BRISK / Cabot Institute Summer School on Uncertainty
Bristol: 15 - 19 July 2013

EXPERT ELICITATION

Willy Aspinall

University of BRISTOL | Cabot Institute

brisk
Bristol environmental risk research centre

Three elicitation and opinion pooling methods will be introduced:

The Cooke Classical Model and EXCALIBUR software

The new Expected Relative Frequency Model

Paired comparison with Probabilistic Inversion analysis

Vesuvius, and the threat to Naples

36 WORLD NEWS 1·23

New computer study shows widespread devastation in final phase of eruption

5mins after it begins
Column of red-hot ash, gas and rock collapses on to mountain, blasting towards populated areas

15mins
Pyroclastic flow sweeps through homes built illegally on slopes

20mins
Flows burst through towns from shire line to Pompeii

AD79 eruption: plaster casts of victims who were burnt to death are now a tourist attraction

Magma erupts into the sky at 400mph

VESUVIUS

Illicit homes sprawl down slopes

Magma travels at 60 mph

Up to 300,000 could be killed

TYRRHENIAN SEA

"Today people aren't prepared for an evacuation. It's wrong to bet on carrying out the entire evacuation in just three to five days. If there was an eruption tomorrow it's quite possible we'd see huge traffic jams, car crashes and people using guns to make their escape," he said. "We need a plan to assist the evacuation of vulnerable people, like pensioners, first. People need to know what to do and if they have to understand they won't be able to get out by car. You can walk 10 miles in three hours, that's quite a good safety margin."

Some Neapolitans are hoping Vesuvius will stir soon. Guides who work near its summit believe a volcano that splutters, grows and occasionally spits fire and lava is a better crowd-puller than a dull, impo-

THE SUNDAY TIMES MARCH 4, 2007

Vesuvius blast could kill 300,000

John Follain

THE next eruption of Vesuvius could kill at least 300,000 people, nearly 20 times as many as the AD79 disaster that buried the ancient city of Pompeii, according to Italian government research.

More than half a million people live in the so-called "red zone" of 18 towns in a four-mile radius of the volcano and most would die if an evacuation could not be completed in time, the research says.

The findings are from a study by some of Europe's leading vulcanologists and public health experts, including Dr Peter Baxter of Cambridge University's Department of Public Health.

The destruction of Pompeii, the worst affected city, has inspired many books and films, including Robert Harris's 2003 bestseller, which features Pliny the Elder and which is to be adapted by the director Roman Polanski in a £100m movie.

Some 2.5m tourists visited Pompeii last year, where the people by bus from each of th 18 towns.

Professor Giuseppe Luong of the University of Naples, former director of the Vesuviu Observatory which monitors th volcano, believes plans a inadequate and local people a ill-informed about them.

Alternative approaches to pooling expert opinions:

simple averaging

committee

decision conferencing

the Delphi method

expert self-weighting

mathematical theory of scoring rules

> **Cooke's "Classical" model for pooling opinions and implementation in the EXCALIBUR program**

Cooke R.M. *Experts in Uncertainty*, Oxford University Press (1991).

STRUCTURED EXPERT ELICITATION: GOALS

The process by which experts come to agreement sensu stricto in science is the scientific method itself. Whilst expert judgments can be regarded as scientific data, a structured expert elicitation formalism cannot pre-empt the scientific method, and therefore cannot have enforced agreement as a valid scientific goal.

Following, loosely, Cooke and Goossens (2008), there are three broadly different goals for which a structured judgment method may be undertaken, in a decision-support role:

- **To arrive at an administrative or political consensus (compromise) on scientific issues**
- **To provide a census of scientists' views**
- **To develop a rational evidence-based consensus on some particulars of a scientific issue of concern**

Rational consensus

Rational consensus refers to a group decision process, as opposed to a group census or consensus procedure. The participants agree on a method by which the 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 the result as his personal degree of belief.

To be rational, this method must comply with necessary generic conditions devolving from the scientific method. Cooke (1991) formulates the necessary conditions or principles, which any method warranting the designation "scientific" should satisfy, as:

Scrutability/accountability: all data, including experts' names and assessments, and all processing tools are available for peer review and results must be open and 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' competencies are not pre-judged, prior to processing the results of their assessments.

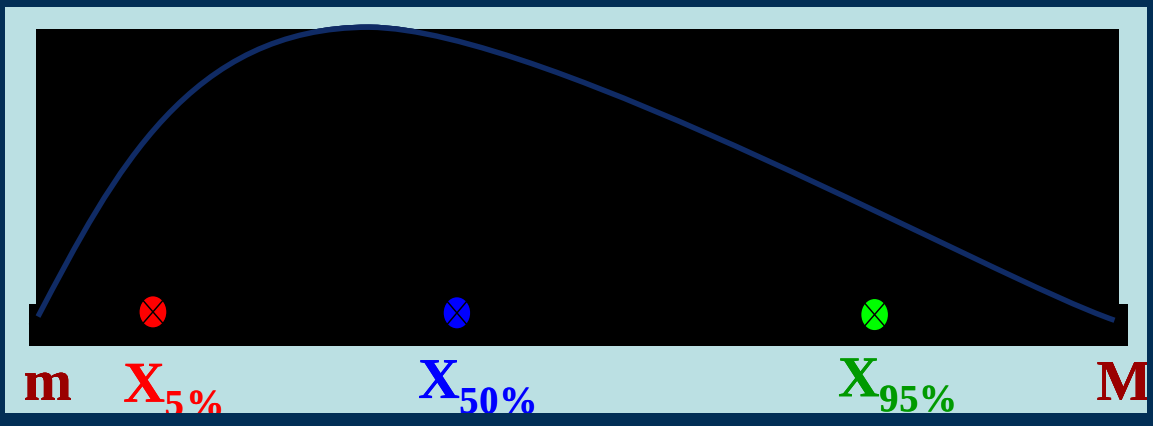
The basis of Cooke's "classical" model

Given a set of known (or knowable) seed items, for each expert calculate his:

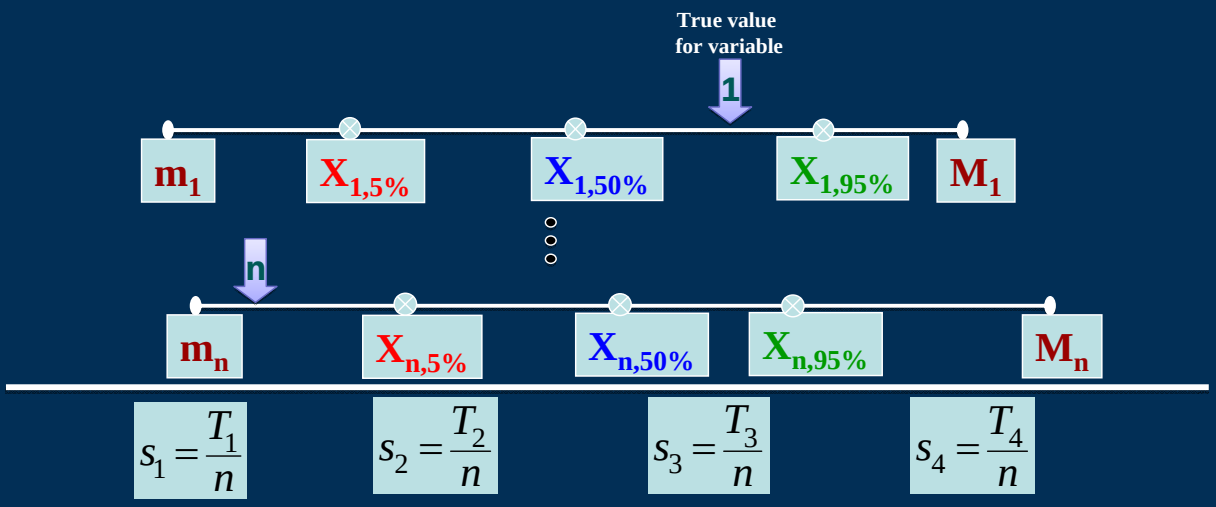
- **Calibration score**
  test hypothesis H_0: "This expert is well calibrated", leading to likelihood of acceptance at some defined significance level
- **Entropy or informativeness score**
  estimate individual's information distribution relative to a uniform distribution
- **Weighting**
  compute individual's (normalised) weight from product
  Calibration * Entropy

The basis of Cooke's "classical" model

For every item, each expert gives his/her estimates of (three or more) quartiles. The pooling algorithm defines an intrinsic range [m, M] to span the group's responses



Calibration over a number of seed items



T_i is the number of times the true values lie in the i^{th} interval out of n

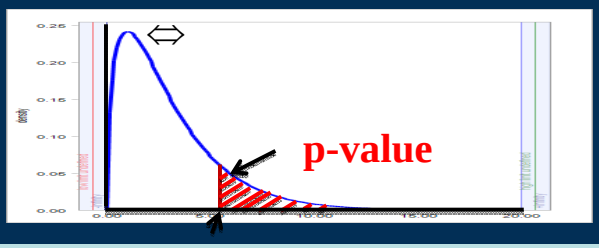
$$I(s \mid \text{True Probabilities}) = s_1 \ln(\frac{s_1}{.05}) + s_2 \ln(\frac{s_2}{.45}) + s_3 \ln(\frac{s_3}{.45}) + s_4 \ln(\frac{s_4}{.05})$$

Calibration (Cont'd)

$$I(s \mid \text{True Probabilities}) = s_1 \ln(\frac{s_1}{.05}) + s_2 \ln(\frac{s_2}{.45}) + s_3 \ln(\frac{s_3}{.45}) + s_4 \ln(\frac{s_4}{.05})$$

Hypothesis: Expert is well Calibrated H_0: (S1,S2,S3,S4)= (.05,.45,.45,.05)

Density of Chi-square distribution with 3 degrees of freedom

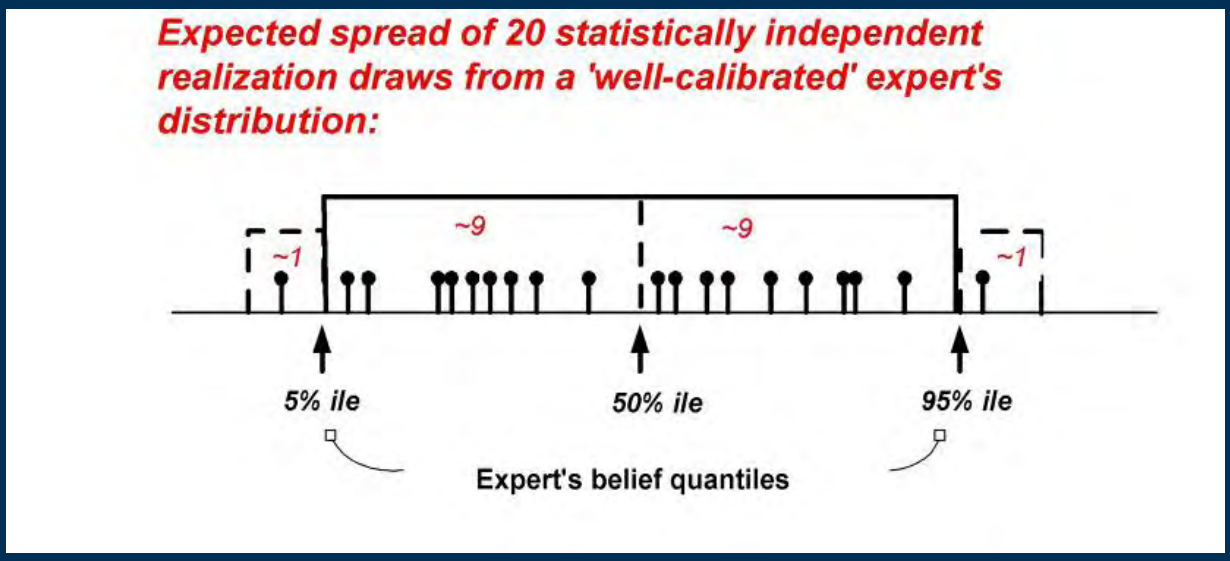


$$C_j = 1 - \chi_R^{\,2}(2 * M * I(s_j, p) * Power)$$

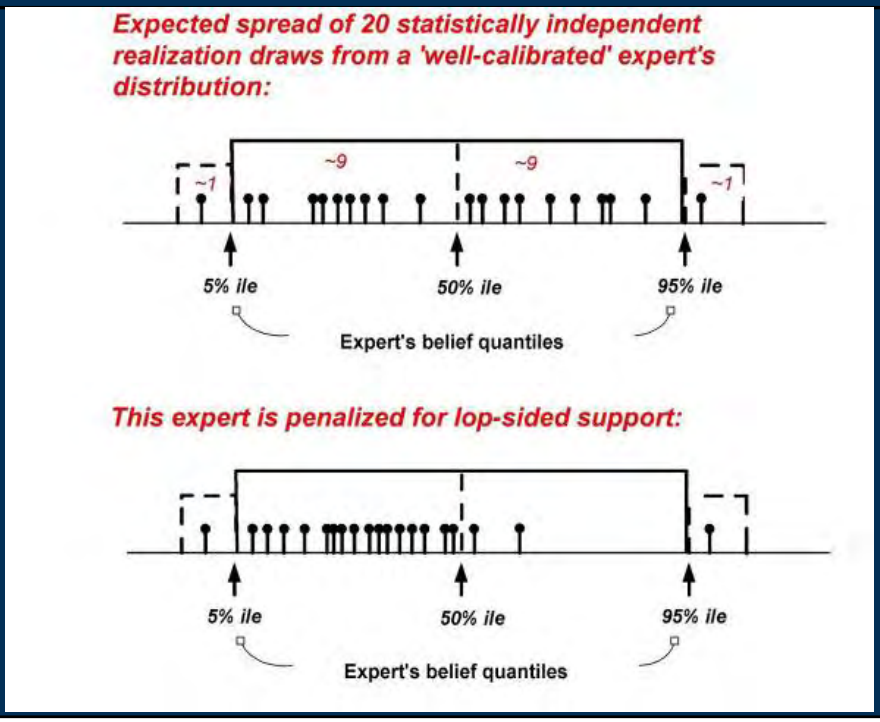
…where *j* denotes the expert, *R* is no. of quantiles (= degrees of freedom), *M* is the number of seed variables used in calibration, and *I(s,p)* is a measure of information.

***Cj* corresponds to the asymptotic probability of seeing a deviation between *s* and *p* at least as great as *I(s,p)*, under the hypothesis.**

Basis of null hypothesis for Classical Model probability calibration – a well-calibrated (statistically accurate) expert



"Poor" performance



Second penalty score: "Informativeness" or Entropy

- **Entropy score**
  estimate individual's information score relative to a uniform or loguniform density function from:

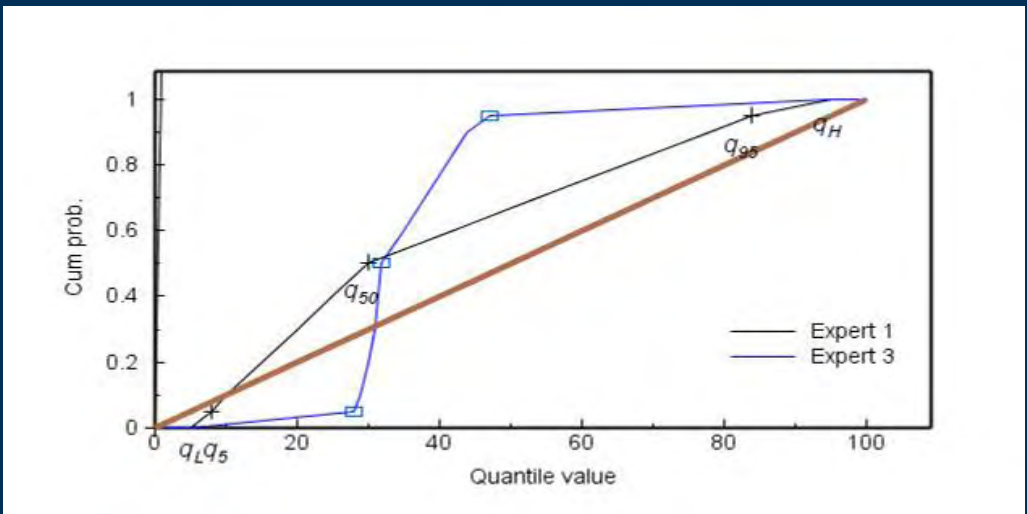
$$I_j(s_j, p) = \frac{1}{n}\sum_{i=1}^{n} s_i \ln\left(\frac{s_i}{p_i}\right)$$

where s_i is a sample distribution obtained from the expert on the seed variables, and p_i is a suitable reference density function, depending on the appropriate scaling for the item.

How concentrated are the expert's uncertainty distributions?

Entropy score

Two experts' item distributions relative to a uniform background distribution (brown line):



Expert 3 (blue squares) is more informative on this item

Expert weighting

- **Individual's expert weighting**

compute individual's weight from product of his Calibration and Entropy scores (where the latter is now estimated from all variables, seeds and unknowns):

$$W_j = C_j * I_j(s_j, p)$$

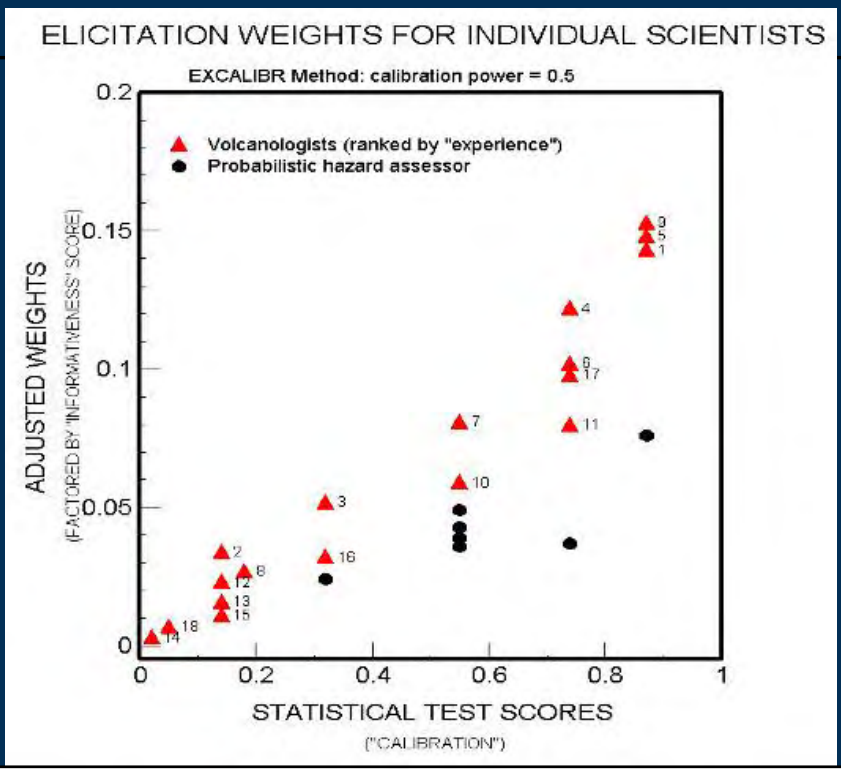
and normalise the *Wj* across all experts to get relative weights.

Classical model weighting profile

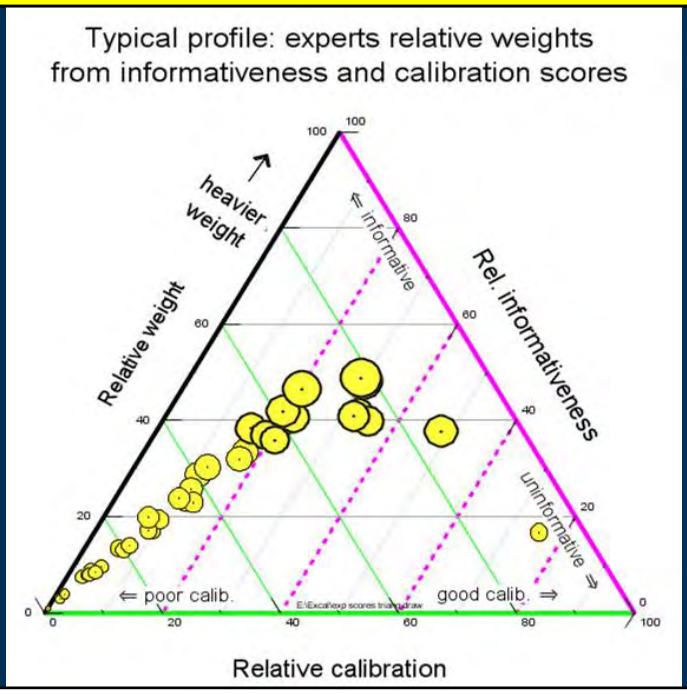
ranking experts by their performance-based weights - typical expert group "profile"

note: experience / reputation is no guarantee of a good score

x-axis is individual's statistical accuracy score; y-axis is the same factored by informativeness score, and normalized for group to sum to unity



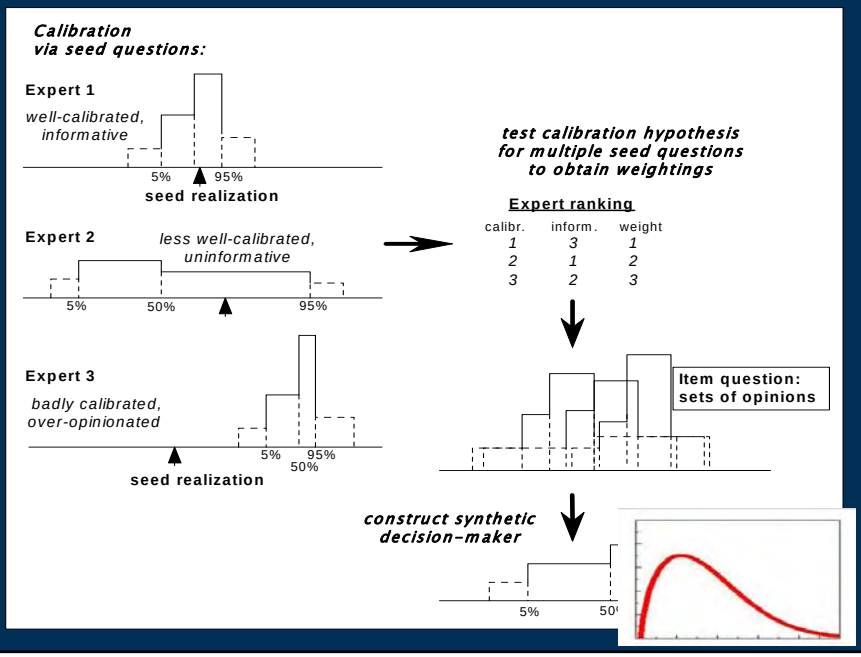
Expert weights from the Classical Model



Pooling experts' distributions with weights

An rational consensus on any question of interest Q_i can then be obtained from the weighted combination of the distributions representing the opinions about Q_i of a group of experts:

$$DM_i = \Sigma\, W_j * Q_i$$

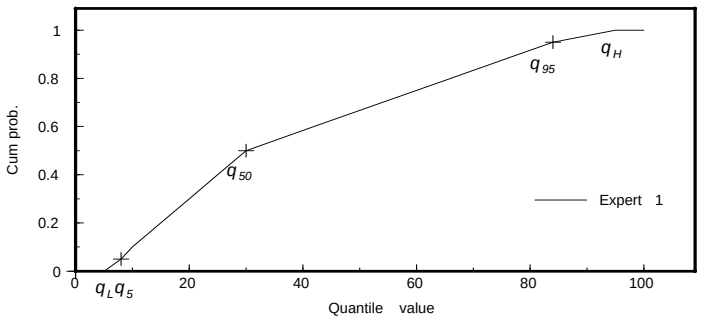


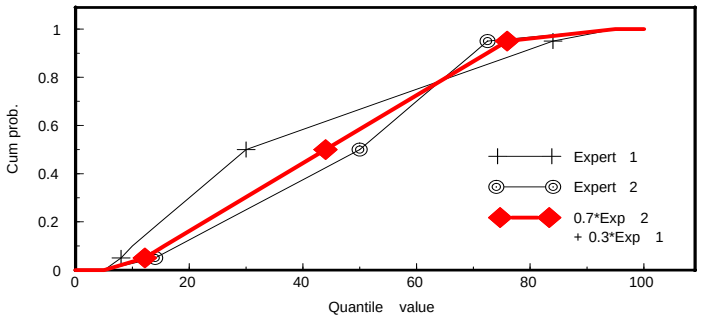
POOLING EXPERTS

Upper panel: simple representation of an interpolated distribution of quantiles for one expert. With suitable overshoot adjustment, q_L and q_H define the intrinsic range (from the extreme quantile values provided by all experts).

The distribution of Expert 1 is approximated by linear interpolation over the quantile information (q_L , 0), (q_5 , 0.05), (q_{50} , 0.5), (q_{95} , 0.95), and (q_H , 1) i.e. with minimum information with respect to the uniform distribution on the intrinsic range which satisfies this expert's quantiles.

Lower panel: a weighted combination of two experts' minimum information distributions, in which Expert 1 has weight 0.3 while Expert 2 has weight 0.7. This illustrates the process by which the Decision Maker's interpolated quantile distribution is derived from the weights ascribed to the experts in the Classical Model.





EXCALIBUR Procedure

The main steps in the EXCALIBUR approach:

- **A group of experts is selected.**
- **Expressing views as elemental uncertainty distributions, experts assess a set of variables ('seed items'), true values of which are known or become known post hoc.**
- **Experts' responses are scored with regard to statistical likelihood that distributions over the set of seed items correspond to the observed or measured results - and also scored by a measure of informativeness compared to uniform background distribution.**
- **The two scores are combined to form a weight for each expert.**
- **Experts are elicited individually regarding their uncertainty judgments in relation to questions of interest (the 'target items').**
- **Performance-based or equal weights scores are applied to individual responses to obtain weighted pooling of uncertainty distribution for each target items.**

EXCALIBUR applications

Early applications:

Space
(propulsion system reliability)
Space
(space debris impact)
Space
(strength of composites)
Industrial
(flange connection failures)
Industrial
(fuelling crane failure)
Hydrology
(groundwater contamination; reservoir erosion modelling)

More recent applications:

Volcanology
(eruption risks…..)

Seismology
(earthquake hazards)

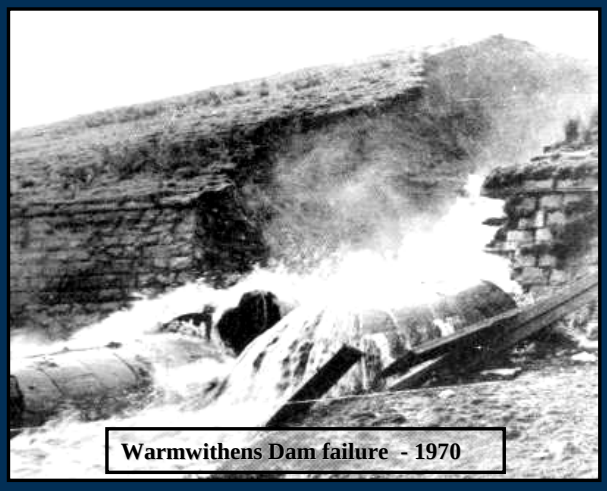
Climate change
(radwaste storage)

Bioterror
(malicious biological agents……)

Medical

- (risk models for SARS; vCJD in blood products; XMRV; chronic wasting disease; urinary fertility products, etc)

From air to water…… ..risk assessment and reservoir safety in the UK.



Warmwithens Dam failure - 1970

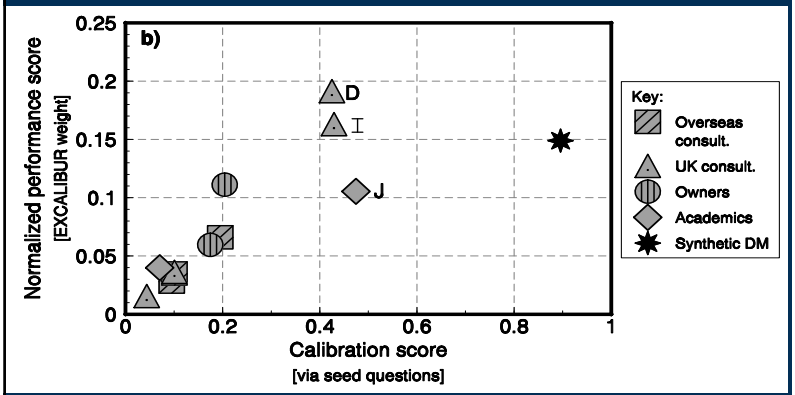
Expert group – circa 1917

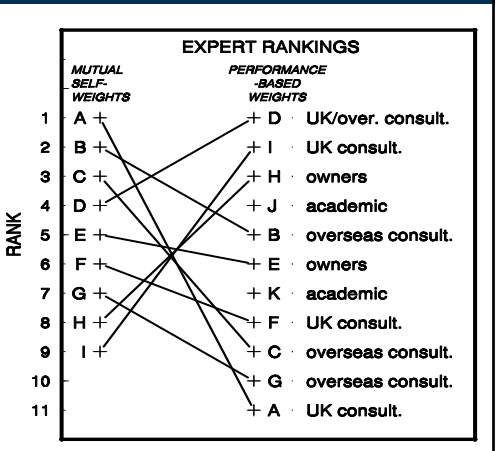
Cowlyd Reservoir inspection party - 1917

Objective: to develop a generic quantitative model for accelerated internal erosion in Britain's population of 2,500 ageing dams, using elicited quantities for key variables

Reservoir engineers: performance-based scores, and mutual weightings

Note big discrepancies between performance-based ranking and *a priori* ranking from mutual weighting exercise (RH panel)





See: **Burgman, M.A., McBride, M., Ashton, R., Speirs-Bridge, A., Flander, L., Wintle, B., Fidler, F., Rumpff, L. and Twardy, C. (2011). Expert status and performance. PloS One 6, e22998**
doi:10.1371/journal.pone.0022998

Expert judgements are essential when time and resources are stretched or we face novel dilemmas requiring fast solutions.

Typically, experts are defined by their qualifications, track record and experience.

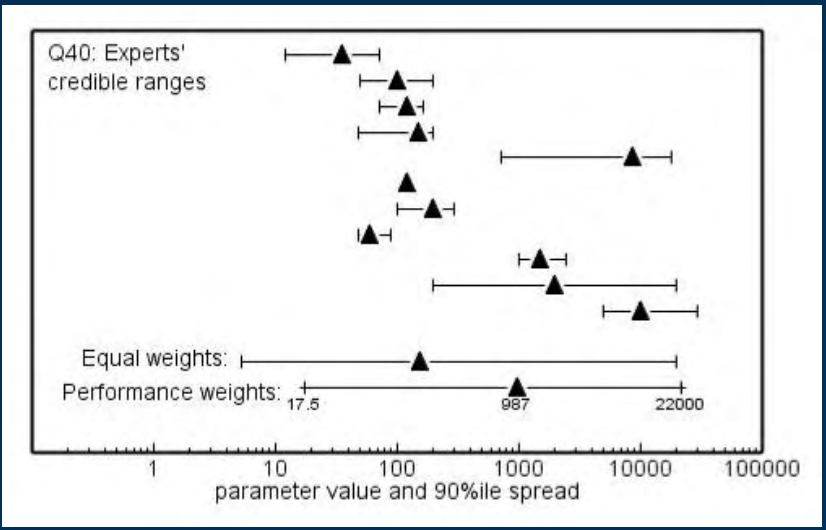
The social expectation hypothesis argues that more highly regarded and more experienced experts will give better advice.

Results indicate that the way experts regard each other is consistent, but unfortunately, ranks are a poor guide to actual performance.

Expert advice will be more accurate if technical decisions routinely use broadly-defined expert groups, structured question protocols and feedback.

Experts' spreads for one parameter

Experts' opinions on the time-to-failure (in days from first detection) of the 10%ile slowest cases, and two alternative ways of pooling weighted opinions – Equal weights and Performance-based DMs



Note the "two schools of thought" effect…

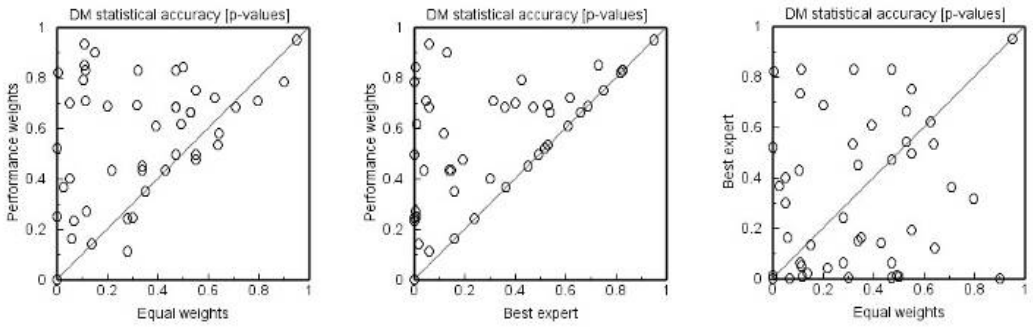
and the strong 'opinionation' of many experts

Experts CAN quantify uncertainty as subjective probability (case histories to 2005)

TU DELFT Expert Judgment database 45 applications (anno 2005):	# experts	# variables	# elicitations
Nuclear applications	98	2,203	20,461
Chemical & gas industry	56	403	4,491
Groundwater / water pollution / dike ring / barriers	49	212	3,714
Aerospace sector / space debris /aviation	51	161	1,149
Occupational sector: ladders / buildings (thermal physics)	13	70	800
Health: bovine / chicken (*Campylobacter*) / SARS	46	240	2,979
Banking: options / rent / operational risk	24	119	4,328
Volcanoes / dams	231	673	29079
Others	19	56	762
TOTALS	***521***	***3688***	***67001***

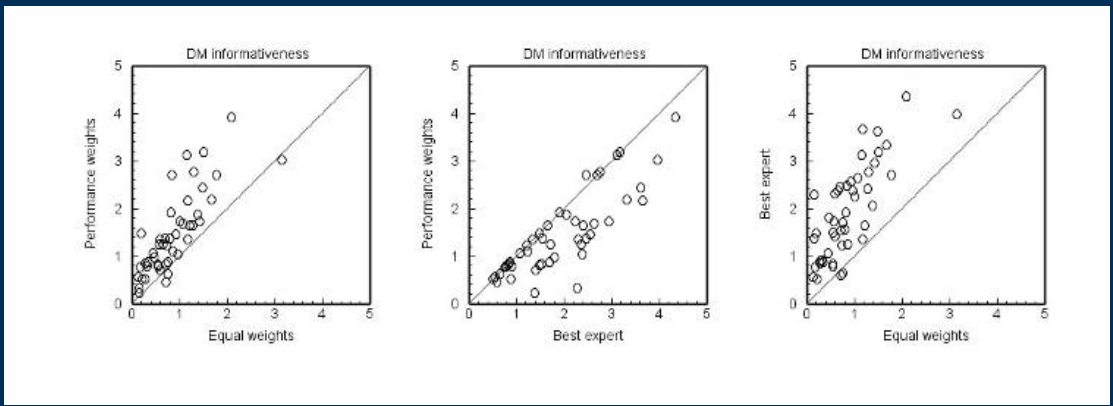
Summarising the TU Delft database of applications – Decision Maker DM statistical accuracy

Cooke, R. M. & Goossens, L. L. H. J. Reliability Engineering & System Safety 93, 657-674 doi:10.1016/j.ress.2007.03.005 (2008).



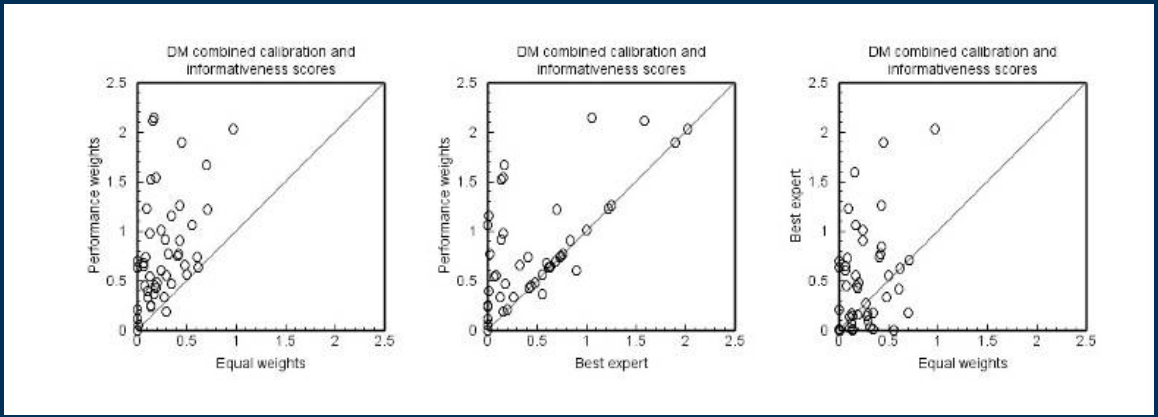
Performance weighted DM usually does better than Equal Weights for statistical accuracy, but not invariably, and as good as or better than individual "Best expert"; Best expert vs Equal weights is a toss-up …

The TU Delft database of applications – DM informativeness



Performance weighted DM almost always does better than Equal Weights for informativeness, unsurprisingly, but rarely better than the individual "Best expert", again unsurprisingly; similarly Best expert is more informative than the Equal weights DM …

The TU Delft database of applications – DM combined score



For overall Classical Model score (i.e. statistical accuracy with informativeness), Performance weighted DM is almost always better than Equal Weights, and usually as good as or better than the "Best expert" DM; which of Best expert and Equal weights DM is better is a lottery …

Analysing expert elicitations with Cooke's "Classical Model"

The procedure relies on cornerstones of the scientific method:

Empirical control - evaluates weights for experts on basis of measures of performance

Accountability - inputs are traceable in terms of scientific inputs of individuals

Reproducibility - can replicate and review all calculations used

Advantages:

Impartiality - experts are treated equally prior to calibration

Equity – individual experts' scores are maximised by stating true scientific views

Diagnostic - procedure can highlight discrepancies in reasoning or inconsistencies in interpretation

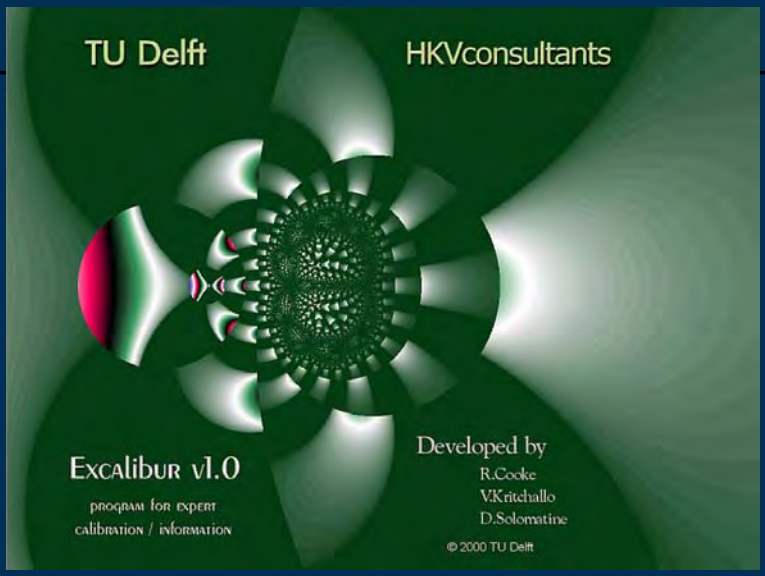
……this approach produces a "rational consensus", and sits squarely within the Bayesian paradigm for decision-support

In almost all circumstances, and at all times, we find ourselves in a state of uncertainty

- Bruno de Finetti

….and scientists will continue to be perplexed, bemused and uncertain!

Conclusion: structured expert elicitation with formalized differential performance-based opinion weighting offers a rational way to deal with most forms of scientific uncertainty, when other solutions are not available.



In the practical, we will use the Classical Model procedure and EXCALIBUR package to calibrate you as "experts", and elicit some important expert judgments from you …… ☺