TSE WP working paper series

12-293

Research Group: Environmental Economics and Natural Resources

March, 2012

“Combining Experts’ Judgments: Comparison of Algorithmic Methods using Synthetic Data”

James K. Hammitt and Yifan Zhang

Ecole d’économie de Toulouse (TSE) - Manufacture des Tabacs
Aile Jean-Jacques Laffont - 21, allée de Brienne - 31000 TOULOUSE
Tél : +33 (0) 5 61 12 85 89 - Fax : +33 (0) 5 61 12 86 37 - **www.TSE-fr.eu** - contact@TSE-fr.eu

Toulouse School of Economics

Combining Experts' Judgments: Comparison of Algorithmic Methods using Synthetic Data

James K. Hammitt

Harvard University (Center for Risk Analysis)
718 Huntington Ave., Boston, MA 02115, USA

Toulouse School of Economics (LERNA-INRA)
21 allée de Brienne, 31000 Toulouse, France

jkh@harvard.edu

Yifan Zhang
Harvard University (Department of Biostatistics)

March 2012

Abstract

Expert judgment (or expert elicitation) is a formal process for eliciting judgments from subject-matter experts about the value of a decision-relevant quantity. Judgments in the form of subjective probability distributions are obtained from several experts, raising the question how best to combine information from multiple experts. A number of algorithmic approaches have been proposed, of which the most commonly employed is the equal-weight combination (the average of the experts' distributions). We evaluate the properties of five combination methods (equal-weight, best-expert, performance, frequentist, and copula) using simulated expert-judgment data for which we know the process generating the experts' distributions. We examine cases in which two well-calibrated experts are of equal or unequal quality and their judgments are independent, positively or negatively dependent. In this setting, the copula, frequentist, and best-expert approaches perform better and the equal-weight combination method performs worse than the alternative approaches.

1. Introduction

Expert judgment (or expert elicitation) is a formal process for eliciting experts' beliefs or opinions about the value of a quantity that may be used as input to a model to inform policy decisions or for other purposes. A defining feature is that each expert provides a subjective probability distribution that summarizes his beliefs about the value of the quantity. While the method has been used for several decades, it is receiving increasing attention. Major expert-judgment studies have been conducted in recent years to quantify mortality effects of airborne particulate matter (Cooke et al. 2007, Roman et al. 2008), key parameters associated with climate change and its impacts (Morgan and Keith 1995, Morgan et al. 2001, Morgan et al. 2006, Zickfield et al. 2007), and future development of climate-friendly energy technologies (Curtwright et al. 2008), among others. The US Environmental Protection Agency (EPA) recently issued a draft white paper discussing the development and appropriate agency application of expert judgment (EPA 2009).

Expert-judgment studies elicit a probability distribution for a quantity of interest (a 'target' quantity or variable) from each of several experts, raising the question of how best to use the multiple distributions. Assume without loss of generality that the target variable is an input to a policy model (e.g., a risk-analytic, benefit-cost, integrated-assessment, or other model intended to inform a decision). The policy model propagates uncertainty about input variables (represented by probability distributions) and produces a probability distribution for the output variable (or variables).

One approach is to run the policy model multiple times, using each expert's distribution as input, then report the resulting set of output distributions to the decision maker. A second approach is to combine these multiple output distributions into a single distribution using some algorithmic method. A third approach is to combine the experts' distributions for the target variable and to use that single distribution as input to the policy model, producing a single output distribution. In general, the second and third approaches will yield different output distributions unless the policy model is linear in the target quantity.

Some authors argue strongly for the first approach, noting that the degree of similarity among experts' distributions may convey important information to a decision maker (e.g., Keith 1996). This approach has limitations, however. If several target quantities are elicited from experts, reporting output distributions corresponding to each combination of input distributions could yield a very large number of output distributions for the decision maker to somehow evaluate. More generally, there is always a question of the most useful balance between aggregating information through models and providing more disaggregate information to decision makers. Taken to an extreme, opposition to aggregation would suggest discarding the policy model and simply presenting decision makers with all available information about each input variable. To the extent that it is useful to aggregate information, it is not clear why one should aggregate information about quantities from some sources (e.g., direct measurements, model estimates reported in literature) but not from expert judgments.

Many approaches to aggregating probability distributions have been developed (Cooke 1991, Clemen and Winkler 1999, 2007). Perhaps most often, aggregation is done informally by the analyst or consumers of the analysis. Alternatively, judgments can be combined using an algorithmic approach or some type of consensus-seeking process. The most common algorithmic approach is to calculate the simple average of the experts' distributions.

This paper is directed toward evaluating the properties of algorithmic combination methods. We compare two state-of-the-art combination methods, the performance (classical) method (Cooke 1991) and the copula method (Jouini and Clemen 1996) with each other and with several alternatives (equal-weight, best-expert, frequentist combination). We employ simulated expert-judgment data, for which we know the properties of the process by which the data are generated; hence we can vary these properties to determine how each of them affects the performance of the combination rules. The use of synthetic data complements studies that use real expert-judgment data (Kallen and Cooke 2002, Cooke 2008, Cooke and Goossens 2008, Lin and Cheng 2009, Flandoli et al. 2011). Another approach to evaluating the performance of alternative combination rules is the analytic method developed by Hora (2010).

The following section presents the methods used, including the model to simulate expert-judgment data, the combination rules that are assessed, and the evaluation criteria. Results are presented in Section 3 and compared with results from studies using field data and with analytic results. Conclusions follow in Section 4.

2. Summary of Methods

We consider a decision maker who elicits subjective probability distributions from each of several experts. Simulated experts provide distributions for one 'target' variable and ten 'seed' variables. The target variable is the quantity of interest, for which the expert-judgment study is conducted. Seed variables are quantities for which the values are not known to the expert but are or will become known to the decision maker. Seed variables allow the decision maker to learn about the quality of the experts' distributions and to use this information in combining them.

We evaluate several combination rules (also described as 'decision makers'). Three are linear combinations of the experts' probability distributions: Cooke's (1991) performance (classical) decision maker (in which the weights depend on the quality of the experts' judgments on seed variables), a simple (equally weighted) average, and a best-expert method (in which the expert with the largest weight in the performance method receives weight one and the other experts receive weight zero). The other combination rules are a copula method (Jouini and Clemen 1996) in which the decision maker updates his (non-informative) prior distribution using the experts' distributions (the name derives from the use of a copula to characterize the decision maker's beliefs about the mutual dependence of the experts' judgments) and a frequentist approach that interprets the experts' distributions as imperfect observations of the variable. In contrast to the linear combinations, the copula and frequentist approaches require information about dependence among the experts' distributions. Because this information would generally not be available to a decision maker, it is estimated using the experts' distributions for the seed variables. (An alternative would be to elicit judgments about dependence from the experts, Clemen et al. 2000.)

The combined distributions are evaluated in terms of calibration (consistency of distributions with realizations), informativeness (tightness of distributions), and the product of these terms. These concepts are defined formally below.

We model experts' distributions as if the expert observes the sum of the variable plus an expert-specific random error (i.e., the expert observes the variable with noise). Each expert's error terms are normally distributed with mean zero and each expert knows his observation-error distribution. An expert's observation errors are independent across variables. Hence each expert's distribution for a variable is normal with mean equal to the value he observes and variance equal to the variance of his individual-specific error.

In general, experts may differ in quality and in the similarity of their knowledge. Quality of experts' distributions is characterized here by two properties, calibration and informativeness. We model experts as perfectly calibrated: their error terms are unbiased (mean zero) and each expert knows his own error variance. We consider cases in which experts have equal and unequal error variances. In cases where experts have unequal error variances, the expert with smaller variance has higher quality (his distribution is more informative).

The assumption of perfect calibration is strong and in conflict with much empirical evidence suggesting that individuals providing probability distributions either in their domains of expertise or in areas where they have no special competence (e.g., ‘almanac questions’) are often overconfident, providing credible intervals that include the realized value much less often than the stated probability (e.g., Morgan and Henrion 1990 report that 20 to 45 percent of realizations fall outside the stated 98 percent credible intervals). Several alternative justifications for the assumption can be offered. First, it provides an idealized benchmark case in which to evaluate the performance of different combination rules. In contrast, there is an infinity of ways in which experts can be imperfectly calibrated, comprising alternative degrees of miscalibration on each of several dimensions (e.g., overconfidence, bias, skewness, peakedness, excessively light or heavy tails). In future research, it would be useful to explore how results are affected by the extent and characteristics of imperfect calibration.

Second, in some cases experts do appear to be reasonably well calibrated. Winkler and Poses (1993) found that physicians’ estimates of the probability that patients admitted to an intensive care unit would be discharged alive were reasonably well calibrated, especially among primary attending physicians and those with regular intensive care unit experience. Despite their lack of prior experience in reporting judgmental survival probabilities, the physicians’ performance compares favorably with meteorologists’ assessed probabilities of precipitation. In a small study, Walker et al. (2003) found that exposure experts asked to predict measured benzene concentrations in EPA Region V were reasonably well-calibrated. Lin and Bier (2008) analyzed calibration in 27 expert-judgment studies including approximately 200 experts in total, confirming the existence of systematic differences in calibration among studies and experts, with some cases suggesting reasonable calibration.

Third, experts are often provided training in advance of the elicitation to alert them to the dangers of reliance on heuristics such as anchoring and adjustment that can lead to overconfidence or other forms of imperfect calibration (Tversky and Kahneman 1974, Morgan and Henrion 1990). Experts receiving this training may be better calibrated than subjects in many of the studies that demonstrate overconfidence.

Fourth and finally, a decision maker could adjust experts’ distributions for the target variable, potentially improving calibration by using information about their performance on seed variables or other factors (Hammitt and Shlyakhter 1999). Our analysis can be interpreted as applying to experts' distributions after these adjustments have been made.

A critical consideration in aggregating information from multiple experts is the extent of dependence among their judgments. If multiple experts provide independent information, then an appropriate aggregate can be highly informative (e.g., the average of N independently and identically distributed estimates of a quantity has variance N-fold smaller than the variance of each estimate). Alternatively, if experts share much of the knowledge relevant to estimating a parameter value (e.g., a common scientific literature and disciplinary perspective), the information contained in the union of their judgments may be little more than that contained in a single expert's judgment (in effect, each expert may report his idiosyncratic perception of a consensus). Clemen and Winkler (1985) provide bounds on the number of independent experts whose combined information is equivalent to that of a larger number of dependent experts.

We simulate expert distributions and combinations for a variety of cases defined by the covariance matrix of observation errors. We vary the relative quality of the experts and the dependence of their information. We report results for two experts and include cases in which the experts have equal or unequal error variances and zero, positive, or negative dependence among their observation errors for a common variable. The case of positive dependence between experts seems most realistic, given that subject-matter experts share knowledge, theoretical frameworks, and assumptions about their subject. Kallen and Cooke (2002) analyzed dependence among experts' judgments (by comparing experts' medians with realized values) for 28 expert-judgment studies and found they could reject the hypothesis of independence at the 5 percent significance level for about half the studies. They report the average correlation between experts in two studies as 0.55 and 0.32. Their analysis also suggests heterogeneity of inter-expert correlations. They report the estimated correlation matrix for one study with eight experts. The 28 estimated correlation coefficients range from 0.23 to 0.76; four exceed 0.7 and two are less than 0.3.

The case of two experts is unrealistic as most studies include about five or ten.[1] This restriction simplifies the study design, analysis, and reporting, as it restricts the degrees of

[1] Cooke and Goosens (2008) describe 45 expert-judgment studies, in which the median number of experts is eight and the upper quartile is 11. Hora (2004) provides evidence that calibration of distributions obtained by pooling experts judgments increases substantially as the number of experts increases to about five, with modest additional improvement as the number increases to 10.

freedom in the experimental design. With more than two experts the set of possible cases that can be studied increases rapidly unless one imposes restrictions on how error variances and correlations may differ among experts. In subsequent work, it would be valuable to increase the number of experts.

Combined distributions are evaluated in terms of calibration, informativeness, their product (called the ‘combined score’), and by the probabilities that the distribution is superior to the distribution produced by each alternative combination rule (i.e., has the larger combined score). Simulated experts report distributions for ten seed variables (a quantity frequently used in practice; Cooke and Goossens 2008) and one target variable. We replicate each expert’s and each decision maker’s distribution for the target variable 30 times in order to assess calibration. Without loss of generality, the value of each variable is set to zero (i.e., our analysis can be interpreted as studying the difference between the mean of a distribution and the value of the variable).

The simulation proceeds as follows:

1. Draw random observation errors for each expert for ten seed variables and 30 replicates of one target variable using the specified covariance matrix. Generate each expert’s distributions for seed and target variables as normal with mean equal to the realized observation error and variance equal to the expert’s error variance. Using the seed-variable distributions, estimate the covariance matrix of experts’ errors (assuming each expert’s error has mean zero) and calculate the calibration, informativeness, and combined score of each expert. Combine experts’ distributions using performance, copula, frequentist, equal-weight, and best-expert combination rules. Evaluate calibration, informativeness, and combined score of each decision maker and compare combined scores among all pairs of decision makers.

2. Replicate step 1 100 times. Calculate the frequency with which each decision maker has the largest combined score in each pair-wise comparison.

3. Replicate step 2 100 times. Calculate deciles of the distribution of the frequency with which each decision maker has the larger combined score for each pair-wise comparison. Pooling all 10,000 replicates, calculate deciles of calibration, informativeness, and combined score for each expert and each decision maker.

Evaluation Criteria

Consider the case of N experts indexed by $i = 1, 2, \ldots,$ N. Each expert provides a probability distribution for $S + T$ variables, indexed by t. Variables $1, 2, \ldots, S$ are seed variables and variables $S + 1, S + 2, \ldots, S + T$ are replicates of the target. For a single realization of step 1 of the simulation, let ε_{it} denote the realization of the observation error for expert *i* and variable *t*. Knowing that his observation error is normal with mean zero and variance σ_i^2, expert *i* reports that his distribution for variable *t* is $N(\varepsilon_{it}, \sigma_i^2)$.

Calibration, informativeness, and combined score are defined by Cooke (1991). They are properties of a set of one or more distributions, either the distributions provided by an expert or a combined distribution.

Calibration is a measure of the extent to which a distribution accurately portrays the frequency distribution of realizations. Divide each distribution into *b* bins and let p_k be the probability content of bin *k* (in our analysis, $b = 10$ and the bins are divided by the deciles of the distribution so $p_k = 0.1$ for $k = 1, 2, \ldots, 10$). For the target variable, aggregate over the *T* realizations and let s_k be the frequency with which bin *k* includes the value of the variable.[2] The calibration score of expert *i* is defined as the right-tail probability of a χ^2 variable with $b - 1$ degrees of freedom,[3]

$$C = 1 - \chi^2_{b-1}\left[2T \cdot \sum_{k=1}^{b} s_k log\left(\frac{s_k}{p_k}\right)\right]. \qquad (1)$$

C is bounded by zero and 1, achieving its maximum in the case of perfect calibration (when $s_k = p_k$ for all *k*). An expert's calibration score for the seed variables is calculated similarly, letting s_k be the frequency with which the values of the seed variables fall in bin *k* and replacing *T* with *S* (the number of seeds).

Informativeness is a measure of the concentration or spread of a distribution. It is defined as the relative information of the expert's distribution compared with a minimal information density function. The minimal information density for a variable is taken to be uniform on the 'intrinsic range' for that variable (defined below).

Consider an expert's distribution for variable *t* with intrinsic range [*l*, *u*]. Let f^k be the kth fractile[4] of the distribution. Then

$$I = \left[\sum_{k=1}^{b} p_k log\left(\frac{p_k}{\frac{f^k - f^{k-1}}{u - l}}\right)\right]. \qquad (2)$$

[2] Note that if $s_k = 0$, the value of the corresponding term in the summation of equation (1) is zero.

[3] The term in brackets is asymptotically distributed χ^2 in the number of variables T.

[4] Let $f^0 = l$ and $f^b = u$.

I takes its minimum value of 0 if the distribution is uniform on the intrinsic range. Its value increases as the difference between some pairs of adjacent fractiles becomes smaller, yielding a more spiked distribution.

The intrinsic range [*l*, *u*] for each variable is defined by applying the ‘10 percent overshoot’ rule to the data (Cooke and Goossens 2008). It is constructed as follows for variable *t*: (1) Let l_i and u_i be expert *i*’s 0.05 and 0.95 fractiles. (2) Let $l_0 = \min_i \{l_i, \theta\}$ and $u_0 = \max_i \{u_i\ \theta\}$ where θ is the value of variable *t*. (3) Define $l = l_0 - 0.1\ (u_0 - l_0)$ and $u = u_0 + 0.1\ (u_0 - l_0)$. In words, the intrinsic range is defined by the smallest 0.05 fractile and largest 0.95 fractile (or the value of the variable when it is not contained by these extreme fractiles), extended in each direction by 10 percent of the difference between them.

The combined score is a summary measure of the quality of a set of one or more distributions. It is defined as the product of the calibration and mean (over multiple variables) information score.

Combination rules

The combined distributions for the target variable (i.e., ‘decision makers’) are defined as follows. The equal-weight decision maker is the simple average of the experts’ distributions for the target variable. The best-expert decision maker is the target-variable distribution provided by the expert having the largest combined score (calculated over the seed variables) among all experts.

The performance decision maker (Cooke 1991) is a weighted average of the distributions of the experts. The weights are proportional to the experts’ combined scores (calculated over the seed variables) except that experts’ whose calibration score falls below a cut-off α receive weight zero. Positive weights are normalized to sum to unity. The value of α is determined by maximizing the average combined score of the combined distributions for the seed variables.

The frequentist decision maker combines distributions by treating each expert as providing an observation of the target variable plus random error. The frequentist decision maker is normal with mean equal to the inverse-variance weighted average of the experts’ means

$$m = \frac{s'Ms}{s's} \tag{3}$$

and variance

$$v = \frac{s'Rs}{s'Vs} \tag{4}$$

where *s* is a column vector with $s_i = 1/\sigma_i$, *M* is a diagonal matrix with $M_{ii} = \varepsilon_i$, *R* is a matrix with $R_{ij} = corr(\varepsilon_i, \varepsilon_j)$, *V* is a matrix with $V_{ij} = 1/(\sigma_i\ \sigma_j)$, and prime denotes transpose. The elements of *s*, *R* and *V* are estimated from the experts' seed-variable distributions under the assumptions that each expert's error terms are normally distributed, independent across variables, have mean zero, constant variance, and that the correlation coefficient between experts' error terms for a variable is common for all variables and all pairs of experts. For the simulated expert-judgment data, all these assumptions are correct.

The copula decision maker (Jouini and Clemen 1996) is motivated by Bayes' rule. Under this approach, the decision maker is taken to have a non-informative prior distribution for the target variable that he updates taking the experts' distributions as data. Hence the copula decision maker is proportional to the likelihood of the experts' judgments,

$$P(\theta|h_1, \ldots, h_n) \propto c[1 - H_1(\theta), \ldots, 1 - H_n(\theta)] \prod_{i=1}^{n} h_i(\theta) \tag{5}$$

where *P* is the copula decision makers' density function for the target variable θ, $h_i(\theta)$ and $H_i(\theta)$ are expert i's density and cumulative distributions for θ, and $c[\cdot]$ is the copula function that encapsulates the decision maker's information about the dependence among experts' observation errors. We adopt the multivariate normal copula

$$c[y] = \frac{exp[-y'(R^{-1} - I)y/2]}{\sqrt{|R|}} \tag{6}$$

where *y* is a vector with elements $y_i = 1 - H_i(\theta)$, *R* is the correlation matrix estimated from the experts' judgments of the seed variables (identical to the matrix used by the frequentist decision maker), and *I* is the identity matrix. It can be shown that the frequentist and copula decision makers are identical when the experts are independent or have equal variances (σ_i all equal).[5]

Because the copula decision maker's density function is proportional to the joint likelihood of the experts' densities, it exhibits a strong form of the 'zero-preservation property': if any expert assigns probability zero to some range of values of θ, then the copula decision maker also assigns probability zero to those values (Cooke 1991). Conversely, the

[5] Note that when both conditions are assumed to hold, $R = I$, and so $c[y] = 1$ (equation (6)) and the copula (and frequentist) decision makers' density functions are equal to the product of the experts' densities.

copula decision maker tends to concentrate probability on values to which all experts assign significant probability (see Hammitt and Shlyakhter 1999 and Kallen and Cooke 2002 for examples). In contrast, linear-opinion pools like the equal-weight and performance decision makers tend to spread probability over all values to which at least some experts assign significant probability.

3. Results

Simulations are reported for a panel of two experts and for six covariance matrices. The covariance matrices provide cases in which the experts have equal or unequal error variances and error terms that are independent, positively, or negatively dependent (with correlation coefficient of 0, 1/2, and -1/2, respectively).[6]

Simple descriptive statistics (the mean and variance) of the combined distributions are reported in the two panels of Table 1. Columns correspond to the six covariance matrices (reported at the head of each column). Within each cell, the first number is the sample mean and the numbers in parentheses are the interquartile range from the simulation. For all decision makers and across all covariance matrices, the mean averages very nearly zero (its true value); the largest absolute deviation is 0.004. The variability of the mean (as characterized by its interquartile range) varies across covariance matrices. The means for the copula and frequentist decision makers tend to be among the least variable and those for the performance and best-expert decision makers among the most variable. This pattern is stronger when the experts' errors are negatively correlated (columns C and F); when errors are positively correlated (columns B and E), there is little difference in variability of the mean across decision makers. Finally, the equal-weight decision maker has the least-variable mean when the experts have equal error variances but among the most-variable means when they have unequal variances.

Across covariance matrices, the variance of the combined distribution is smallest (on average) for the copula and frequentist decision makers and largest for the equal-weight decision maker. The variance of the combined distribution is larger (smaller) when the expert errors are positively (negatively) correlated than when they are uncorrelated, except for the best-expert decision maker when the experts have equal error variances (for which the variance of the combined distribution is necessarily one). The variance is of course larger

[6] The value of 1/2 is compatible with the values (0.55 and 0.32) that Kallen and Cooke (2002) estimate for the two expert-judgment studies they report.

when the experts have unequal variances (1 and 4) than when they both have variance one. The distribution of the variance is right-skewed, with mean values greater than or equal to the third-quartile in 8 of the 20 cells shown in Table 1.

Performance characteristics of the combined distributions are summarized in Table 2. The three panels report the calibration, informativeness, and combined score for each decision maker as a function of the covariance matrix of the experts' errors.

With regard to calibration, the copula and frequentist decision makers are very similar and the performance decision maker is somewhat better. The best-expert decision maker is always better calibrated than the others and the equal-weight decision maker is substantially more poorly calibrated than any of the others. The superiority of the best-expert to the performance and equal-weight decision makers is anticipated from Hora's (2004) proof that a linear combination of well-calibrated experts must be less well calibrated than the experts themselves (unless their judgments are identical). For the performance and equal-weight decision makers, calibration is substantially better when the experts' errors are positively correlated (columns B and E) than when they are independent or negatively correlated. The equal-weight decision maker is very poorly calibrated when the experts' errors are negatively correlated. Calibration of the other decision makers does not vary much across error distributions.

In terms of informativeness, the copula and frequentist decision makers are very similar and are more informative than any of the others. The best-expert tends to be the next most informative followed by the performance decision maker. The equal-weight decision maker is uniformly the least informative. This pattern is expected because the copula and frequentist approaches tend to concentrate probability while the linear combinations spread it. Informativeness of the copula and frequentist decision makers is highly sensitive to the error distribution. These decision makers are more informative when the experts' errors are negatively correlated and less informative when the experts' errors are positively correlated. When the experts have equal error variances, the performance and best-expert decision makers are more informative when the experts are negatively correlated and less informative when they are positively correlated; the dependence on correlation is less evident when the experts have unequal error variances. For all decision makers, information is larger when the experts have unequal rather than equal error variances, which reflects the fact that the intrinsic range (against which information is calculated) is larger in this case (holding dependence constant, variance of the combined distribution is always larger when the experts have unequal rather than equal variances, see Table 1).

The copula and frequentist decision makers have similar combined scores that are larger than those of the other decision makers. These are followed by the best-expert then the performance decision maker. The equal-weight decision maker has uniformly the smallest combined score. With the exception of the equal-weight decision maker, the other decision makers tend to have larger combined scores when the experts are negatively rather than positively correlated and when the experts have unequal rather than equal variances.

Results of the pair-wise comparisons of decision makers are reported in Table 3. The six panels correspond to the covariance matrices in the corresponding columns of Table 2. The number in each cell is the estimated probability that the row decision maker is superior to the column decision maker (i.e., has the larger combined score). Asterisks indicate that the estimated probability is significantly different from 50 percent at the 5 percent level.

Over all covariance patterns, the copula and frequentist decision makers perform similarly. Their results are generally equal or superior to those of the performance, equal-weight, and best-expert decision makers. In cases where their performance differs significantly, the copula decision maker outperforms the frequentist decision maker (panels A and C, with equal variance and independent or negatively dependent expert errors). Although the copula and frequentist decision makers are identical when the experts' errors are assumed to be independent or to have equal variance (as noted above), the simulated decision makers use estimates of the covariance matrix (calculated from the seed variables) and hence are not identical in panels A, B, C, and D.

The equal-weight decision maker is substantially inferior to the alternative combination rules. Its probability of being superior to an alternative decision maker is about 30 percent in the cases with positive dependence among the experts' errors (panels B and E) and less than about 10 percent in all other cases.

The best-expert decision maker performs well against the performance and equal-weight decision makers. Its probability of providing the better distribution is always about 80 percent or larger. Its advantage is smallest in cases with positive dependence among the experts' errors (panels B and E). The best-expert decision maker is comparable to the copula and frequentist decision makers. It performs significantly worse than these alternatives in panel C (equal variance, negative dependence of experts' errors) and somewhat better in panel E (unequal variance, positive dependence).

The performance decision maker is superior to the equal-weight, and inferior to the best-expert, decision maker for all six covariance matrices. It is comparable to the copula and frequentist decision makers when the experts' errors are positively dependent (panels B and E) but inferior for other cases. As noted above, positive dependence seems likely to be the most realistic case.

Comparison with empirical and other studies

Our analysis of the comparative performance of five combination methods using synthetic expert-judgment data complements prior studies that have compared a smaller number of combination methods using data collected in real expert-judgment studies. Cooke and Goossens (2008) compare the performance, equal-weight, and best-expert combination rules using 45 expert-judgment studies that include multiple seed variables and encompass a wide variety of topics. Evaluating the performance of the three decision makers using the combined score calculated on the full set of seed variables for each study, they find that the performance decision maker is better than both equal-weight and best-expert decision makers in 27 of 45 cases. In an additional 15 cases, the performance decision maker and best expert coincide (i.e., the performance decision maker puts unit weight on a single expert) and have a higher combined score than the equal-weight decision maker. The equal-weight decision maker is best in only one case and better than the best-expert in 18 cases. The best-expert decision maker is best in two cases. Overall, they find the equal-weight decision maker is slightly less well calibrated and significantly less informative than the performance decision maker.

In response to Clemen's (2008) observation that performance should be tested out of sample (i.e., on variables other than those used to estimate the weights for the performance decision maker), Cooke (2008) reports a follow-on comparison of performance and equal-weight decision makers in which weights for the performance decision maker are calculated using half the seed variables in a study and performance is evaluated using the other half of the seed variables for each study. This analysis is restricted to 13 studies having at least 16 seed variables and yields 26 comparisons (using each half of the seed variables alternatively to estimate weights and to evaluate combination rules). In this out-of-sample comparison, the performance decision maker has a larger combined score than the equal-weight decision maker in 20 of 26 cases (results for the best-expert are not reported).

Using a subset of the Cooke and Goosens (2008) data, Lin and Cheng (2009) conduct a cross-validation exercise in which each seed variable is used in sequence as the target and the performance and best-expert decision makers are identified using the remaining seeds. They find that performance and equal-weight methods perform similarly, with the performance decision maker and equal-weight decision maker having the largest average combined scores for 16 and 17 cases out of 40, respectively. The best expert has the largest average combined score for 8 cases (for one cases the performance decision maker and best expert coincide).

Flandoli et al. (2011) use data from five expert-judgment studies. For each study, they treat 70 percent of the seed variables as seeds and the remaining 30 percent as targets. Averaging over all possible splits between seeds and targets, they find the performance decision maker has a higher combined score for three datasets and the equal-weight decision maker has a higher combined score for two.

Kallen and Cooke (2002) provide the only study of which we are aware that compares copula and performance decision makers using field data. In contrast to the present study, they use Frank's copula (as originally suggested by Jouini and Clemen 1996). They use data from two expert-judgment studies to compare the copula, performance, equal-weight, and best-expert decision makers. In both cases, the copula decision maker has the smallest calibration and highest information. Its combined score is the smallest in one case and second smallest in the other (slightly exceeding the combined score of the equal-weight decision maker). They observe that Frank's copula can lead to numerical instability and suggest that use of a multivariate normal copula (as suggested by Clemen and Reilly 1999 and used here) may improve performance.

Hora (2010) provides an analytic approach to evaluate performance of combination rules and illustrates its application to equal-weight and geometric combinations of experts' distributions that are normal, independent or positively dependent, and perfectly or imperfectly calibrated. He finds that geometric combination (like the copula and frequentist measures) performs best when experts are independent and well calibrated but its performance deteriorates compared with the equal-weight combination as (positive) dependence increases or expert calibration decreases. Consistent with the present study, he also finds that geometric combinations tend to have lower variance (higher informativeness) than the equal-weight combination.

Our results using synthetic data are consistent with most of the studies using field data in finding that the equal-weight decision maker performs worse than alternatives. In contrast, in our analysis the best-expert decision maker outperforms the performance decision maker, with probabilities of having a higher combined score exceeding 80 percent for all six covariance matrices. In the analyses using field data, the best-expert is usually inferior to performance and equal-weight decision makers. The difference may be explained by the use of perfectly calibrated experts in the synthetic data. As shown by Hora (2004), a linear combination of well-calibrated experts is necessarily less well calibrated than the individual experts; hence the best synthetic expert compares well with both performance- and equal-weight decision makers. Human experts are often not well calibrated and so a linear combination can improve calibration and potentially overall performance.

4. Conclusion

Evaluation of alternative combination rules for experts' judgments, using simulation methods so that the process generating the experts' distributions is known and its properties can be experimentally varied, can contribute to understanding the properties of methods for combining judgments and have implications for the choice of method for application. We simulate what may be considered a best case for expert judgment in which the experts are perfectly calibrated. In subsequent work, it would be useful to compare combination rules using more realistic synthetic data where the experts are not perfectly calibrated and to determine how the type of miscalibration (e.g., overconfidence, bias, excessively light tails) influences the performance of alternative combination rules. In addition, future studies should explore how relative performance of the combination rules depends on the number of experts.

We evaluate five combination rules in six contexts characterized by equal or unequal quality of the experts (as represented by the variance of their error terms) and by positive, negative, or zero dependence among their judgments. With the exception of the equal-weight combination rule, all of the other rules require information on experts' quality that can be obtained by evaluating their judgments on seed variables (under the maintained assumption that expert performance on the seed variables is predictive of their performance on the target). Among these combination rules, we find that the copula, frequentist, and best-expert approaches generally perform better than the performance combination method. Across all cases considered, the equal-weight combination rule, which is the approach most often applied in practice, is clearly worst. It should be noted that the performance deficit of the equal-weight combination rule is smaller when experts' error terms are positively correlated (rather than independent or negatively correlated), a condition that seems likely to obtain in practice.

The copula and frequentist decision makers may not perform as well relative to other combination methods in cases with more, or less well-calibrated, experts. The copula decision maker concentrates probability on values to which all experts assign significant probability and very little probability to values to which any expert assigns small probability (including zero to regions to which any expert assigns zero probability). This contributes to its higher informativeness than other combination methods (consistent with Kallen and Cooke 2002 and Hora 2004). With more experts, or with less well-calibrated experts, it is more likely that some expert will assign very small probability to values the other experts judge plausible, making these combination rules highly sensitive to the set of experts that are included. For example, a copula combination of 16 experts' judgments on climate sensitivity to greenhouse gases (Morgan and Keith 1995) is extremely sensitive to whether a particular expert is included or excluded while the equal-weight combination is only slightly affected by this choice (Hammitt and Shlyakhter 1999).[7] In general, the equal-weight combination is likely to be the least sensitive of the combination methods considered to variation in the experts who are initially selected.

Overall, our results suggest that expert-judgment studies should use one or more of the alternative methods as a substitute, or at least a complement, to the equal-weight combination. To do so, studies must elicit experts' judgments on seed variables that can be used to evaluate their performance, individually (for the performance and best-expert decision makers) or jointly (for the copula and frequentist decision makers).

[7] This expert provided a very tight distribution with its support entirely outside the other experts' inter-quartile ranges.

References

Clemen, R.T., Comment on Cooke's classical method, *Reliability Engineering and System Safety* 93: 760-765, 2008.

Clemen, R.T., and T. Reilly, Correlations and copulas for decision and risk analysis, *Management Science* 45: 208-224, 1999.

Clemen, R.T., and R.L. Winkler, Aggregating probability distributions, *Advances in Decision Analysis: From Foundations to Applications* (W. Edwards, R.F. Miles, and D. von Winterfeldt, eds.), Cambridge University Press, 154-176, 2007.

Clemen, R.T., and R.L. Winkler, Combining probability distributions from experts in risk analysis, *Risk Analysis* 19: 187-203, 1999.

Clemen, R.T., and R.L. Winkler, Limits for the precision and value of information from dependent sources, *Operations Research* 33: 427-442, 1985.

Clemen, R.T., G.W. Fischer, and R.L. Winkler, Assessing dependence: some experimental results, *Management Science* 46: 1100-1115, 2000.

Cooke, R.M., *Experts in Uncertainty: Opinion and Subjective Probability in Science*, Oxford University Press, New York, 1991.

Cooke, R.M., Response to discussants, *Reliability Engineering and System Safety* 93: 775-777, 2008.

Cooke, R.M., and L.H.J. Goossens, TU Delft expert judgment data base, *Reliability Engineering and System Safety* 93: 657-674, 2008.

Cooke, R.M., A.M. Wilson, J.T. Toumisto, O. Morales, M. Tanio, and J.S. Evans, A probabilistic characterization of the relationship between fine particulate matter and mortality: elicitation of European experts, *Environmental Science and Toxicology* 41: 6598-6605, 2007.

Curtwright, A., M.G. Morgan, and D. Keith, Expert assessment of future photovoltaic technology, *Environmental Science and Technology* 42: 9031-9038, 2008.

EPA, Science Policy Council, *Expert Elicitation Task Force White Paper*, External Review Draft and Addendum: Selected Recent (2006-2008) Citations, January 6, 2009.

Flandoli, F., E. Giorgi, W.P. Aspinall, and A. Neri, Comparison of a new expert elicitation model with the classical model, equal weights and single experts, using a cross-validation technique, *Reliability Engineering and System Safety* 96: 1292-1301, 2011.

Hammitt, J.K., and A.I. Shlyakhter, The expected value of information and the probability of surprise, *Risk Analysis* 19: 135-152, 1999.

Hora, S.C., Probability judgments for continuous quantities: linear combinations and calibration, *Management Science* 50: 567-604, 2004.

Hora, S.C., An analytic method for evaluating the performance of aggregation rules for probability densities, *Operations Research* 58: 1440-1449, 2010.

Jouini, M.N., and R.T. Clemen, Copula methods for aggregating expert opinions, *Operations Research* 44: 444-457, 1996.

Kallen, M.J., and R.M. Cooke, Expert aggregation with dependence, *Probabilistic Safety Assessment and Management* (E.J. Bonano, A.L. Camp, M.J. Majors, and R.A. Thompson, eds.), Elsevier, 1287-1294, 2002.

Keith, D.W., When is it appropriate to combine expert judgments? *Climatic Change* 33: 139-143, 1996.

Lin, S.-W., and V.M. Bier, A study of expert overconfidence, *Reliability Engineering and System Safety* 93: 711-721, 2008.

Lin, S.-W., and C.-H. Cheng, The reliability of aggregated probability judgments obtained through Cooke's classical model, *Journal of Modelling in Management* 4: 149-161, 2009.

Morgan, M.G., and M. Henrion, *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press, 1990.

Morgan, M.G., and D.W. Keith, Subjective judgments by climate experts, *Environmental Science and Technology* 29: 468A-476A, 1995.

Morgan, M.G., L.F. Pitelka, and E. Shevliakova, Elicitation of expert judgments of climate change impacts on forest ecosystems, *Climatic Change* 49: 279-307, 2001.

Morgan, M.G., P. Adams, and D.W. Keith, Elicitation of expert judgments of aerosol forcing, *Climatic Change* 75: 195-214, 2006.

Roman, H.A., K.D. Walker, T.L. Walsh, L. Conner, H.M. Richmond, B.J. Hubbell, and P.L. Kinney, Expert judgment assessment of the mortality impact of changes in ambient fine particulate matter in the U.S., *Environmental Science and Technology* 42: 2268-2274, 2008.

Tversky, A., and D. Kahneman, Judgment under uncertainty: heuristics and biases, *Science* 185: 1124-1131, 1974.

Walker, K.D., P. Catalano, J.K. Hammitt, and J.S. Evans, Use of expert judgment in exposure assessment: Part 2. Calibration of expert judgments about personal exposures to benzene, *Journal of Exposure Analysis and Environmental Epidemiology* 13: 1-16, 2003.

Winkler, R.L., and R.M. Poses, Evaluating and combining physicians' probabilities of survival in an intensive care unit, *Management Science* 39: 1526-1543, 1993.

Zickfield, K., A. Levermann, T. Kuhlbrodt, S. Rahmstorf, M.G. Morgan, and D. Keith, Expert judgments on the response of the Atlantic meriodional overturning circulation response to climate change, *Climatic Change* 82: 235-265, 2007.

Table 1. Means and variances of combined distributions (mean and interquartile range)

Column label		A	B	C	D	E	F
Covariance matrix of experts' errors		$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & -1/2 \\ -1/2 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$	$\begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$	$\begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$
Mean							
	Performance	0.003 (-.114, 0.116)	0.000 (-0.119, 0.120)	-0.001 (-0.110, 0.108)	-0.001 (-0.137, 0.134)	0.003 (-0.137, 0.142)	-0.001 (-0.138, 0.136)
	Copula	0.003 (-0.087, 0.092)	0.001 (-0.112, 0.113)	0.000 (-0.065, 0.066)	-0.001 (-0.116, 0.117)	0.002 (-0.127, 0.131)	-0.000 (-0.086, 0.084)
	Frequentist	0.003 (-0.087, 0.092)	0.001 (-0.107, 0.109)	0.000 (-0.068, 0.068)	-0.001 (-0.115, 0.114)	0.002 (-0.132, 0.134)	0.000 (-0.089, 0.089)
	Equal weight	0.002 (-0.085, 0.085)	0.001 (-0.106, 0.108)	0.000 (-0.061, 0.062)	-0.001 (-0.140, 0.140)	0.003 (-0.160, 0.166)	-0.001 (-0.110, 0.108)
	Best expert	0.003 (-0.122, 0.124)	0.000 (-0.122, 0.124)	-0.001 (-0.125, 0.121)	0.002 (-0.128, 0.132)	0.004 (-0.126, 0.131)	0.001 (-0.130, 0.131)
Variance							
	Performance	1.088 (1, 1)	1.059 (1, 1.109)	1.113 (1, 1)	2.286 (1, 3.676)	1.989 (1, 2.544)	2.524 (1, 4)
	Copula	0.451 (0.295, 0.571)	0.673 (0.447, 0.845)	0.227 (0.149, 0.227)	0.717 (0.467, 0.912)	0.904 (0.593, 1.144)	0.386 (0.253, 0.489)
	Frequentist	0.455 (0.297, 0.576)	0.698 (0.464, 0.878)	0.233 (0.156, 0.291)	0.746 (0.487, 0.746)	1.086 (0.702, 1.385)	0.441 (0.318, 0.537)
	Equal weight	1.499 (1.406, 1.579)	1.250 (1.203, 1.289)	1.754 (1.615, 1.875)	3.751 (3.520, 3.954)	3.523 (3.114, 3.373)	4.246 (3.931, 4.510)
	Best expert	1 (1, 1)	1 (1, 1)	1 (1, 1)	1.285 (1, 1)	1.214 (1, 1)	1.317 (1, 1)

Table 2. Performance characteristics of combined distributions (mean and interquartile range)

Column label		A	B	C	D	E	F
Covariance matrix of experts' errors		$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 1/2\\1/2 & 1\end{bmatrix}$	$\begin{bmatrix}1 & -1/2\\-1/2 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & 4\end{bmatrix}$	$\begin{bmatrix}1 & 1\\1 & 4\end{bmatrix}$	$\begin{bmatrix}1 & -1\\-1 & 4\end{bmatrix}$
Calibration							
	Performance	0.382 (0.094, 0.633)	0.420 (0.151, 0.675)	0.380 (0.073, 0.647)	0.342 (0.073, 0.577)	0.412 (0.141, 0.671)	0.345 (0.042, 0.618)
	Copula	0.326 (0.055, 0.556)	0.320 (0.052, 0.539)	0.325 (0.055, 0.556)	0.317 (0.050, 0.525)	0.324 (0.055, 0.558)	0.320 (0.053, 0.525)
	Frequentist	0.327 (0.055, 0.562)	0.336 (0.066, 0.566)	0.316 (0.047, 0.539)	0.330 (0.059, 0.563	0.339 (0.072, 0.577)	0.345 (0.077, 0.577)
	Equal weight	0.068 (0.003, 0.073)	0.309 (0.067, 0.511)	0.001* (0.000, 0.000)	0.067 (0.003, 0.069)	0.306 (0.073, 0.500)	0.001* (0.000, 0.000)
	Best expert	0.447 (0.179, 0.726)	0.451 (0.182, 0.726)	0.446 (0.175, 0.726)	0.442 (0.173, 0.703)	0.448 (0.179, 0.703)	0.443 (0.175, 0.703)
Informativeness							
	Performance	0.223 (0.212, 0.256)	0.180 (0.158, 0.204)	0.257 (0.248, 0.295)	0.403 (0.217, 0.578)	0.410 (0.329, 0.549)	0.393 (0.148, 0.604)
	Copula	0.533 (0.401, 0.641)	0.355 (0.239, 0.446)	0.817 (0.686, 0.930)	0.737 (0.605, 0.850)	0.654 (0.522, 0.761)	0.983 (0.850, 1.098)
	Frequentist	0.529 (0.406, 0.637)	0.339 (0.223, 0.432)	0.803 (0.681, 0.914)	0.718 (0.591, 0.833)	0.563 (0.432, 0.672)	0.921 (0.818, 1.016)
	Equal weight	0.121 (0.118, 0.124)	0.124 (0.121, 0.126)	0.119 (0.116, 0.122)	0.199 (0.194, 0.204)	0.210 (0.206, 0.214)	0.192 (0.188, 0.197)
	Best expert	0.245 (0.230, 0.260)	0.198 (0.189, 0.209)	0.281 (0.263, 0.299)	0.544 (0.569, 0.600)	0.524 (0.543, 0.565)	0.565 (0.592, 0.627)

Combined score							
	Performance	0.092 (0.019, 0.156)	0.077 (0.026, 0.122)	0.106 (0.020, 0.182)	0.143 (0.022, 0.221)	0.171 (0.044, 0.280)	0.134 (0.014, 0.174)
	Copula	0.158 (0.027, 0.261)	0.098 (0.016, 0.155)	0.249 (0.043, 0.413)	0.217 (0.036, 0.364)	0.196 (0.034, 0.326)	0.298 (0.051, 0.498)
	Frequentist	0.158 (0.027, 0.262)	0.102 (0.019, 0.161)	0.238 (0.036, 0.402)	0.225 (0.042, 0.375)	0.180 (0.036, 0.294)	0.308 (0.069, 0.512)
	Equal weight	0.008 (0.000, 0.009)	0.038 (0.008, 0.063)	0.000* (0.000, 0.000)	0.013 (0.001, 0.014)	0.064 (0.015, 0.104)	0.000* (0.000, 0.000)
	Best expert	0.110 (0.043, 0.173)	0.089 (0.036, 0.139)	0.125 (0.048, 0.198)	0.214 (0.077, 0.394)	0.243 (0.079, 0.381)	0.249 (0.082, 0.407)
Note: For equal-weight decision maker, mean is outside interquartile range in cells marked with *. Column C: max cal = 0.343, max weight = 0.039; column F: max cal = 0.219, max weight = 0.045.							

Table 3. Probability that row decision maker is superior to column decision maker
(* indicates significantly different from 50 percent at 5 percent significance level)

A. Equal variance, independence

	Copula	Frequentist	Performance	Best expert	Equal weight
Copula	-	60*	61*	55	89*
Frequentist	40*	-	61*	55	90*
Performance	39*	39*	-	2*	86*
Best expert	45	45	98*	-	95*
Equal weight	11*	10*	14*	5*	-

B. Equal variance, positive dependence

	Copula	Frequentist	Performance	Best expert	Equal weight
Copula	-	56	52	47	68*
Frequentist	44	-	53	49	71*
Performance	48	47	-	7*	71*
Best expert	53	51	93*	-	77*
Equal weight	32*	29*	29*	23*	-

C. Equal variance, negative dependence

	Copula	Frequentist	Performance	Best expert	Equal weight
Copula	-	58*	67*	62*	98*
Frequentist	42*	-	66*	60*	98*
Performance	33*	34*	-	0*	90*
Best expert	38*	40*	100*	-	99*
Equal weight	2*	2*	10*	1*	-

D. Unequal variance, independence

	Copula	Frequentist	Performance	Best expert	Equal weight
Copula	-	48	60*	43	89*
Frequentist	52	-	62*	45	90*
Performance	40*	38*	-	10*	89*
Best expert	57	55	90*	-	95*
Equal weight	11*	10*	11*	5*	-

E. Unequal variance, positive dependence

	Copula	Frequentist	Performance	Best expert	Equal weight
Copula	-	53	51	40	71*
Frequentist	47	-	50	38*	72*
Performance	49	50	-	17*	75*
Best expert	60	62*	83*	-	83*
Equal weight	29*	28*	25*	17*	-

F. Unequal variance, negative dependence

	Copula	Frequentist	Performance	Best expert	Equal weight
Copula	-	48	67*	51	98*
Frequentist	52	-	70	54	99*
Performance	33*	30	-	6*	99*
Best expert	49	46	94*	-	100*
Equal weight	2*	1*	1*	0*	-