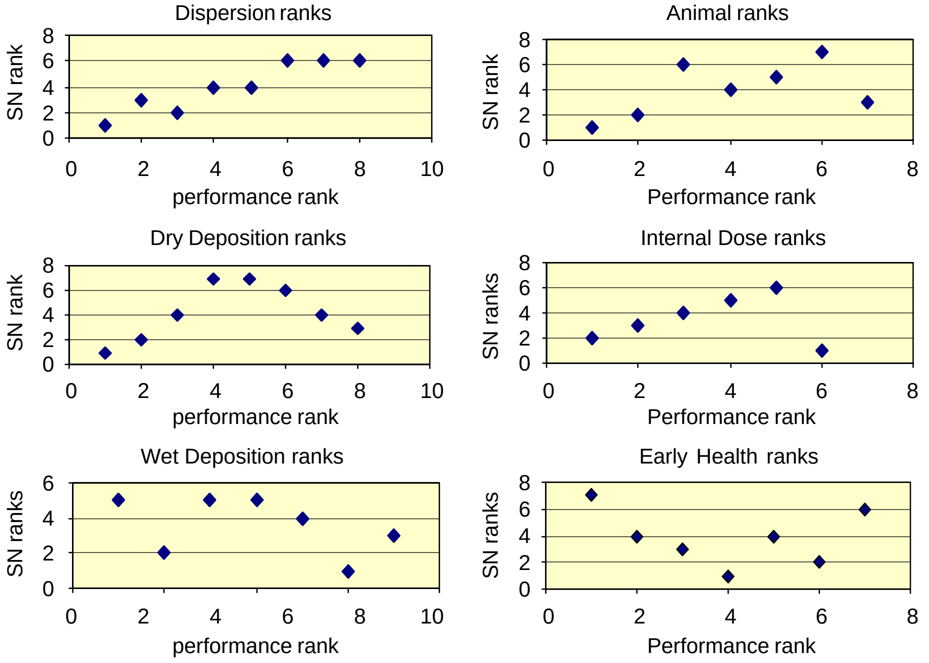

Fig. 3. Experts' performance ranks and social network weight ranks.

Table 4
Effects of removing experts with zero SN weight on calibration scores

	Number of experts removed	SN	PERF	Equal
Early health effects	0	0.002176	0.3889	0.09153
	1		0.02116	0.03462
Dispersion				
	0	0.355	0.8592	0.2593
	3		0.8592	0.1588
Dry deposition	0	0.3565	0.659	0.00169
	3		0.659	0.00169
Wet deposition	0	0.1245	0.2556	0.003239
	3		0.1701	0.05047

the classical model's performance-based weights. If likelihood weights delivered good performance with *fewer* seed variables, this would be a significant advantage. Such a claim has not been advanced, though it could be studied empirically with the methods used in the following section.

Likelihood weights constitute an improper scoring rule, sometimes called the "direct rule" [1]. Indeed, let X be an uncertain quantity with continuous range, and suppose an expert believes density function $g(x)$, and is asked to state an assessed density $f(x)$. If value x is observed, the expert receives score $K \times f(x)$ for some constant K. The expert's expected score is thus

$$\text{Expected score} = K\int f(x)g(x)\mathrm{d}x.$$

If the expert chooses f to maxmize his expected score he will evidently choose

$$f(x) = \delta(x^* - x),$$

where $x^* = \operatorname{argmax} g(x)$ and $\delta(x)$ is the Dirac function assigning unit mass to the point x. Hence, if experts are rewarded in a manner proportional to the likelihood of an observed outcome, an expert who wishes to maximize his/her expected reward is encouraged to give extremely overconfident assessments. In the same vein, one can question whether likelihood scores are reasonable measures of performance. An expert who is poorly calibrated and uninformative may nonetheless have a higher likelihood score than a well-calibrated informative expert. The AOT-AEX case discussed in the next section provides an example.

When several outcomes are observed, we interpret the likelihood of the joint observation as the product of the likelihoods of the individual observations. We thus assume that each expert regards the variables as independent. In cases where no information on dependence is assessed, there is no practical alternative but to proceed with the independence assumption.

In spite of these features, the likelihood weights continue to have an appeal, perhaps owing to the salient role of likelihood in Bayesian and classical statistics. Without contesting the proper role of theoretical disquitions, the present study focuses on performance with real expert data.

7. Results with likelihood weights

The expert data from the TU Delft database consists of quantile assessments from experts. We may implement likelihood weights in two ways, according to how we define the likelihood of the observed values. For each expert, we may either (A) define the likelihood of the observation as the probability of the interquantile interval into which the observation falls, or (B) using the minimal information density fit to the expert's quantiles, define the likelihood as the density at the observed value. To illustrate the difference between these two alternatives, suppose the value 15 is observed. Suppose expert 1 assess his 5% quantile at 10 and his 25% quantile at 20, while expert 2 assesses his 5% quantile at 10 and his 25% quantile at 50. No intermediate quantiles are assessed. On alternative (A) both experts assign the same likelihood to the observation, namely 0.2. Using a uniform background measure with alternative (B), the first expert assigns a likelihood of $0.2/10 = 0.02$; while the second expert assigns likelihood $0.2/40 = 0.005$.

Alternative (B) is more in keeping with the spirit of likelihood weights, though it requires the uniform background measure. In the TU Delft data, this measure is supplied by the analyst and not assessed by experts. Alternative (A) has been analyzed in Van Rooij [14]; which echoes the results found below. We proceed here with alternative (B). In either case, it is preferable if the experts assess a large number of quantiles. In most TU Delft studies, the experts assessed the 5%, 50% and 95% quantiles; however, in five studies the 25% and 75% quantiles were also assessed. These are (references to number in Table 2 of this volume):

1. Amsterdam Option Traders AEX (AOT-AEX), next day opening price for the AEX index [16].
2. Amsterdam Option Traders, risk analysts (AOT-Risk) [16].
3. DSM ground water transport [17].
4. Dike ring risk [18,19].
5. Health effects of fine particulate matter $PM_{2.5}$ [20].

In all cases the uniform background measure was used. Table 5 below compares the DMs based on likelihood weights, with the global (classical model) and equal weighting. In each case the calibration, average relative information and combined score (product of calibration and information scores) are shown. The full data including the expert weights are given in the Appendix. To enable the comparison with likelihood weights, the calculations are sometimes done differently than in Table 2 of Cooke and Goossens [2].[1]

Table 5
Comparison of likelihood, performance based and equal weighting

Study	Expert	Calibr'n (*p*-value)	Ave. rel. inf.	# seeds	Combined score
AOT AEX	L'hood	0.04488	0.3933	34	0.1842
	Global	0.9652	0.5224	34	0.5042
	equal	0.9769	0.2075	34	0.2027
AOT Risk	L'hood	0.8597	1.047	11	0.9005
	Global	0.8272	1.212	11	1.003
	equal	0.324	0.7449	11	0.2413
DSM grndwater	L'hood	0.08694	3.419	10	0.2972
	Global	0.7562	2.787	10	2.107
	equal	0.05891	2.895	10	0.1706
Dikering	L'hood	0.1322	0.6067	47	0.0802
	Global	0.3955	0.6462	47	0.2555
	equal	0.06979	0.7537	47	0.0526
PM2.5	L'hood	0.645	0.2132	12	0.1375
	Global	0.578	0.8065	12	0.4661
	equal	0.645	0.5421	12	0.3497

In two of the five studies (AEX, DSM) the calibration of the likelihood weights is marginally acceptable. A similar remark holds for the equal weights (for DSM, Dike ring). Fig. 4 below shows the calibration or *p*-values and information scores in graphical format. The *p*-values are shown on the left vertical axis, the average relative information with respect to the background measure on the right axis.

Although there is no theorem that global weights outperform equal weights in calibration and information, global DM does optimize for the product of calibration and information, and in practice almost always performs better. The same remark leads us to suspect better performance than likelihood weights, and this is indeed borne out in Fig. 5. It is interesting to compare these three DMs with regard to their likelihood scores. For each DM, we compute the likelihood of the realizations and, for graphical representation, normalize so that the three likelihood scores sum to one. Fig. 5 compares these likelihood scores for the three DMs, and also shows the combined score from the classical model (calibration × information).

It is notable that the DM formed using likelihood weights does *not* generally have a higher likelihood score than the other DMs. This is the case in only AOT-AEX and DSM ground water, the two studies in which the likelihood weight DM's calibration is borderline.

The overall picture is as follows. In terms of calibration and information, likelihood weights' performance is intermediate between that of the global and the equal weight DM. In terms of likelihood scores, the performance of likelihood weights is somewhat erratic.

8. Conclusions

The picture that emerges with regard to SN weights is rather mixed. Clearly, SN theory does not provide an

[1] The calibration scores in Table 5 are computed with all the seed items and without reducing the effective number of seeds [2]. The reason for this is that there is no straightforward way to perform this reduction with likelihood weights. AOT-AEX involved 38 seed variables, and 9 experts, but 4 of the experts assessed less than 34 of the seed variables. These 4 experts are excluded in this comparison. Dike ring involved 47 seed variables. In cases with a large number of seeds, the calibration scores of the experts may be very low and in such cases the effective number of seeds is often reduced to 10 to enable comparisons with other studies. These considerations explain differences between the values in Table 5 and those in Table 2 of (Cooke and Goossens 2006).

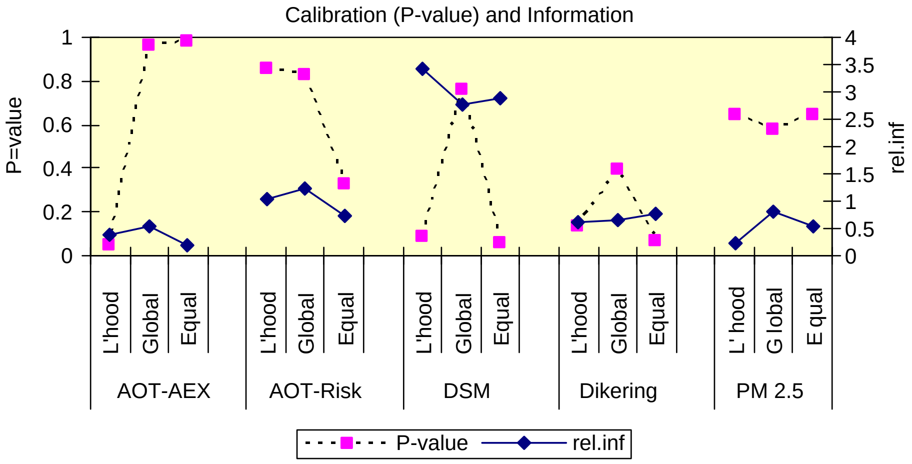

Fig. 4. Comparison of *p*-values, and relative information for likelihood, global and equal weights.

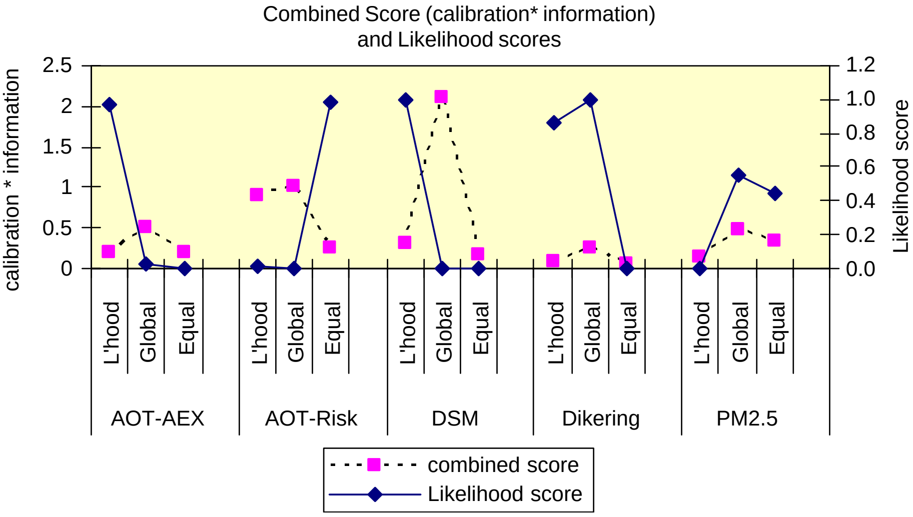

Fig. 5. Comparison of combined scores (calibration × information) and normalized likelihood scores for likelihood, global and equal weights.

alternative to performance-based combination of expert judgments. Indeed, the statistical accuracy of the SN DM is sometimes unacceptably low. On the other hand, it does outperform equal weighting in the majority of cases. In some cases the SN weights lead to a ranking of experts which is similar to their performance ranks, but this pattern is not consistent.

It might be speculated that SN theory would provide an acceptable means for nominating experts. So far as we can judge from this data, such a conclusion would not be supported.

Of course, there are many caveats to these conclusions. This represents a first attempt to derive SN weights. There are doubtless other ways of constructing such weights, based on scientific citations. Some of these were mentioned above and include:

- restricting references to the recent past;
- alternative counts for references in multi-author papers;
- using pair-wise expert interactions.

The results here, though not overwhelmingly positive, do nonetheless motivate further research into social interaction methods for nominating and weighting experts. Indeed, a full expert judgment study with performance measurement requires an investment in time and effort, with a view to securing external validation. If high confidence in a comparable level of validation can be obtained by less intensive methods, this would be very welcome, and would facilitate the application of structured expert judgment in situations where the resources for a full study are not available.

With regard to likelihood weights, the evidence presented here suggests that they do not outperform global weights either with regard to calibration and information, which are optimized in the global weights, nor indeed with regard to likelihood. If, in spite of the theoretical drawbacks noted in Section 4, one adhered to the idea that likelihood is a good measure of performance, then this study suggest that such a person could better default to equal weighting and spare himself trouble of developing seed variables.

Appendix

The following table gives the individual expert and DM scores for the EU-USNRC studies. The combined score is the product of the calibration score and the mean relative information with respect to the background for seed variables. SN denotes the social network, the SN weights are the weights assigned to the individual experts by the social network theory discussed in Section 4. SN DM in column 1 denotes the decision maker resulting from combining the experts with the SN weights (Tables A1 and A2).

Table A1

Study	Calibr'n (*p*-value)	Mean rel. inf. (seeds)	# seeds	Combined score	SN weights
Dispersion					
Exp. 1	5.23E−05	0.6418	23	3.36E−05	0
Exp. 2	7.57E−08	0.7848	23	5.94E−08	0
Exp. 3	0.001498	0.6519	23	0.000976	0
Exp. 4	0.1358	0.5574	23	0.0757	0.645
Exp. 5	0.034	0.961	23	0.03268	0.0323
Exp. 6	0.009073	0.8812	23	0.007995	0.0161
Exp. 7	0.01447	0.8404	23	0.01216	0.0161
Exp. 8	0.02151	0.6411	23	0.01379	0.2905
SN DM	0.355	0.3483	23	0.1236	
Item DM	0.8592	0.444	23	0.3815	
Global DM	0.5187	0.5254	23	0.2725	
Equal DM	0.2593	0.2467	23	0.06397	
Dry deposition					
Exp. 1	3.06E−05	0.7044	14	2.16E−05	0.081
Exp. 2	0.5274	0.1661	14	0.08759	0.405
Exp. 3	0.00169	0.41	14	0.000693	0
Exp. 4	0.00169	0.7231	14	0.001222	0
Exp. 5	2.06E−08	0.7201	14	1.48E−08	0.189
Exp. 6	0.002202	1.341	14	0.002953	0.243
Exp. 7	0.00169	0.7826	14	0.001323	0.081
Exp. 8	0.000877	0.5431	14	0.000476	0.001
SN DM	0.3992	0.1516	14	0.06051	
Item DM	0.659	0.1789	14	0.1179	
Global DM	0.5274	0.1812	14	0.09557	
Equal DM	0.00169	0.1629	14	0.000275	
Wet deposition					
Exp. 1	3.85E−10	2.254	19	8.69E−10	0.16
Exp. 2	0.01293	0.5595	19	0.007233	0
Exp. 3	0.003239	1.096	19	0.00355	0
Exp. 4	1.29E−06	1.672	19	2.15E−06	0.37
Exp. 5	0.00387	0.9804	19	0.003794	0.32
Exp. 6	0.000251	1.683	19	0.000423	0.15
Exp. 7	0.00025	1.737	19	0.000435	0
SN DM	0.1245	0.7161	19	0.08913	
Item DM	0.2556	0.4024	19	0.1029	
Global DM	0.2556	0.393	19	0.1005	
Equal DM	0.003239	0.6491	19	0.002103	
Foodchain Animal					
Exp. 1	0.002442	1.118	8	0.002732	0.025
Exp. 2	0.001995	1.15	8	0.002293	0.196
Exp. 3	0.09031	0.1564	8	0.01412	0.082
Exp. 4	0.7565	1.11	8	0.8396	0.228
Exp. 5	0.01391	1.314	6	0.01829	0.177
Exp. 6	0.6497	1.302	8	0.8461	0.247
Exp. 7	0.02528	1.272	7	0.03215	0.045
SN DM	0.557	0.5123	8	0.2854	
Item DM	0.7565	1.11	8	0.8396	
Global DM	0.7565	1.11	8	0.8396	
Equal DM	0.557	0.3573	8	0.199	
Foodchain Soil/plant					
Exp. 1	0	1.591	31	0	0.321
Exp. 2	4.96E−16	0.5205	31	2.58E−16	0.143

Table A1 (*continued*)

Study	Calibr'n (*p*-value)	Mean rel. inf. (seeds)	# seeds	Combined score	SN weights
Exp. 3	1.06E−07	0.5318	31	5.63E−08	0.321
Exp. 4	1.34E−08	0.7998	31	1.07E−08	0.215
SN DM	3.08E−07	0.2489	31	7.68E−08	
Item DM	9.53E−07	0.3972	31	3.79E−07	
Global DM	4.22E−06	0.3317	31	1.40E−06	
Equal DM	3.08E−07	0.2117	31	6.53E−08	
Internal dosimetry					
Exp. 1	0.003235	1.66	39	0.00537	0.1875
Exp. 2	0.7346	0.8151	55	0.5988	0.25
Exp. 3	1.70E−10	1.947	50	3.31E−10	0.025
Exp. 4	8.39E−17	2.363	39	1.98E−16	0.275
Exp. 5	4.55E−06	1.182	39	5.38E−06	0.0375
Exp. 6	0.009419	0.8617	28	0.008116	0.225
SN DM	0.07101	0.5997	55	0.04259	
Item DM	0.7346	0.8151	55	0.5988	
Global DM	0.8318	0.7745	55	0.6442	
Equal DM	0.1125	0.5164	55	0.05812	
Early health					
Exp. 1	0.000185	0.8381	15	0.000155	0.234
Exp. 2	0.000284	1.381	15	0.000393	0
Exp. 3	2.44E−06	1.016	15	2.48E−06	0.298
Exp. 4	0.000356	0.9652	15	0.000343	0.053
Exp. 5	1.69E−12	1.123	15	1.89E−12	0.021
Exp. 6	4.46E−05	0.5796	15	2.58E−05	0.053
Exp. 7	0.000319	0.4182	15	0.000133	0.341
SN DM	0.002176	0.2181	15	0.000475	
Item DM	0.3889	0.4345	15	0.169	
Global DM	0.3889	0.3872	15	0.1506	
Equal DM	0.09153	0.167	15	0.01528	

Table A2

Expert	Calibr'n (*p*-value)	Ave. rel. inf.	# seeds	Combined score	Likelihood weights	Global weights
AOT-AEX						
Exp. 1	0.8686	0.39	34	0.3388	2.283E−06	0
Exp. 2	0.8377	0.2166	34	0.1815	8.349E−04	0
Exp. 3	0.5538	0.4177	34	0.2313	1.261E−08	0
Exp. 4	0.9652	0.5224	34	0.5042	6.427E−01	1
Exp. 5	0.9403	0.5776	34	0.5431	3.565E−01	0
L'hood	0.04488	0.3933	34	0.1842		
Global	0.9652	0.5224	34	0.5042		
Equal	0.9769	0.2075	34	0.2027		
AOT-Risk						
Exp. 1	0.281	1.273	11	0.3577	6.74E−01	0
Exp. 2	0.8272	1.212	11	1.003	2.45E−01	1
Exp. 3	0.1609	1.446	11	0.2327	6.41E−05	0
Exp. 4	0.08609	1.063	11	0.09155	2.08E−03	0
Exp. 5	0.4949	1.451	11	0.718	7.88E−02	0
L'hood	0.8597	1.047	11	0.9005		
Global	0.8272	1.212	11	1.003		
Equal	0.324	0.7449	11	0.2413		
DSM gr						
Exp. 1	0.000139	4.445	10	0.0006161	3.253E−14	0
Exp. 2	0.000697	3.905	10	0.002721	1.191E−02	0
Exp. 3	0.44	3.802	10	1.673	9.020E−02	0.74
Exp. 4	1.27E−11	6.217	10	7.87E−11	7.165E−28	0
Exp. 5	0.1466	1.704	10	0.2498	1.952E−04	0.11

Table A2 (*continued*)

Expert	Calibr'n (*p*-value)	Ave. rel. inf.	# seeds	Combined score	Likelihood weights	Global weights
Exp. 6	0.007621	4.831	10	0.03681	9.561E−06	0
Exp. 7	0.08694	3.797	10	0.3301	8.977E−01	0.15
L'hood	0.08694	3.419	10	0.2972		
Global	0.7562	2.787	10	2.107		
Equal	0.05891	2.895	10	0.1706		
Dikering						
Exp. 1	1.47E−05	1.093	47	1.61E−05	3.519E−06	0
Exp. 2	1.30E−05	1.254	47	1.63E−05	9.079E−10	0
Exp. 3	0.000144	0.8015	47	0.0001153	9.690E−02	0
Exp. 4	1.56E−08	1.46	47	2.28E−08	8.754E−19	0
Exp. 5	2.04E−11	1.572	47	3.21E−11	1.200E−18	0
Exp. 6	0.0341	0.4371	47	0.0149	1.111E−09	0
Exp. 7	5.28E−15	0.9633	47	5.09E−15	1.412E−28	0
Exp. 8	4.81E−05	1.061	47	5.10E−05	3.389E−12	0
Exp. 9	3.83E−11	1.403	47	5.38E−11	8.047E−14	0
Exp. 10	0.3955	0.6462	47	0.2555	9.031E−01	1
Exp. 11	3.09E−18	2.133	47	6.59E−18	2.965E−24	0
Exp. 12	3.25E−19	2.471	47	8.04E−19	7.753E−27	0
Exp. 13	6.78E−08	1.531	47	1.04E−07	4.953E−12	0
Exp. 14	0	2.065	47	0	3.793E−35	0
Exp. 15	6.49E−08	1.24	47	8.05E−08	1.864E−11	0
Exp. 16	0.001114	0.8198	47	0.000913	1.086E−09	0
Exp. 17	3.27E−09	1.111	47	3.64E−09	2.757E−12	0
L'hood	0.1322	0.6067	47	0.0802		
Global	0.3955	0.6462	47	0.2555		
Equal	0.06979	0.7537	47	0.0526		
PM2.5						
Exp. 1	0.000508	1.68	12	0.0008531	3.02E−06	
Exp. 2	0.1195	1.486	12	0.1776	9.94E−02	0.9
Exp. 3	0.08127	0.8755	12	0.07115	7.77E−02	
Exp. 4	0.08554	0.2331	12	0.01994	7.24E−01	0.1
Exp. 5	2.90E−05	2.673	12	7.74E−05	7.46E−02	
Exp. 6	0.000634	1.244	12	0.0007879	2.45E−02	
L'hood	0.645	0.2132	12	0.1375		
Global	0.578	0.8065	12	0.4661		
Equal	0.645	0.5421	12	0.3497		

References

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

[2] Cooke RM, Goossens LHJ. TU Delft expert judgment data base, 2007, this issue, doi:10.1016/j.ress.2007.03.005.

[3] Stiber NA, Small MJ, Pantazidou M. Site-specific updating and aggregating of Bayesian belief network models for multiple experts. Risk Anal 2004.

[4] Kallen MJ, Cooke RM. Expert aggregation with dependence. In: Bonano EJ, Camp AL, Majors MJ, Thompson RA, editors. Probabilistic safety assessment and management. Amsterdam: Elsevier; 2002. p. 1287–94.

[5] Clemen RT, Jouini MN. Copula models for aggregating expert opinions. Oper Res 1996;44(3):444–57.

[6] Cooke RM, Goossens LHJ, Kraan BCP. Methods for CEC/USNRC accident consequence uncertainty analysis of dispersion and deposition—performance based aggregating of expert judgements and PARFUM method for capturing modelling uncertainty. Report EUR 15856, 1995.

[7] Harper FT, Goossens LHJ, Cooke RM, Hora SC, Young ML, Päsler-Sauer J, et al. Probabilistic accident consequence uncertainty analysis: dispersion and deposition uncertainty assessment. Report NUREG/CR-6244, EUR 15855, Washington, DC/USA, Brussels-Luxembourg; 1995.

[8] Goossens LHJ, Boardman J, Harper FT, Kraan BCP, Cooke RM, Young ML, et al. Probabilistic accident consequence uncertainty analysis: external exposure from deposited material uncertainty assessment. Report NU-REG/CR-6526, EUR 16772, Washington, DC/USA, Brussels-Luxembourg; 1997.

[9] Brown J, Goossens LHJ, Harper FT, Kraan BCP, Haskin FE, Abbott ML, et al. Probabilistic accident consequence uncertainty analysis: food chain uncertainty assessment. Report NUREG/CR-6523, EUR 16771, Washington, DC/USA, Brussels-Luxembourg; 1997.

[10] Goossens LHJ, Harrison JD, Harper FT, Kraan BCP, Cooke RM, Hora SC. Probabilistic accident consequence uncertainty analysis: internal dosimetry uncertainty assessment. Report NUREG/CR-6571, EUR 16773, Washington, DC/USA, Brussels-Luxembourg, 1998.

[11] Haskin FE, Harper FT, Goossens LHJ, Kraan BCP, Grupa JB, Randall J. Probabilistic accident consequence uncertainty analysis: early health effects uncertainty assessment. Report NUREG/CR-6545, EUR 16775, Washington, DC/USA, Brussels-Luxembourg; 1997.

[12] Little M, Muirhead C, Goossens LHJ, Harper FT, Kraan BCP, Cooke RM, et al. Probabilistic accident consequence uncertainty analysis: late (somatic) health effects uncertainty assessment. Report NUREG/CR-6555, EUR 16774, Washington, DC/USA, Brussels-Luxembourg; 1997.

[13] Goossens LHJ, Kraan BCP, Cooke RM, Jones JA, Ehrhardt J. Probabilistic accident consequence uncertainty analysis using COSYMA: countermeasures uncertainty assessment. Report EUR 18821, Brussels-Luxembourg; 2001.

[14] Van Rooij MM. Performance of expert judgment methods with expert modeling. Masters thesis, Department of Mathematics, Delft University of Technology, Delft, April 15, 2005.

[15] Hanneman RA. Introduction to social network method, 2001.

[16] Van Overbeek FNA. Financial experts in uncertainty. Masters thesis, Department of Mathematics, Delft University of Technology, Delft; 1999.

[17] Claessens M. An application of expert opinion in ground water transport (in Dutch). DSM Report R 90 8840, TU Delft, 1990.

[18] Frijters M, Cooke RM, Slijkuis K, van Noortwijk J. Expert judgment uncertainty analysis for inundation probability (in Dutch). Ministry of Water Management, Bouwdienst, Rijkswaterstaat, Utrecht; 1999.

[19] Cooke RM, Slijkhuis KA. Expert judgment in the uncertainty analysis of dike ring failure frequency. In: Wallace R, Blischke DN, Prabhakar M, editors. Case studies in reliability and maintenance. New York: Wiley; 2003. p. 331–52 ISBN: 0-471-41373-9.

[20] Tuomisto JT, Wilson A, Evans JS, Tainio M. Uncertainty in mortality response to airborne fine particulate matter: combining European air pollution experts, 2007, this issue, doi:10.1016/j.ress.2007.03.002.