Safety, Reliability and Risk Analysis: Beyond the Horizon – Steenbergen et al. (Eds)
© 2014 Taylor & Francis Group, London, ISBN 978-1-138-00123-7

Experimental results about the assessments of conditional rank correlations by experts: Example with air pollution estimates

O. Morales-Nápoles
TNO, Structural Reliability and TU Delft, DIAM, Delft, The Netherlands

A.M. Hanea
TU Delft, DIAM, Delft, The Netherlands

D.T.H. Worm
TNO, Performance of Networks and Systems, Delft, The Netherlands

ABSTRACT: Science-based models often involve substantial uncertainty that must be quantified in a defendable way. Shortage of empirical data inevitably requires input from expert judgment. How this uncertainty is best elicited can be critical to a decision process, as differences in efficacy and robustness of the elicitation methods can be substantial. When performed rigorously, expert elicitation and pooling of experts' opinions can be powerful means for obtaining rational estimates of uncertainty.

Causes of uncertainty may be interrelated and may introduce dependencies. Ignoring these dependencies may lead to large errors. Dependence modelling is an active research topic, and methods for dependence elicitation are still very much under development. Dependence measures such as rank correlations are commonly used in different types of models. Eliciting rank correlations and conditional rank correlations from experts have been proposed and used in the past. Conditional rank correlations are not elicited directly from experts, rather the experts are asked to estimate some other related quantities. In this paper two methods for eliciting conditional rank correlations via related quantities are compared in order to obtain insight about which of the two renders more accurate estimates of conditional rank correlations. Our data shows that good performance in uncertainty assessments does not automatically translates into good performance in dependence estimates. We show that, analogously to uncertainty estimates, combining experts' estimates of dependence according to their performance results in better estimates of the dependence structure.

1 INTRODUCTION

Dependence measures such as rank correlations are commonly used in different types of models. Whenever field data is available rank correlations may be estimated directly from data. However, many times field data is not available. In such situations, one needs to make use of a structured protocol for the elicitation of expert opinions. Moreover, it has been recognized that when several uncertain quantities are elicited, there is a need to elicit the dependence structure between them (French 2011).

Methods for eliciting rank correlations from experts have been proposed and used in the past. See Cooke & Goossens (1999), Clemen & et al. (1999) and Clemen & et al. (2000) for example. One of the options is directly asking experts for an estimate of the rank correlation between pairs of variables. Another option is asking experts for estimates of some other quantity, for example a conditional probability of exceedance or probabilities of concordance or discordance, and use these to estimate rank correlations (under certain copula assumptions). Though not conclusive, previous results indicate that the most accurate way to obtain a subjective measure of bivariate dependence is simply to ask the expert to estimate the correlation between the two variables in question (Clemen & et al., 2000).

Recently, Non-Parametric Bayesian Networks (NPBN) have been introduced as flexible tools for applications where dependence modeling is important. See for example Ale et al. (2007), Hanea & Ale (2009). The inputs for these models are univariate marginal distributions and rank and conditional rank correlations. When field data is not available, rank and conditional rank correlations have been assessed from experts through the elicitation of Conditional Probabilities of Exceedance

(CPE, see Morales et al. (2008)) or ratios of rank correlations (RRC, see Morales-Nápoles et al. (2013)). To our knowledge, there is no experimental study available that would give some indication as to the accuracy of experts in estimating conditional rank correlations. Hence even less can be expected about evidence of one option being preferable than the other. In this study we describe data collected from a controlled exercise. The question of interest is whether experts can estimate more accurately conditional rank correlations through estimates of conditional probabilities of exceedence or through directly estimating bivariate rank correlation coefficients.

2 THE EXERCISE

A pilot study was conducted at the TU Delft on December 2011. The objective was to obtain data to start addressing the question of interest. We gathered a group of 14 experts. The group consisted of 9 graduate students from the TU Delft with formal training on statistics and 5 researchers from the TU Delft and TNO.

For the exercise two sets of data were used. Both describe the relationship between SO_2 emissions and concentrations of fine particulate matter, $PM_{2.5}$ in Alabama, United States. This problem is of interest because $PM_{2.5}$ exposure has been associated with adverse health effects. We however did not concentrate on these health effects but rather in models describing the relationship between pollutants. This data has been used before in the context of NPBN (Hanea & Harrington 2009).

The two data sets were generated from the model in Figure 1. The model consists of variables $X_1, \dots, X_6$. Variables $X_1, \dots, X_5$ are SO_2 monthly emissions gathered from electricity generating stations. Variable X_6 corresponds to monthly mean concentrations of $PM_{2.5}$ gathered from a collection site. The naming of variables is based on the direction and distance from the monitoring site. Ring "a" consists of power plants within 100 miles of the monitor site, ring "b" to 100–250 miles. The numbers 1, 3, 5, 6 and 8 in the variables' names correspond to several zones from a total of 8 zones whose bisectors are the compass directions NNW, proceeding counterclockwise to NNE.

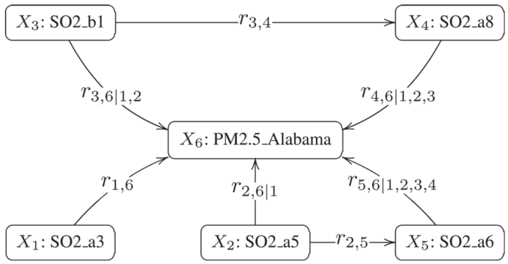


Figure 1. BN of interest.

By specifying one dimensional marginal distributions, conditional rank correlations, the directed acyclic graph and a family of copulae, the joint distribution is completely and uniquely determined. For more details regarding NPBNs the reader is referred to Hanea et al. (2006). It is sufficient here to say that choosing the normal copula to realise the rank and conditional rank correlations assigned to the arcs of a NPBN offers many computational advantages. For this reason the normal copula assumption is usually used and validated when data allows. In this exercise, marginal distributions for both data sets were the same but the dependence structures were different. Model 1 was quantified directly from the original data. This model was validated for the normal copula assumption with the techniques presented in Hanea & Harrington (2009). All rank correlations in Model 1 are positive. For Model 2 the magnitude of the rank correlations was changed. Additionally, one of the unconditional rank correlations in Model 2 was chosen to be negative. Both data sets used in the experiment were generated under the normal copula assumption. This was explicitly mentioned to the participants.

Two large data sets were produced with Models 1 and 2 and sent to the 14 experts with some background information about the data and the type of questions to be asked in order to assess the dependence measures of interest. One week later a half day workshop was held at the TU Delft where both subjects (the data and the methods to be used) were further discussed with the group. The group of experts was further divided into two groups of 7 experts each. Group 1 consists of experts A, B, C, H, I, K, N and group two of experts D, E, F, G, J, L, M. Group 1 was asked for the dependence measures of interest in Model 1 with CPE and in Model 2 with RRC. Inversely, group 2 was asked CPE to obtain the measures of interest in Model 2 and RRC for Model 1 (see Table 2). We will denote further the four cases as M1RRC, M1CPE, M2RRC and M2CPE

The type of questions asked for CPE were:

1. *Consider model i. There are $N_{1,i}$ samples (out of* 500,000) *for which variable SO2_a*3 *is at least* 10,466 *(median). Consider the indices of all variables corresponding to this subset. In other words, conditionalize on this subset. In how many of these indices will the value of PM2.5_ Alabama be at least* 14.82 *(median)*?

⋮

5. *Consider model i. There are $N_{5,i}$ samples (out of* 500,000) *for which variable SO2_a*3 *is at*

least 10,466 *(median), SO2_a5 is at least* 7,256 *(median), SO2_b*1 *is at least 26,091 (median), SO2_a*8 *(median) is at least* 3,429 *(median) and SO*2*_a*6 *is at least* 21,908 *(median). Consider the indices of all variables corresponding to this subset. In other words, conditionalize on this subset. In how many of these indices will the value of PM*2.5*_Alabama be at least* 14.82 *(median)*?

For RRC:

6. *Consider model i. What is the rank correlation between SO*2*_a*3 *and PM*2.5*_Alabama?*

⋮

10. *Consider model i. What is the ratio of the rank correlation between SO*2*_a*6 *and PM*2.5*_Alabama to the rank correlation between SO*2*_a*3 *and PM*2.5*_Alabama*?

A total of 20 questions were asked to each participant, 10 of which were additionally used as calibration variables. The classical method for structured expert judgment (Cooke 1991) was used to investigate experts' performance as uncertainty assessors.

3 RESULTS

3.1 *Calibration & information*

The classical model is a performance-based linear pooling model based on statistical hypothesis testing. It aggregates individual experts' PDFs in order to obtain one combined PDF for each variable. Experts give pre-defined quantiles of distributions, typically 5%, 50% and 95%. Experts can be weighted equally or according to their (relative) expertise, as indicated by their performance on seed variables. Seed (calibration) variables are variables from the experts' field whose realizations are (or will be) known to the analysts, but unknown to the experts.

The individual experts' weights are based on two quantitative measures of performance: calibration and information. Calibration measures the statistical likelihood that the realizations of the seed variables correspond, in a statistical sense, with an expert's assessments. If this likelihood score is below a certain cut-off level, the expert is unweighted. The cut-off could be chosen by the analyst or determined by optimizing the performance of the combined virtual expert.

The calibration score takes values between 0 and 1, with a high score implying that the expert's PDFs are statistically supported by the set of seed variables. Information represents the degree to which an expert's PDFs are concentrated, relative to some chosen background measure, and it is always positive. Good uncertainty assessors are those exhibiting good calibration and high information. The virtual (combined) expert resulting from the combination of experts' opinions will also have a calibration and an information score. The individual experts' performance-based weights are proportional to the product of calibration and information. For a detailed discussion about the classical method see Cooke (1991).

Table 1 shows calibration and information scores for groups 1 and 2. At the bottom, the equal weight combination (all experts assigned equal weights regardless of their performance on seed variables) and the performance base combination (Global) are also shown. Notice that for group 1, the equal weight combination is better calibrated (by far) than individual experts. For group 2 the equal weight combination is better calibrated than every individual expert except expert F who has a rather high calibration score. The price for high calibration in the case of the equal weight combinations is wide distributions (low information score).

The performance weight combination in the case of group 1 is better calibrated than individual experts and also than the equal weight combination. The information score is however on the order of the less informative experts from this group, but still higher that the information of the equally weighted combination. In group 1 the performance based combination would be formed with experts A and B. Given experts' A and B calibration and information scores (and their respective products) one can notice that the normalized weight of expert A is 0.93, whereas that of expert B is 0.07. Even though both experts are added to the combination, the features of expert's A distribution will dominate. For group 2, the performance based combination would give all weight to expert F.

Table 1. Calibration and information scores for experts.

Group 1			Group 2		
Id.	Calibr.	Inform.	Id.	Calibr.	Inform.
A	0.0139	2.092	D	0.0357	2.745
B	0.0013	1.662	E	0.0063	1.497
C	1.3×10^{-8}	1.89	F	0.7069	0.7571
H	4.1×10^{-6}	2.336	G	5.9×10^{-4}	1.86
I	4.9×10^{-7}	1.474	J	2.4×10^{-10}	2.49
K	0.0011	1.209	L	0.0028	1.169
N	3.5×10^{-8}	2.378	M	0.00131	3.84
Eq.	0.2282	0.0263	Eq.	0.5503	0.3009
Gl.	0.8283	1.459	Gl.	0.7069	0.7571

3.2 *Individual estimates of conditional rank correlations*

For each case of interest M1RRC, M1CPE, M2RRC and M2CPE we have a total of 49 estimates. That is, seven expert estimating each of the seven (conditional) rank correlations shown in figure 1. We denote the true estimates (unknown to the expert) $r_{i,j|D}$ where D corresponds to the conditioning set corresponding to the arc of interest from the NPBN in figure 1. The corresponding estimate for different experts is denoted as $r^e_{i,j|D}$. The absolute difference between the true estimates and experts' individual answers for each particular case: $\delta_e^c = | r_{i,j|D} - r^e_{i,j|D} |$ where $c \in$ {M1RRC, M1CPE, M2RRC, M2CPE}, is shown in Figure 2.

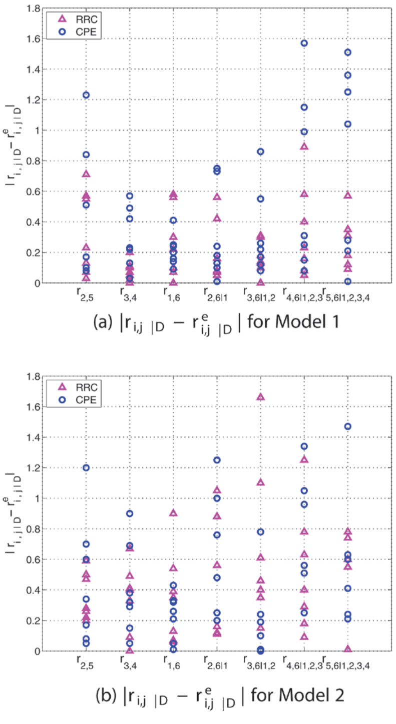
Values for $\delta_e^{\text{M1CPE}}, \delta_e^{\text{M2RRC}}$ and δ_e^{M2CPE} are similar and larger than δ_e^{M1RRC}. The largest value for δ_e^{M1RRC} is smaller than 1. This does not hold for all other models. The average across experts $\bar{\delta}_{\text{M1CPE}} = 0.43, \bar{\delta}_{\text{M2RRC}} = 0.46$ and $\bar{\delta}_{\text{M2CPE}} = 0.49$ are similar andlarger than $\bar{\delta}_{\text{M1RRC}} = 0.23$. The average across all observations in Figure 2 is $\delta_e^c = 0.4\cdot$

One way to investigate whether the estimates for cases M1RRC, M1CPE, M2RRC and M2CPE are statistically different is through a two-sample Kolmogorov-Smirnov test. This is based on the distribution of $D_{m,n} = \text{l.u.b.} | S_m(x) - T_n(x) |$, where $S_m(x)$ and $T_n(x)$ are the empirical distribution functions of two samples of size m and n respectively of mutually independent random variables having a common distribution function F (see Feller (1948) and Feller (1950)).

A two-sample Kolmogorov-Smirnov test is performed for all six possible pairs of distributions to be compared δ_e^c. The hypothesis that the δ_e^{M1RRC} and δ_e^{M2RRC} have the same distribution is rejected at the 5% level. Similar results are obtained for δ_e^{M1RRC} and δ_e^{M2CPE}. For all other pairs the hypothesis that the distributions are different cannot be rejected on the basis of this test. It is worth stressing that the distribution that seems significantly different than the others is that of δ_e^{M1RRC}. Recall that Model 1 was the true model that was quantified with the original data and hance made more physical sense.

Another way to investigate the "homogeneity" of data is through analysis of variance (ANOVA). The hypothesis to be tested is H_0: $\bar{\delta}_{\text{M1CPE}} = \bar{\delta}_{\text{M2RRC}} = \bar{\delta}_{trmM2CPE} = \bar{\delta}_{\text{M1RRC}}$. The hypothesis is tested by comparing two unbiased estimates of σ^2 which is the unknown but equal variance across sub-populations. One of these is based on the variation from sample to sample and the other one on the variations within samples, that is, *Total SS* = *SST* + *SSE*. Where *Total SS* stands for total sum of squares, *SST* for sum of squares for treatments or between groups and *SSE* for sum of squares of errors or within a group. If each term in the sum of squares in *SST* and *SSE* comes from a normal distribution with equal variance then the ratio $\frac{SST/(k-1)}{SSE/(n_1+\ldots+n_k-k)}$ should come from an F-distribution with $k-1$ and $(n_1+\ldots+n_k-k)$ degrees of freedom. The p-value for the hypothesis that $\bar{\delta}_{\text{M1CPE}} = \bar{\delta}_{\text{M2RRC}} = \bar{\delta}_{\text{M2CPE}} = \bar{\delta}_{\text{M1RRC}}$ is 0.0016. This casts statistical evidence to reject the null hypothesis[1].

In order to get further insight into the differences of means the Tukey test based on "allowances" is used (see Ramachandran & Tsokos (2009, ch.10) and Duncan (1955, p. 29–31)). Tukey's procedure estimates confidence intervals or "allowances" for H_0: $\bar{\delta}_i - \bar{\delta}_j$ from a randomized design such as the one way ANOVA procedure described briefly above. Tukey showed that if $\bar{\delta}_i^*$, $i = 1, \ldots, k$ denote the sample means computed with equal sample size and $\bar{\delta}_i$ the true means, then the probability that all $\binom{k}{2}$ differences $\bar{\delta}_i - \bar{\delta}_j$ will simultaneously satisfy the inequalities:



(a) $|r_{i,j|D} - r^e_{i,j|D}|$ for Model 1

(b) $|r_{i,j|D} - r^e_{i,j|D}|$ for Model 2

Figure 2. Absolute difference between the true estimates and experts individual answers $| r_{i,j|D} - r^e_{i,j|D} |$ for Models 1 and 2 with RRC and CPE.

[1] If sample sizes are equal ANOVA is robust to violation of normality and equal variance assumptions (Ramachandran & Tsokos 2009, ch.10).

$$(\bar{\delta}_i^* - \bar{\delta}_j^*) - q_{\alpha,k,(n-1)k}\sqrt{\frac{SSE/(n_1+\ldots+n_k-k)}{n}} \le \bar{\delta}_i - \bar{\delta}_j \le$$
$$(\bar{\delta}_i^* - \bar{\delta}_j^*) + q_{\alpha,k,(n-1)k}\sqrt{\frac{SSE/(n_1+\ldots+n_k-k)}{n}}$$

is $(1-\alpha)$, where $q_{\alpha,k,v}$ is the upper α critical value of the Studentized range distribution based on k, v degrees of freedom. If for a given i and j zero is not contained in the interval above inequality, H_0: $\bar{\delta}_i = \bar{\delta}_j$ may be rejected at the α significance level. The Tukey procedure applied to the data shown in Figure 2 would render $\bar{\delta}_{\text{M1RRC}}$ significantly different fromn $e\delta_{\text{M1CPE}}$, $\bar{\delta}_{\text{M2RRC}}$ and $\bar{\delta}_{\text{M2CPE}}$ at the 5% level. The same hypothesis of equality of means would not be rejected for all other pairs of means for the same significance level.

Finally, it is worth mentioning that there appears to be a positive correlation between δ_e^c and the cardinality of the set D. This would indicate that larger errors would tend to appear in the elicitation of higher order conditional rank correlations. These correlations are 0.23, 0.35, 0.06 and 0.21 for M1RRC, M1CPE, M2RRC and M2CPE respectively. Nevertheless, given the sample size, only values higher than 0.27 (at a 0.05 level of significance for a nondirectional—two-tailed—test) can be considered as significantly different than zero (Fisher & Yates 1974).

3.3 *Individual models*

The main question of interest is whether experts can approximate a multivariate model through one or the other technique investigated in this paper to a desired level of accuracy. In Hanea et al. (2010) the determinant of the correlation matrix of a NPBN is proposed as a summary measure of dependence to be used in a data mining procedure. The main reason is that the determinant may be written as a function of the partial correlations corresponding to the arcs of the NPBN. Notice that the arcs in a NPBN are associated with conditional rank correlations. Under the normal copula assumption, the corresponding partial correlations may be calculated. From the partial correlation specification and the (conditional) independence statements embedded in the graph, the correlation matrix of interest may be computed.

Theorem Let K be the determinant of an n-dimensional correlation matrix ($K \ge 0$). For any partial correlation NPBN specification

$$K = \prod_{e \in E(NPBN)} \left(1 - \rho^2_{i,j|D_{i,j}}\right) \tag{1}$$

where $\rho_{i,j|D_{i,j}}$ is the partial correlation associated with the arc between nodes i and j, with conditioning set $D_{i,j}$, and the product is taken over the set of edges ($E(NPBN)$) in the NPBN.

The determinant K of an n-dimensional correlation matrix for the partial correlation NPBN specification can take values between 0 and 1. Where 1 corresponds to independence and 0 to perfect linear dependence. In order to address the question of interest one would like a protocol by which one could decide if experts have approximated sufficiently the target correlation matrix. The first idea could be to follow the protocol described in Hanea et al. (2010) for data mining which is based on the determinant of the correlation matrix. For example one might sample a number of times from the normal distribution with correlation matrix corresponding to the model of interest. For each sample compute the correlation matrix and its determinant thus obtaining an empirical distribution. Then observe whether experts' estimated correlation matrix falls within the empirical distribution. However, the determinant of the correlation matrix in the case of expert judgments is not an appropriate test statistic. This may be seen in the following lemma.

Lemma Fix $K \in (0,1)$ and a NPBN structure (the DAG of a NPBN). There exist infinitely many different partial correlation NPBN specifications, hence infinitely many correlation matrices with determinant K.

Proof. Let K be in (0, 1) and assume we have a NPBN structure given, but the values of the partial correlations are still free to choose. Rewrite equation 1 as $K = \prod_{k=1}^{M}(1-\rho_k^2)$ where M is the number of edges in the NPBN structure. We can specify any value $\in (-1,1)$ for the ρ_k. This is an M-dimensional function, whose equality to K describes an $M-1$ dimensional hyperplane, which will have infinitely many points. More rigorously: First of all, we can choose any, thus infinitely many $\rho_1 \in (-\sqrt{(1-K)}, \sqrt{(1-K)})$. Then $(1-\rho_1^2) > K$. Then $\prod_{k=2}^{M}(1-\rho_k^2)$ should be equal to $K_1 := K/(1-\rho_1^2)$, and K_1 takes values $\in (K,1)$. Now we can continue this process by choosing infinitely many $\rho_2 \in (-\sqrt{(1-K_1)}, \sqrt{(1-K_1)})$. Continue in the same way until we reach ρ_M which may be set equal to $\pm(\sqrt{1-K/\prod_{k=1}^{M-1}(1-\rho_k^2)})$.

The lemma above entails that different experts may provide different dependence estimates that will yield the same (or approximately the same) determinant for the NPBN structure of interest. This is obviously not desirable if the procedure previously described based on the determinant of experts' correlation matrices was to be used to decide upon experts dependence estimates. Even though in practice the number of correlation matrices with the same determinant is not strikingly high in all situations, especially in higher dimensions, for determinants larger than 10^{-2}, we would prefer a different measure of performance. Instead of a summary measure of dependence we would like to use a measure of distance for our protocol for deciding if experts have approximated sufficiently the target correlation matrix. In this example we know that the underlying copula corresponds to the normal copula. Hence we would look at measures of distance for the multivariate Gaussian distribution. In Abou Moustafa et al. (2010) several measures of distance between Gaussian densities are discussed. For the rest we consider the Heillinger distance $d_H(N_1, N_2) = \sqrt{1 - \eta(N_1, N_2)}$ where $N_1(\mu_1, \Sigma_1)$ and $N_2(\mu_2, \Sigma_2)$ are two gaussian densities with covariance matrices Σ_1, Σ_2, and vector means μ_1, μ_2, and η is as in equation 2:

$$\eta(N_1, N_2) = \frac{det(\Sigma_1)^{\frac{1}{4}} det(\Sigma_2)^{\frac{1}{4}}}{det(\frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2)^{\frac{1}{2}}} \times \exp\{-\frac{1}{8}(\mu_1 - \mu_2)^T \frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2(\mu_1 - \mu_2)\} \qquad (2)$$

When assuming the normal copula the marginal distributions are transformed to standard normals, hence the exponent term in equation 2 vanishes and Σ_1, Σ_2 correspond to correlation matrices. Moreover, in our case the Heillinger distance satisfies the usual axioms of a metric: it equals zero iff $\Sigma_1 = \Sigma_2$, it is symmetric, and it satisfies the triangle inequality. Its maximum value is 1, which it reaches if $det(\Sigma_1) = 0$ (there is perfect dependence between certain variables) and $det(\Sigma_2) = 1$ (independence) or vice versa. Table 2 presents results of computing the Heillinger distance between the real rank correlation matrix and the elicited rank correlation matrix, per model and elicitation technique of interest for each of the 14 experts.

The three smallest distances (< 0.3) are observed between the dependence structure given by experts G, M and D for M1RRC. The largest distances (> 0.9) are observed between expert L dependence structure for M1RRC, and between expert F dependence structure for M2CPE. The correlation matrices for the two experts with smallest value of d_H per model are shown in Table 3. In general smaller values for d_H are observed for model 1 regardless of the elicitation technique.

Averaging across experts M1CPE shows smaller average value (0.49) for d_H. The average value across experts for M1RRC is (0.52). The averages for M2RRC and M2CPE are 0.75 and 0.69 respectively. An ANOVA analysis as in previous section based on d_H would indicate no significant difference between the four groups.

Following our previous discussion, a procedure for deciding whether an expert has approximated sufficiently the target correlation matrix is to construct the empirical distribution of $d_H(\Sigma_m, \Sigma_{m,sample})$ where Σ_m corresponds to the target correlation matrix and $\Sigma_{m,sample}$ to the correlation matrix

Table 2. Results from expert judgment elicitation of dependence.

Expert	A	B	C	H	I	K	N
	m = M1 t = M1CPE						
$d_H(\Sigma_m, \Sigma_{e,t})$	0.54	0.35	0.68	0.72	0.36	0.38	0.37
	m = M2 t = M2RRC						
$d_H(\Sigma_m, \Sigma_{e,t})$	0.87	0.48	0.90	0.68	0.82	0.83	0.69
Expert	D	E	F	G*	J	L	M
	m = M1 t = M1RRC						
$d_H(\Sigma_m, \Sigma_{e,t})$	0.29	0.49	0.88	0.13	0.68	0.91	0.25
	m = M2 t = M2CPE						
$d_H(\Sigma_m, \Sigma_{e,t})$	0.40	0.68	0.91	0.51	0.83	0.84	0.68

Table 3. Correlation matrices for experts with smallest d_H per model.

$$\Sigma_{M1} = \begin{array}{cccccc} 1 & 0 & 0 & 0 & 0 & 0.49 \\ & 1 & 0 & 0.58 & 0 & 0.21 \\ & & 1 & 0 & 0.59 & 0.10 \\ & & & 1 & 0 & 0.31 \\ & & & & 1 & 0.19 \\ & & & & & 1 \end{array}$$

$$\Sigma_{G,M1RRC} = \begin{array}{cccccc} 1 & 0 & 0 & 0 & 0 & 0.41 \\ & 1 & 0 & 0.48 & 0 & 0.12 \\ & & 1 & 0 & 0.45 & 0.20 \\ & & & 1 & 0 & 0.33 \\ & & & & 1 & 0.12 \\ & & & & & 1 \end{array}$$

$$\Sigma_{M,M1RRC} = \begin{array}{cccccc} 1 & 0 & 0 & 0 & 0 & 0.51 \\ & 1 & 0 & 0.76 & 0 & 0.21 \\ & & 1 & 0 & 0.76 & 0.10 \\ & & & 1 & 0 & 0.33 \\ & & & & 1 & 0.32 \\ & & & & & 1 \end{array}$$

$$\Sigma_{M2} = \begin{array}{cccccc} 1 & 0 & 0 & 0 & 0 & 0.10 \\ & 1 & 0 & -0.57 & 0 & 0.58 \\ & & 1 & 0 & 0.90 & 0.30 \\ & & & 1 & 0 & 0.10 \\ & & & & 1 & 0.34 \\ & & & & & 1 \end{array}$$

$$\Sigma_{D,M2CPE} = \begin{array}{cccccc} 1 & 0 & 0 & 0 & 0 & 0.31 \\ & 1 & 0 & -0.51 & 0 & 0.33 \\ & & 1 & 0 & 0.91 & 0.31 \\ & & & 1 & 0 & -0.11 \\ & & & & 1 & 0.23 \\ & & & & & 1 \end{array}$$

$$\Sigma_{B,M2RRC} = \begin{array}{cccccc} 1 & 0 & 0 & 0 & 0 & 0.51 \\ & 1 & 0 & 0 & 0 & 0.41 \\ & & 1 & 0 & 0.81 & 0.21 \\ & & & 1 & 0 & 0.26 \\ & & & & 1 & 0.36 \\ & & & & & 1 \end{array}$$

estimated from a sample of size M from the normal copula with correlation matrix $\Sigma_{m,sample}$, and test if a particular value (calculated per expert, per model) is below agiven percentile (significance level) of this distribution. The empirical distribution of interest is obtained by a bootstrapping procedure. By such a procedure the correlation matrix $\Sigma_{G,M1RRC}$ is found significant at a 0.05 level of significance with sample size up to 300.

4 COMBINATION OF EXPERTS DEPENDENCE ESTIMATES

In Table 1 we showed how combining experts with regard to their assessments of the one dimensional uncertainty distributions, taking into account both calibration and information scores, may result in combined estimates for the distributions that perform better than any one of the experts. We now propose an approach of combining experts' dependence assessment in a similar fashion. First, we define the calibration of an experts' estimate of the dependence structure (correlation matrix) via the Heillinger distance: considering model m, actual correlation matrix Σ_m corresponding to this model and an expert's estimation Σ_e of the correlation matrix we define the *d-calibration* score as: $1 - d_H(\Sigma_m, \Sigma_e)$.

The d-calibration takes values between 0 and 1, with a high score implying that the expert's correlation matrix is statistically close to the actual correlation matrix. In order to measure this d-calibration in an expert judgment session we need to include variables with a known dependence structure, in addition to some unknown dependence structure we would like to assess as a main goal of the session.

Now we can discuss combining dependence structures. Observe that the set of all $n \times n$ correlation matrices is a convex subset of the set of all $n \times n$ matrices, meaning that the normalized weighted sum of correlation matrices will yield a correlation matrix[2].

As in the case of estimating one dimensional uncertainty distributions mentioned earlier, experts can be weighted equally or according to their (relative) performance in dependence assessments, as indicated by their d-calibration score.

If the d-calibration score is below a certain cut-off level, the expert is un-weighted and thus not used in the combined dependence estimation. The cut-off could be chosen by the analyst or determined by optimizing performance of the combined virtual expert. Here we choose the latter.

Table 4. d-calibration scores of combined expert judgment of dependence.

	m = M1 t = M1CPE	m = M2 t = M2RRC
Equal	0.74	0.37
Global	0.76	0.52
	m = M1 t = M1RRC	m = M2 t = M2CPE
Equal	0.66	0.37
Global	0.95	0.60

This yields the results as displayed in Table 4. 'Equal' denotes equal weighting of the experts estimates, while 'Global' denotes the weighting according to d-calibration score with the optimal cut-off.

In order to compare these d-calibration scores to Table 2, note that we need to compute 1 minus the values in Table 2 to obtain the d-calibration scores of individual experts.

We can make several observations: Only for M1CPE does the equal weighting come close to the global weighting in terms of d-calibration. In the other three settings the global weighting d-calibration score is significantly higher than the equal weighting d-calibration score. For M1CPE, the best individual expert's d-calibration score equals 0.65, which means that the equal weighting gives a better score than individual experts. In all other settings there are individual experts that outperform the equal weighting. For the global weighting in M1CPE, the estimates of the experts B, I, K and N are combined, while in M1RRC, the estimates of experts G and M are combined. Especially in the latter case, a significant improvement is made in d-calibration when considering individual experts versus global weighting. From 0.87 (expert G) to 0.95 (weighted combination of G and M). The correlation matrix corresponding to the global weight solution is presented in equation 3. This may be compared with individual expert estimates in Table 3. For both M2RRC and M2CPE, the global weighting gives full weight to the best expert, expert B and D respectively.

$$\Sigma_{Global,M1RRC} = \begin{matrix} 1 & 0 & 0 & 0 & 0 & 0.46 \\ & 1 & 0 & 0.61 & 0 & 0.16 \\ & & 1 & 0 & 0.60 & 0.16 \\ & & & 1 & 0 & 0.33 \\ & & & & 1 & 0.21 \\ & & & & & 1 \end{matrix} \qquad (3)$$

While expert B is among the best d-calibrated expert (in his group) for both the CPE and RRC elicitation techniques, this is not the case for

[2] Weighted combinations of correlation matrices might not, in general, preserve the conditional independence embedded in the graph.

any of the other experts. Expert L has the worst d-calibration score for both CPE and RRC.

Further exploration of these and other methods for combining experts estimates of dependence structures is necessary to find out which has the highest potential to lead to best dependence estimates.

5 DISCUSSION AND FINAL COMMENTS

We have presented an exercise aiming at answering the question of whether estimates of conditional rank correlations may be elicited more accurately from ratios of rank correlation or conditional probabilities of exceedence. This question is meaningful if experts are able to elicit the required quantities to a certain accuracy. In this particular exercise at least the estimates from expert G (M1RRC) would confer the required level of accuracy. On average experts' assessments for M1RRC performed significantly better that the other three groups. This tendency however is not preserved for individual models. No significant difference was observed for individual models across the elicitation techniques explored.

We have suggested some first steps in tackling the issue of combining experts' dependence assessment. Some further comments are in line. Comparing tables 1 and 2 we may see that good calibration (at least in the sense of the classical model) does not warranty that experts' would provide best dependence estimates. From group 1 expert A would dominate the linear pool. However her values for d_H are in the higher side (0.54 for M1RRC and 0.87 M2RRC). The situation for group 2 is more acute since expert F would be the recommended "combination" based on performance measures for one dimensional uncertainty distributions. Her d_H values would be 0.88 for M1RRC and 0.93 for M2RRC. In an uncertainty analysis, when the dependence between quantities of interest is relevant, it would be desirable to provide the best advise possible also with respect to dependence. The suggested approach via the Heillinger distance presents a first step towards this goal. As observed, the global weight combination of dependence shown in equation 3 would also confer the required level of accuracy. The d-calibration philosophy might be combined with the classical calibration methods for one dimensional uncertainty distributions. This is the subject of authors' current research.

ACKNOWLEDGEMENTS

This research was founded by the Enabling Technology Program of TNO under project Graphical Methods for Systems Risk and Reliability.

REFERENCES

Abou Moustafa, K.T., F. De La Torre, & F. P. Ferrie (2010). *Designing a Metric for the Difference between Gaussian Densities*, Volume 83. Berlin: Springer.

Ale, B., L. Bellamy, R.d. Boom, J. Cooper, R. Cooke, L. Goossens, A. Hale, D. Kurowicka, O. Morales, A. Roelen, & J. Spouge (2007). Further developments of a causal model for air transport safety (cats); building the mathematical heart. In *ESREL.*, pp. 1431–1439.

Clemen, G. & et al. (1999). Correlations and copulas for decision and risk analysis. *Management Science 45*, 208–224.

Clemen, G. & et al. (2000, August). Assesing dependencies: Some experimental results. *Management Science 2000 Informs 46*(8), 1100–1115.

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

Cooke, R. & L. Goossens (1999, July). Procedures guide for structured expert judgment. Technical Report EUR18820, European Comission: Nuclear Science and Technology, Brussels-Luxemburg.

Duncan, D.B. (1955). Multiple range and multiple f tests. *Biometrics 11*(1), 1–42.

Feller, W. (1948). On the Kolmogorov-Smirnov limit theorems for empirical distributions. *Ann. Math. Statist. 19*(2), 177–189.

Feller, W. (1950). Errata: On the Kolmogorov-Smirnov limit theorems for empirical distributions. *Ann. Math. Statist. 21*(2), 301–302.

Fisher, A. & F. Yates (1974). Statistical tables for biological, agricultural, and medical research (2nd ed.). *Edinburgh: Oliver and Boyd, Ltd.*, Table VII.

French, S. (2011). Aggregating expert judgments. *Real Academia de Ciencias Exactas Fis´ icas y Naturales 105*(1), 181–206.

Hanea, A. & W. Harrington (2009). Ordinal data mining for fine particles with non parametric continuous Bayesian belief nets. In *MMR 2009 Mathematical Methods In Reliability: Theory, Methods, Applications.*, Moscou, pp. 167–171.

Hanea, A., D. Kurowicka, & R. Cooke (2006). Hybrid method for quantifying and analyzing Bayesian belief nets. *Quality and reliability Engineering International 22*, 709–729.

Hanea, A.M., D. Kurowicka, R.M. Cooke, & D.A. Ababei (2010, March). Mining and visualising ordinal data with non-parametric continuous bbns. *Comput. Stat. Data Anal. 54*(3), 668–687.

Hanea, D. & B. Ale (2009). Risk of human fatality in building fires: A decision tool using Bayesian networks. *Fire Safety Journal 44*(5), 704–710.

Morales, O., D. Kurowicka, & A. Roelen (2008). Eliciting conditional and unconditional rank correlations from conditional probabilities. *Reliability Engineering & System Safety 93*(5), 699–710. Expert Judgement.

Morales-Nápoles, O., D.J. Delgado-Hernández, D. De-León-Escobedo, & J.C. Arteaga-Arcos (2013). A continuous Bayesian network for earth dams' risk assessment: methodology and quantification. *Structure and Infrastructure Engineering*, 1–15.

Ramachandran, K.M. & C.P. Tsokos (2009). *Mathematical Statistics with Applications*. ELSEVIER.