PSAM 4 Proceedings, (Moselh and Bari eds) Springer 1998,1937-1942)

EVALUATION OF WEIGHTING SCHEMES FOR EXPERT JUDGMENT STUDIES

Goossens, L.H.J., Cooke R.M., Kraan, B.C.P
Delft University of Technology

1. Introduction

Experts judgment is increasingly recognized as a valuable source of scientific data. Like any scientific measurement, the acquisition, use, and validation of expert judgment data must proceed in a traceable way according to rigorous methodological rules. That having been said, the exact nature of these methodological rules is a subject of ongoing scientific discussion.

Delft University of Technology under contract to the EU and in cooperation with other EU institutes, is completing a study with the United States Nuclear Regulatory Commission (Harper et al 1995) using formal expert judgment methods for retrieving uncertainty distributions over the major parameters in accident consequence models for nuclear power plants. Expert assessments were aggregated to yield one combined uncertainty distribution over each assessed variable of interest. Two methods of combining expert assessments were applied: equal and performance based weighting schemes. The pro's and con's of different weighting schemes remain a subject of research. For this reason, there was a need in parallel to undertake a more formal analysis of the merits of different schemes. This resulted in review (Goossens et al 1996) of applications both within the EU-USNRC project, and elsewhere, in which seed variables have been applied in the field of 'Technological risk'. Studies in other areas (project risk, financial risk) and studies undertaken for academic research only are excluded. The list of eligible studies is updated below:

1. Crane risk (DSM in collaboration with TU Delft, Akkermans 1989)
2. Space debris (TU Delft for the European Space Agency, Meima 1990)
3. Safety analysis composite materials (European Space Agency in collaboration with TU Delft, Offerman 1990)
4. Groundwater transport (DSM chemical plant in collaboration with TU Delft, Claessens, 1990)
5. Atmospheric Dispersion and Deposition (TU Delft for EU as a pilot project for EU-USNRC joint study, Cooke 1991A, replicated by TNO with independent experts, Cooke 1994)
6. Dose response relations for hazardous substances (TU Delft for Dutch Ministery of Environment, Goossens et al 1992)
7. Water Pollution (TU Delft for Dutch Min. of Environment, VROM 1994)
8. EU-USNRC dispersion and deposition modules (TU Delft and SANDIA for EU and USNRC, Harper et al 1995)
9. Failure of underground gas pipelines (TU Delft for the Dutch Gasunie, Cooke et al 1996, Cooke and Jager, 1998)
10. Failure of moveable water barriers (Dutch Ministery of Water Management in collaboration with TU Delft, van Elst 1997)

11. Safety factors for airline pilots (Aspinall and Associates for British Air, Aspinall 1996)
12. Expert Judgment at Montserrat (Aspinall and Associates for governor Montserrat, Aspinall 1996)
13. Expert Judgment for serviceability limit states (Ter Haar et al 1998).

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 the last three studies this was not possible. Some studies involved multiple expert panels.

2. Structured expert judgment

The goal of applying structured expert judgment techniques is to enhance rational consensus. Necessary conditions for achieving this goal are laid down as methodological principles.

> **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.

These 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 performance, as measured on calibration or seed variables. The name “classical model” derives from a strong analogy between calibration measurement and classical statistical hypothesis testing and is contrasted with Bayesian models.

The performance based weights use two quantitative measures of performance, calibration and information. The former requires the use of calibration or seed variables; variables whose true values are unknown to the experts at the time of the elicitation, but whose values are known post hoc. Sometimes calibration variables will be ‘near future’ versions of the variables of interest, and will be observed within the time frame of the study. More often, calibration variables are not themselves variables of interest, but are included in the elicitation to enable performance based weighting. The designation “seed” variables is then suggested by their role in ‘seeding’ the combination model. 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.

Calibration measures the statistical likelihood that actual experimental results correspond, in a statistical sense, with the experts assessments. Information represents the degree to which an expert’s is distribution is concentrated, relative to some user-selected background measure. "Good expertise" corresponds to good calibration (high statistical likelihood) and high information. The weights in the classical model are proportional to the product of statistical likelihood and information. For more detail see (Goossens et al 1989, Cooke 1991A).

Note that seed variables are needed to enable empirical control of combination schemes, regardless whether these schemes themselves use seed variables in deriving weights. Hence, interest in the best way for choosing seed variables and how to evaluate performance on the basis of seed variables is not confined to practitioners using performance based combination schemes. In the authors' view, seed variables are essential for satisfying the criteria for rational consensus, regardless of the scheme used for combining expert judgments.

In the classical model calibration and information are combined to yield an overall or combined score (of the 'decision maker', DM) 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.
3. Calibration is scored as 'statistical likelihood with a cut-off'. An expert is associated with a statistical hypothesis, and the calibration 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 ther 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 (DM).

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, reflected in the seed variables. The performance of the experts on the seed variables is taken as indicative for the performance on the variables of interest. The choice of these seed variables is therefore critical.

3. Review of applications

Table 1 below presents information on numbers of experts, items, seed items, and the performance of the best expert, the equal weight DM and the performance based DM. Seed variables are distinguished according to their affinity to the variables of interest. "Domain variables" have the same physical dimensions as the variables of interest. They represent measurements of past realizations, or 'near field' realizations. "Adjacent variables" are of different dimension from the variables of interest, but are drawn from the experts' relevant knowledge base. They represent variables about which experts should be able to give an 'educated guess'. Some studies involved more than one expert panel. All studies with the exception of 2 and 10 involved experts with university training. Studies 8a and 8b involved extensive training in subjective probability assessment; the other studies involved only cursory training.

Case	#experts	#vbls/#seed	dom/adj		perform weights	equal weights	best expert
1	8	39/12	adj	calibr’n	0.84	0.5	0.005
Crane risk				inform’n	1.367	0.69	2.458
				combi’n	**1.148**	**0.345**	**0.012**
2	7	58/26	dom	calibr’n	0.78	0.9	0.0001
Space				inform’n	0.32	0.15	2.29
debris				**combi’n**	**0.25**	**0.14**	**0.0003**
3	6	22/12	dom	calibr’n	0.27	0.12	0.005
Composite				inform’n	1.442	0.929	2.549
materials				**combi’n**	**0.4**	**0.111**	**0.013**
4	7	48/10	dom	calibr’n	0.7	0.05	0.4
Grndwater				inform’n	3.008	3.16	3.966
transport				**combi’n**	**2.106**	**0.158**	**1.586**
5a	11	91/36	dom	calibr’n	0.68	0.71	0.36
dispersion				inform’n	0.827	0.715	1.532
panel TUD				**combi’n**	**0.562**	**0.508**	**0.552**
5b	11	91/36	dom	calibr’n	0.69	0.32	0.53
dispersion				inform’n	0.875	0.751	1.716
panel TNO				**combi’n**	**0.604**	**0.24**	**0.909**
5c	4	56/24	dom	calibr’n	0.45	0.34	0.45
dry				inform’n	1.647	1.222	1.647
deposition				**combi’n**	**0.741**	**0.415**	**0.741**
6a	7	43/10	adj	calibr’n	0.24	0.28	0.24
acrylo-				inform’n	3.186	1.511	3.186
nitrile				**combi’n**	**0.764**	**0.423**	**0.764**
6 b	6	28/10	adj	calibr’n	0.11	0.28	0.06
ammonia				inform’n	1.672	1.075	2.627
panel				**combi’n**	**0.184**	**0.301**	**0.158**
6 c	4	28/6	adj	calibr’n	0.14	0.14	0.02
sulphur tri				inform’n	3.904	2.098	4.345
oxide				**combi’n**	**0.547**	**0.294**	**0.087**
7	11	21/11	adj	calibr’n	0.35	0.35	0.16
water				inform’n	1.87	1.75	2.76
pollution				**combi’n**	**0.66**	**0.48**	**0.33**
8a	8	74/23	dom	calibr’n	0.9	0.16	0.13
dispersion				inform’n	1.087	0.862	1.242
panel				**combi’n**	**0.978**	**0.129**	**0.161**
8 b	8	56/14	dom	calibr’n	0.52	0.001	0.52
dry				inform’n	1.339	1.184	1.339
deposition				**combi’n**	**0.697**	**0.001**	**0.697**
9a	15	48/28	both	calibr’n	0.93	0.11	0.06
environm.				Inform’n	1.628	1.274	2.411

panel				**combi'n**	**1.514**	**0.14**	**0.145**
9b	12	58/11	both	calibr'n	0.7	0.06	0.2
corrosion				inform'n	1.219	1.304	2.762
panel				**combi'n**	**0.853**	**0.078**	**0.552**
10	8	35/15	adj	calibr'n	0.43	0.22	0.04
moveable				inform'n	1.234	0.57	1.711
barriers				**combi'n**	**0.531**	**0.125**	**0.068**

Table 1

To appreciate the numbers in Table 1 we must take into account the numbers of seed variables, the numbers of experts, and the robustness of the results against seed variables and experts. Only gross differences in calibration should be regarded as significant; changes by a factor 2 or 3 may arise in performing robustness on seed variables. Information scores are more stable, and a difference of a factor 2 is usually robust. Information scores cannot be compared across studies.

4. Conclusion

In spite of the above caveats we hazard some conclusions. In a number of studies (1, 2, 3, 6b, 6c 9a, 10) we see what might be called the "overconfident experts" pattern: the best expert is much more informative and much less well calibrated than the performance-based and equal weight decision makers. The latter two are roughly equivalent in terms of calibration but the performance based decision maker is more informative.

There are significant departures from this pattern however. In some cases a small number of experts succeeds in combining high information with good calibration. In these cases the performance based DM gives high weight to these experts, and this can lead to large differences between the performance based and equal weight decision makers (4, 8a, 8b, 9a, 9b). In other cases the high performance experts are sufficiently representative of the whole group that the equal weight decision maker is not dramatically worse than the performance based decision maker, (5a, 5b, 5c 6c, 7) and maybe a little better (6b). Only in 5b is the best expert better than the performance based DM.

References

Aspinall W., Expert judgment case studies, Cambridge Program for Industry, Risk management and dependence modeling, Cambridge 1996

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

Cooke, R.M. Experts in Uncertainty, Oxford University Press, Oxford, 1991A.

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

1991B.

Cooke RM. Uncertainty in dispersion and deposition in accident consequence modeling assessed with performance-based expert judgment, Reliability Engineering and System Safety, 1994, Vol.45, pp.35-46.

Cooke, R.M. and Jager, E., Failure frequency of underground gas pipelines: methods for assessment with structured expert judgment, appearing in Risk Analysis, 1998.

Goossens LHJ, Cooke RM, and 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 1996.

Goossens LHJ, Cooke RM, Woudenberg F and 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.

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. Prepared for U.S. Nuclear Regulatory Commission and Commission of European Communities, NUREG/CR-6244, EUR 15855, Washington/USA, and Brussels-Luxembourg, 1995 (Volumes I, II, III).

Meima B., Expert opinion and space debris, Technological Designer's Thesis, Faculty ot Technical Mathematics and Informatics, Delft University of Technology, Delft, 1990.

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

Ter Haar, T.R., Retief, J.V. and Dunaiski, P.E. Towards a more rational approach of the serviceability limit states design of industrial steel structures paper no. 283, 2end World conference on steel in construction, San Sebastian, Spain, 1998.

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