Available online at www.sciencedirect.com
ScienceDirect
Reliability Engineering and System Safety 93 (2008) 699–710
RELIABILITY ENGINEERING & SYSTEM SAFETY
www.elsevier.com/locate/ress

Eliciting conditional and unconditional rank correlations from conditional probabilities

O. Morales[a,*], D. Kurowicka[a], A. Roelen[b]

[a]Delft Institute of Applied Mathematics, Delft University of Technology, Mekelweg 4, 2628CD Delft, The Netherlands
[b]National Aerospace Laboratory, Amsterdam, The Netherlands

Available online 21 March 2007

Abstract

Causes of uncertainties may be interrelated and may introduce dependencies. Ignoring these dependencies may lead to large errors. A number of graphical models in probability theory such as dependence trees, vines and (continuous) Bayesian belief nets [Cooke RM. Markov and entropy properties of tree and vine-dependent variables. In: Proceedings of the ASA section on Bayesian statistical science, 1997; Kurowicka D, Cooke RM. Distribution-free continuous Bayesian belief nets. In: Proceedings of mathematical methods in reliability conference, 2004; Bedford TJ, Cooke RM. Vines—a new graphical model for dependent random variables. Ann Stat 2002; 30(4):1031–68; Kurowicka D, Cooke RM. Uncertainty analysis with high dimensional dependence modelling. New York: Wiley; 2006; Hanea AM, et al. Hybrid methods for quantifying and analyzing Bayesian belief nets. In: Proceedings of the 2005 ENBIS5 conference, 2005; Shachter RD, Kenley CR. Gaussian influence diagrams. Manage Sci 1998; 35(5) [15].] have been developed to capture dependencies between random variables. The input for these models are various marginal distributions and dependence information, usually in the form of conditional rank correlations. Often expert elicitation is required. This paper focuses on dependence representation, and dependence elicitation. The techniques presented are illustrated with an application from aviation safety.

1. Introduction

Graphical dependence models offer a compact and intuitive representation of high dimensional probability distributions. This property has made them attractive for applications in artificial intelligence, decision theory and uncertainty analysis. The models to be discussed in the present setting are non-parametric Bayesian belief nets (BBNs) (Section 3.2). The basic concepts and definitions required for the study of the graphical models will be briefly presented in Section 2.

These models contain nodes representing continuous random variables with invertible distribution functions and directed edges representing dependencies between the nodes, as unconditional and conditional rank correlations. Whenever possible, we retrieve these inputs from data. However, in many applications, information about the marginal distribution might be available but information about the joint distribution is not. We must then have recourse to expert judgment. In the worst case, not even a marginal distribution might be retrieved from data and expert judgment is used for this as well.

The issue of eliciting and combining experts' opinions as marginal probabilities has been widely discussed (see for example [1,2]) and will not be reviewed here. The focus of this paper is the elicitation of dependencies in the form of unconditional and conditional rank correlations; the combination of experts' dependence assessments is not an issue in this paper and hence will not be discussed. The elicitation of unconditional rank correlations has been discussed elsewhere [3–5]. Here the emphasis is placed on the probabilistic approach. Conditional probabilities of exceedance in Section 4.1 are extended in Section 4.2 to elicit conditional rank correlations.

In a project commissioned by the Dutch Ministry of Transport, Public Works and Water Management for aviation safety, a model for "*Missed Approach*" was recently developed and quantified with the methods

*Corresponding author.
E-mail addresses: o.moralesnapoles@ewi.tudelft.nl (O. Morales), d.kurowicka@ewi.tudelft.nl (D. Kurowicka).

0951-8320/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ress.2007.03.020

described here. This application model will be presented in Section 5. Section 6 presents conclusions and recommendations for future work.

2. Preliminary concepts and definitions

In this section we briefly present basic concepts and definitions used later on in the paper. The *product moment correlation* of random variables X and Y with finite expectations $E(X)$, $E(Y)$ and finite variances $var(X)$, $var(Y)$ is

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{var(X)var(Y)}}.$$

The *rank correlation* of random variables X, Y with cumulative distribution functions F_X and F_Y is

$$r_{X,Y} = \rho_{F_X(X),F_Y(Y)} = \frac{E(F_X(X)F_Y(Y)) - E(F_X(X))E(F_Y(Y))}{\sqrt{var(F_X(X))var(F_Y(Y))}}.$$

The rank correlation is the product moment correlation of the ranks of variables X and Y, and measures strength of monotonic relationship between variables. The conditional rank correlation of X and Y given Z is

$$r_{X,Y|Z} = r_{\widetilde{X},\widetilde{Y}},$$

where $(\widetilde{X}, \widetilde{Y})$ has the distribution of (X, Y) given $Z = z$.

The (conditional) rank correlation is the dependence measure of interest because of its close relationship with conditional copulas used in non-parametric BBNs. One disadvantage of this measure, however, is that it fails to capture non-monotonic dependencies.

Rank correlations may be realized by copulas, hence the importance of these functions in dependence modelling. A bivariate copula C is a distribution on the unit square $[0,1]^2$ with uniform marginal distributions on $[0,1]$. Random variables X and Y are joined by *copula* C if their joint distribution can be written as

$$F_{X,Y}(x,y) = C(F_X(x), F_Y(y)).$$

We can always find a unique copula that corresponds to any given continuous joint distribution. For example, if Φ_ρ is the bivariate standard normal cumulative distribution function with correlation ρ and Φ^{-1} the inverse of the univariate standard normal distribution function then

$$C_\rho(u,v) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v)); \quad u, v \in [0,1]$$

is called the *normal copula*. Notice that ρ is a parameter of the normal copula. The relationship between the correlation of the normal copula r (the rank correlation of the normal variables) and the parameter ρ (the product moment correlation of the normal variables) is known and given by the following formula [6]:

$$\rho = 2\sin\left(\frac{\pi}{6}r\right). \tag{2.1}$$

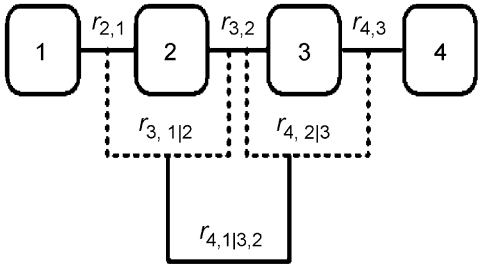

Fig. 1. D-Vine on four variables with (conditional) rank correlations assigned to its edges.

(Conditional) Copulas provide a natural way of constructing multivariate distributions with given marginals and given dependence structure. We will concentrate here on families of copula that have the zero independence property; that is, the property that correlation zero entails independence.[1]

The *partial correlation* can be computed recursively from correlations. The partial correlation of X_1 and X_2, with respect to $X_3, X_4, \ldots, X_n$ is [8]

$$\rho_{1,2;3,\ldots,n} = \frac{\rho_{1,2;4,\ldots,n} - \rho_{1,3;4,\ldots,n} \cdot \rho_{2,3;4,\ldots,n}}{((1-\rho^2_{1,3;4,\ldots,n}) \cdot (1-\rho^2_{2,3;4,\ldots,n}))^{1/2}}. \tag{2.2}$$

In general partial correlation is not equal to conditional correlation, however, for the joint normal distribution the partial and conditional correlations are equal. In the next section we begin our presentation with a brief description of non-parametric continuous BBNs.

3. Vines and non-parametric continuous BBNs

Vines and BBNs represent a joint distribution specified by marginal distributions and conditional bivariate dependence statements. One advantage of BBNs vs. vines is that the former preserve the intuitive representation of influence diagrams. This section describes vines and non-parametric BBNs.

3.1. Vines

A vine [6,9,10] is a graphical model for dependence modelling. The nodes of the vine represent random variables with invertible distribution function and the edges may be used to specify conditional bivariate dependencies. Formally, a *vine* on n variables is a nested set of trees where the edges of the jth tree become the nodes of the $(j+1)$th tree for $j = 1, \ldots, n-1$. A *regular* vine on n variables is a vine in which two edges in tree j are joined by an edge in tree $j+1$ only if these edges share a common node. A D-vine is a special case of a regular vine in which each node in T_1 has degree at most 2, hence each node in the first tree has at most two neighbors (see Fig. 1).

Each edge in the regular vine may be associated with a conditional rank correlation. In general these conditional

[1] For a review on copulas, see [7].

rank correlations may depend on the values of the conditioning nodes, but in the present implementation, all conditional rank correlations are constant. All assignments of rank correlations to edges of a vine are consistent and each one of these correlations may be realized by a copula. The vine enables the construction of a joint distribution from bivariate and conditional bivariate distributions.[2] If one chooses the normal copula to realize the (conditional) rank correlations assigned to the edges of a vine and the marginal distributions are standard normal, then we call such vine a *standard normal vine*. The standard normal vine gives us a very convenient way of specifying standard joint normal distribution by specifying $\binom{n}{2}$ algebraically independent numbers from $(-1,1)$. This is in contrast to the specification of a correlation matrix that must satisfy the constraint of positive definiteness [10].

3.1.1. Example

Let us consider a standard normal D-vine on three standard normal variables and assume that the following rank correlations were specified: $r_{2,1}$, $r_{3,2}$ and $r_{3,1|2}$. The correlation matrix of the joint normal distribution corresponding to this normal vine can be calculated as follows:

- Let $\rho_{2,1}$, $\rho_{3,2}$ and $\rho_{3,1;2}$ be the product moment correlations obtained by applying Eq. (2.1) to $r_{2,1}$, $r_{3,2}$ and $r_{3,1|2}$, respectively.
- Since for the normal distribution partial correlation is equal to conditional correlation $\rho_{3,1|2}=\rho_{3,1;2}$, then from Eq. (2.2) we can compute $\rho_{3,1}$ as

$$\rho_{3,1}=\rho_{3,1|2}\cdot((1-\rho_{2,1}^2)(1-\rho_{3,2}^2))^{1/2}+\rho_{2,1}\rho_{3,2}.$$

3.2. Non-parametric continuous BBNs

Non-parametric BBNs and their relationship to vines were presented in [11] and extended in [12]. A non-parametric continuous BBN is a directed acyclic graph whose nodes represent continuous univariate random variables and whose arcs are associated with parent–child (un)conditional rank correlations. For each variable i with parents $i_1,\ldots,i_{p(i)}$ associate the arc $i_{p(i)-k}\rightarrow i$ with the conditional rank correlation:

$$\begin{cases} r_{i,i_{p(i)}}, & k=0,\\ r_{i,i_{p(i)-k}|i_{p(i)},\ldots,i_{p(i)-k+1}}, & 1\leqslant k\leqslant p(i)-1. \end{cases} \tag{3.1}$$

The assignment is vacuous if $\{i_1,\ldots,i_{p(i)}\}=\emptyset$. These assignments together with a copula family indexed by correlation and with conditional independence statements embedded in the graph structure of a BBN are sufficient to construct a unique joint distribution. Moreover, the conditional rank correlations in 3.1 are algebraically independent, hence any number in $(-1,1)$ can be attached to the arcs of a non-parametric continuous BBN. In Fig. 2 one sees that variables 1–3 are independent and their dependence with the variable 4 is described in terms of three (conditional) rank correlations.

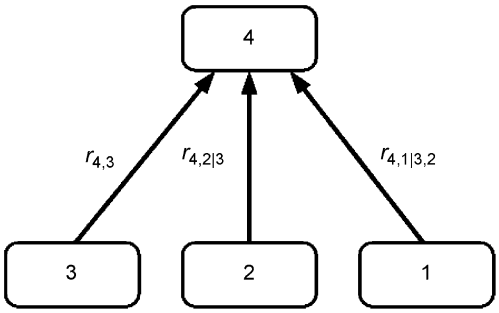

Fig. 2. A simple example of BBN on four variables.

One can use the copula-vine approach sketched in the appendix to represent the multidimensional joint distribution specified by a BBN [11,12]. D-vines become an important instrument as the sampling procedure for this BBN is based on the sampling procedure for the D-vine in Fig. 1 (the sampling procedure is presented in the appendix).

Any copula with an invertible conditional cumulative distribution function[3] may be used as long as the chosen copula possesses the zero independence property. In order to specify a joint distribution for a BBN (or for a vine), marginal distributions and conditional rank correlations must be specified. Next a procedure for eliciting conditional rank correlations for the BBN in Fig. 2 will be discussed, generalizations to other BBNs and vines are possible and follow the ideas presented in the next section.

4. Elicitation of conditional and unconditional rank correlations

Previous studies [3,4] indicate that information about dependencies such as rank correlations may be used as input parameters in risk and decision models. Ref. [5] identifies three groups of techniques for eliciting dependencies:

1. *Statistical approaches*: One option is to elicit from experts dependence statements with a predefined scale. This scale should later be translated to a rank correlation by the analyst. The informal translation of verbal qualifiers marks this method as a useful starting

[2]The reader may see in the appendix how to sample a joint distribution represented by the vine in Fig. 1. At this point the reader may also observe that a Markov-Dependence tree is a special case of a vine where all conditional rank correlations are set to zero. In other words, in a Markov-Dependence tree the random variables that are not joined by an edge in the tree are conditionally independent given variables on the path between them. The reader may see that vines relax the assumptions about conditional independence for Markov-Dependence trees to allow for conditional dependence.

[3]A copula with an analytic form for the conditional and inverse conditional cumulative distribution function accelerates the sampling procedure.

point to help expert think about dependence between variables. An other option is to let the expert directly assess a rank correlation, if indeed the expert is comfortable with the notion of rank correlation.

2. *Probabilistic approaches*: Experts are queried probability statements such as a joint probability, a conditional probability or a probability of concordance. By making assumptions about the joint distribution the assessments can later be translated to a rank correlation.
3. *Conditional quantile approach*: The expert is given the information that $Y = y_A$ that corresponds to the Ath quantile of Y, and is then queried about the expected quantile of X given $Y = y_A$. The relationship with $r_{X,Y}$ is determined from the non-parametric regression representation
$E(F_X(x_A)|Y = y_A) = r_{X,Y}(F_Y(y_A) - 0.5) + 0.5$, where F_X and F_Y are the cumulative distribution functions of X and Y, respectively. Ref. [3] suggests to have each expert make several conditional estimates and then use least squares to estimate $r_{X,Y}$.

The methods briefly discussed above have been used in the elicitation of unconditional rank correlations. In the non-parametric continuous BBN approach conditional rank correlations are also required. In practice conditional probabilities are more frequently used than conditional rank correlations. The conditional probability techniques used in previous studies can be naturally extended to elicit higher order dependence in the form of conditional rank correlations. However, there is as yet no empirical evidence on how well experts can estimate conditional rank correlations directly as opposed to the elicitation of conditional rank correlation through conditional probabilities. The next subsections elaborate in the conditional probability technique for eliciting (un)conditional rank correlations.

4.1. Conditional probabilities of exceedance and rank correlations

In this subsection the conditional probability method for estimating rank correlations is presented in more detail. The BBN in Fig. 2 will be used as example. To elicit the rank correlation $r_{4,3}$, we ask the expert the following:

1. *Suppose that the variable X_3 was observed above its qth quantile. What is the probability that also X_4 will be observed above its qth quantile?*

This question requires expert's estimate of $P_1 = P(F_{X_4}(X_4) > q | F_{X_3}(X_3) > q)$. Figs. 3 and 4 show the relationship between P_1 and the rank correlation $r_{4,3}$ for the normal and minimum information copulas, $q = \{0.5, 0.7\}$.

For the minimum information copula,[4] this relationship is determined using the simulation program UNICORN[5] at discrete points and smoothed using a locally weighted scatter plot smooth with least squares quadratic polynomial fitting.

For the normal copula we can use the relationship between the normal copula and the normal distribution from Section 2. To calculate the exceedance probability one can integrate numerically the bivariate normal density

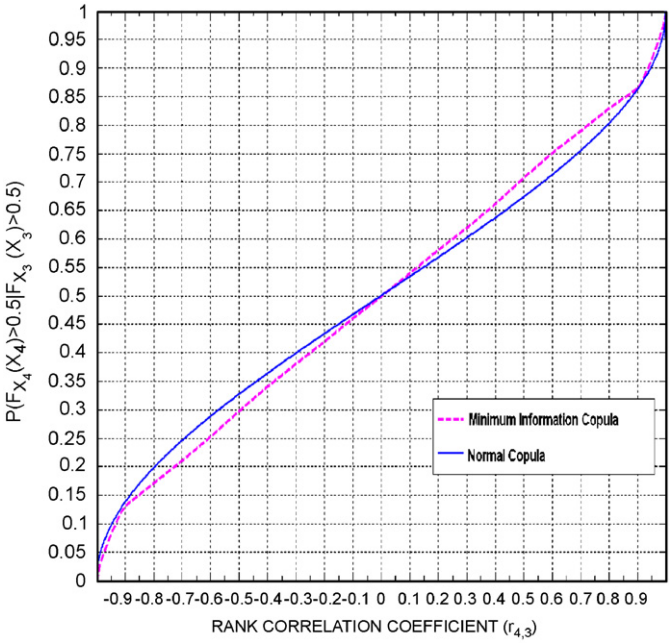

Fig. 3. Relationship between $P(F_{X_4}(X_4) \geqslant 0.5 | F_{X_3}(X_3) \geqslant 0.5)$ and $r_{4,3}$ for normal and minimum information copula.

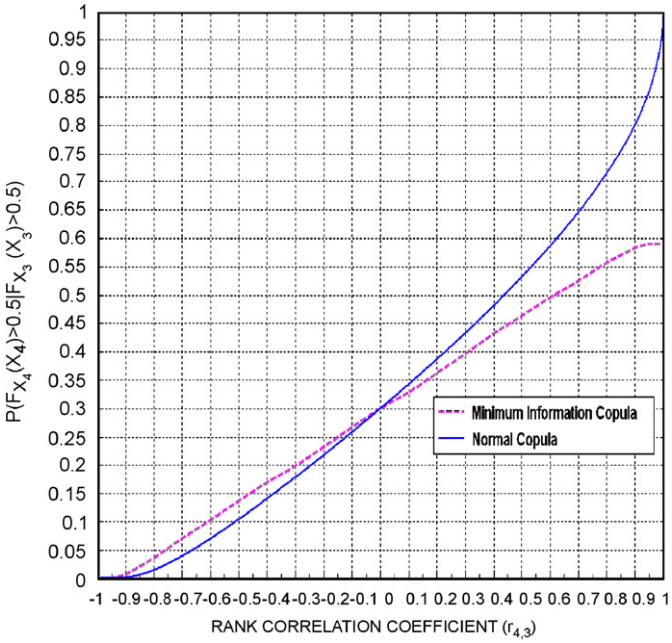

Fig. 4. Relationship between $P(F_{X_4}(X_4) \geqslant 0.7 | F_{X_3}(X_3) \geqslant 0.7)$ and $r_{4,3}$ for normal and minimum information copula.

[4]With respect to the independent copula [13].

[5]For details see [6].

$\phi(x_3, x_4, \rho_{4,3})$ over the region corresponding to the quantile's exceedance region $[\Phi^{-1}(q), \infty)^2$, where Φ^{-1} is the inverse standard normal cumulative distribution function (see formula (4.1)). The analyst then finds the ρ which satisfies the expert's conditional probability assessment and transforms this to the corresponding rank correlation using the inverse function of Eq. (2.1).

$$\frac{1}{1-q}\int_{\Phi^{-1}(q)}^{\infty}\int_{\Phi^{-1}(q)}^{\infty}\phi(x_3, x_4, \rho_{4,3})\,dx_3\,dx_4. \tag{4.1}$$

Because of the zero independence property, zero correlation entails that for any q, $P_1 = 1-q$. A conditional probability value in the interval $[0, 1-q)$ corresponds to negative correlation and positive correlation is attained when $P_1 > 1-q$. Choosing a value for q different than 0.5 makes the resulting rank correlation more dependent on the choice of copula, as is evident by comparing Figs. 3 and 4. In particular, $P(F_{X_4}(X_4) \geqslant 0.5 | F_{X_3}(X_3) \geqslant 0.5)$ may take any value in the interval $(0, 1)$ for both copulas which is not the case for $P(F_{X_4}(X_4) \geqslant 0.7 | F_{X_3}(X_3) \geqslant 0.7)$. The choice of the copula has a strong impact on the conditional probability when $q \neq 0.5$.

The conditional probability method has been extensively used together with structured expert judgment elicitation techniques in an uncertainty analysis conducted jointly by the European Union and the US Nuclear Regulatory Commission (see references in [5]). The approach followed was the one described in this subsection with $q = 0.5$ assuming the minimum information copula realizing the rank correlations in the joint distribution. Assessing higher order dependencies in the form of conditional rank correlations requires much more computational effort and we will then opt to use the normal copula and $q = 0.5$. In the next section a procedure to elicit conditional rank correlations will be presented continuing with the BBN in Fig. 2.

4.2. Conditional probabilities of exceedance and conditional rank correlations

In this subsection we extend the elicitation procedure from unconditional to conditional rank correlations. As already mentioned we will use the 50th percentile while eliciting exceedance probabilities and the normal copula to find the relationship between the probability of exceedance and the (conditional) rank correlation.

The elicitation of the rank correlation $r_{4,3}$ was described in Section 4.1. We will consider two situations when P_1 is equal to 0.25 and 0.75. We can read from Fig. 3 that these assessments correspond to rank correlations of -0.7 and 0.7, respectively. To assess the conditional correlation $r_{4,2|3}$ we will ask the expert the following question:

2. *Suppose that not only variable X_3 but also X_2 were observed above their medians. What is now your probability that also X_4 will be observed above its median value?*

An answer to this question is equivalent to an estimate of $P_2 = P(F_{X_4}(X_4) > 0.5 | F_{X_3}(X_3) > 0.5, F_{X_2}(X_2) > 0.5)$. The probability that the expert can provide in this situation will depend on the estimate given in question 1 (Section 4.1). The reader may see this by observing that if the expert regards variables X_2 and X_4 as independent given X_3, then the answer to question 2 is identical to the answer to question 1. If the expert regards variables X_3 and X_4 as completely positively (negatively) correlated then he/she would have answered $P_1 = 1$ ($P_1 = 0$) and question 2 would not have been necessary at all, as X_4 would be completely explained by X_3. Any answer for P_1 different than 0, 0.5 or 1 means that the expert believes that X_3 explains at least in part X_4 and hence X_2 can only explain part of the dependence that was not explained already by X_3.

In our example, to determine the possible values for P_2 and its relationship with the conditional correlation $r_{4,2|3}$ we consider a normal D-vine on variables X_4, X_3 and X_2. As mentioned earlier, the rank correlation $r_{4,3}$ has been already calculated using expert's assessment in question 1 (Section 4.1). In the particular case of the BBN in Fig. 2, variables X_3 and X_2 are independent, hence $r_{3,2}$ is equal to zero. Since all rank correlations specified on the BBN are algebraically independent, $r_{4,2|3}$ can take any value in $(-1, 1)$. The correlation matrix of the joint normal distribution corresponding to this normal vine can be found as in Example 3.1.1 and should have the form:

$$\Sigma_{4,3,2} = \begin{pmatrix} \rho_{4,4} & \rho_{4,3} & \rho_{4,2} \\ \rho_{4,3} & \rho_{3,3} & \rho_{3,2} \\ \rho_{4,2} & \rho_{3,2} & \rho_{2,2} \end{pmatrix} = \begin{pmatrix} 1 & \rho_{4,3} & \rho_{4,2} \\ \rho_{4,3} & 1 & 0 \\ \rho_{4,2} & 0 & 1 \end{pmatrix}. \tag{4.2}$$

We denote the density function of the normal distribution with the correlation matrix $\Sigma_{4,3,2}$ calculated from the normal vine specification as $\phi(x_4, x_3, x_2, \rho_{4,3}, \rho_{4,2|3})$. Hence, given the value for $r_{4,3}$ a relationship between P_2 and $r_{4,3|2}$ can be determined by transforming to $\rho_{4,3|2}$ using formula (2.1) and computing the triple integral (4.3):

$$\frac{1}{0.5 \cdot 0.5}\int_0^{\infty}\int_0^{\infty}\int_0^{\infty}\phi(x_4, x_3, x_2, \rho_{4,3}, \rho_{4,2|3})\,dx_4\,dx_3\,dx_2. \tag{4.3}$$

Fig. 5 shows the relationship between P_2 and $r_{4,2|3}$ when the expert's previous estimate for P_1 was 0.25 and 0.75. One can see that the probability of exceedance is constrained by the expert's previous estimate. If $P_1 = 0.25$ then P_2 is constrained to the interval (0,0.51); if $P_1 = 0.75$ then P_2 is constrained to the interval (0.49,1). In both cases conditional independence corresponds to the expert not modifying his/her previous estimate (0.25 and 0.75, respectively). Suppose the expert's assessments were $P_1 = 0.75$ ($r_{4,3} = 0.7$ from Fig. 3) and $P_2 = 0.65$ ($r_{4,2|3} = -0.4$ from Fig. 5), to assess the last conditional correlation $r_{4,1|3,2}$ we will ask the expert the following question:

3. *Suppose that not only variable X_3 but also X_2 and X_1 were observed above their medians. What is now your probability that also X_4 will be observed above its median value?*

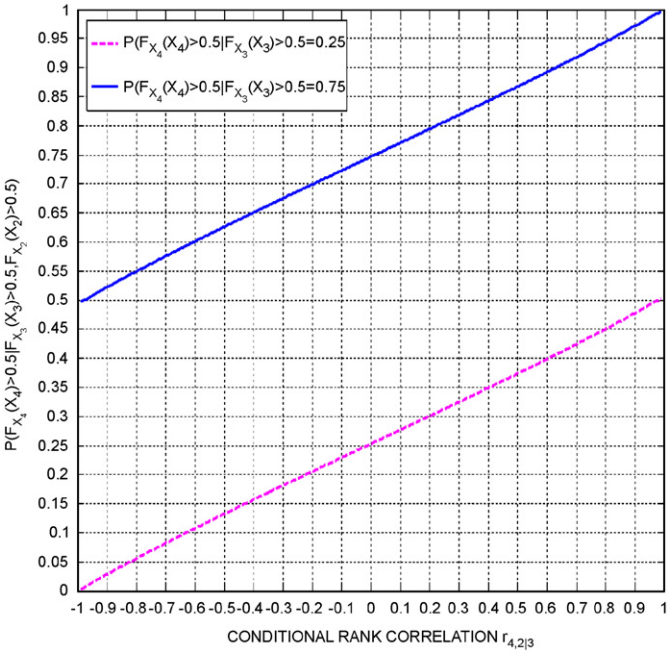

Fig. 5. Relationship between $P(F_{X_4}(X_4)\geqslant 0.5|F_{X_3}(X_3)\geqslant 0.5, F_{X_2}(X_2)\geqslant 0.5)$ and the rank correlation coefficient $r_{4,3|2}$ for the tri-variate standard normal distribution.

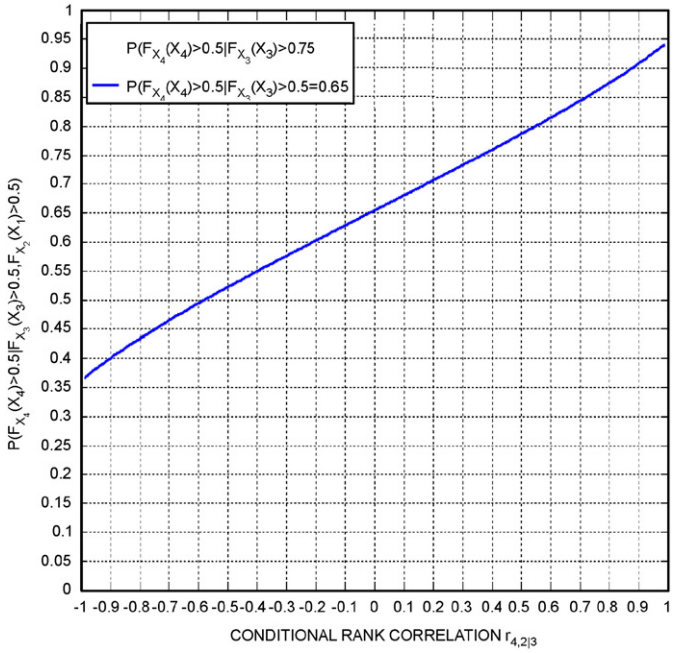

Fig. 6. Relationship between $P(F_{X_4}(X_4)\geqslant 0.5|F_{X_3}(X_3)\geqslant 0.5, F_{X_2}(X_2)\geqslant 0.5, F_{X_1}(X_1)\geqslant 0.5)$ and the rank correlation coefficient $r_{4,1|3,2}$ for the four variate standard normal distribution.

Question 3 is equivalent to estimating $P_3 = P(F_{X_4}(X_4)\geqslant 0.5|F_{X_3}(X_3)\geqslant 0.5, F_{X_2}(X_2)\geqslant 0.5, F_{X_1}(X_1)\geqslant 0.5)$. As before, the rank correlations $r_{4,3}$ and $r_{4,2}$ have been specified in questions 1 and 2, respectively. Again from Fig. 2 it is observed that $r_{3,1}$ and $r_{2,1}$ are both zero, hence the correlation matrix of the joint normal distribution corresponding to the D-vine on X_1, X_2, X_3 and X_4 should look as in Eq. (4.4). The density of this four variate standard normal distribution will be denoted as $\phi(x_4, x_3, x_2, x_1, \rho_{4,3}, \rho_{4,2}, \rho_{4,1|3,2})$:

$$\Sigma_{4,3,2,1} = \begin{pmatrix} \rho_{4,4} & \rho_{4,3} & \rho_{4,2} & \rho_{4,1} \\ \rho_{4,3} & \rho_{3,3} & \rho_{3,2} & \rho_{3,1} \\ \rho_{4,2} & \rho_{3,2} & \rho_{2,2} & \rho_{2,1} \\ \rho_{4,1} & \rho_{3,1} & \rho_{2,1} & \rho_{1,1} \end{pmatrix} = \begin{pmatrix} 1 & \rho_{4,3} & \rho_{4,2} & \rho_{4,1} \\ \rho_{4,3} & 1 & 0 & 0 \\ \rho_{4,2} & 0 & 1 & 0 \\ \rho_{4,1} & 0 & 0 & 1 \end{pmatrix}. \quad (4.4)$$

The relationship between P_3 and $r_{4,1|3,2}$ will be determined by transforming to the corresponding $\rho_{4,1|3,2}$ with formula (2.1) and computing the four-dimensional integral (4.5):

$$\frac{1}{0.5\cdot 0.5\cdot 0.5}\int_0^\infty\int_0^\infty\int_0^\infty\int_0^\infty \times\phi(x_4, x_3, x_2, x_1, \rho_{4,3}, \rho_{4,2}, \rho_{4,1|3,2})\,dx_4\,dx_3\,dx_2\,dx_1. \quad (4.5)$$

Fig. 6 shows the relationship between P_3 and $r_{4,1|3,2}$ for the case under study ($P_1 = 0.75$ and $P_2 = 0.65$). In this case, conditional independence corresponds to 0.65 (corresponding to the expert's conditional probability assessment for question 2). The expert's assessment is constrained to the interval (0.365,0.945). If the expert's assessment in question 3 were $P_3 = 0.55$ we would have $r_{4,1|3,2} = -0.4$. Observe that compared to the previous step, the upper and lower bounds for the conditional probability required from the expert are smaller. In general these bounds depend on the expert's previous assessments and must be computed on-line in a real elicitation to help experts avoid inconsistencies when providing the conditional probability that will be translated to a conditional rank correlation. In the next section an example of an elicitation recently conducted using the approach of conditional probabilities exposed here is presented.

5. The missed approach model

In recent years, the Federal Aviation Authority and the Dutch Ministry of Transport have used causal modelling techniques to investigate integrated safety in air traffic. For this purpose in [14] discrete Bayesian belief networks (BBN) were fully quantified for the cases of *Missed approach* and *Flight crew alertness*. However, two disadvantages with discrete BBNs were encountered:

- When variables were discretized into a number of values considered representative, the size of the conditional

probability tables exploded. As a result a drastic two-valued discretization (usually OK/Not OK) was forced.

- For many variables there was extensive data from the field. When using discrete BBN's, only the source nodes could be quantified with field data; other nodes have their marginal distributions determined by the conditional probability tables. Finding conditional probability tables that were compatible with the existing marginal information was a daunting, sometimes hopeless task.

Because of these problems, there was interest in finding a suitable alternative to discrete BBNs.

Here we will concentrate on the model for missed approach. A missed approach should be initiated when a situation arises that would make the continuation of the approach and landing unsafe. The purpose of a missed approach is to abort a landing in unsafe circumstances to allow the crew to carry out a new approach and landing under safer circumstances. According to [14] "the most common primal causal factor [of approach and landing accidents] was judged to be the omission of action/inapropriate action". Hence, the missed approach model tries to capture the idea of a *Failure to execute a missed approach when conditions are present*.

Fig. 7 presents the original discrete model for missed approach. All nodes in this model have two states. The top events are:

- *Condition for missed approach* that measures whether there is a condition during the approach or landing phase that requires a missed approach according to the operator's Aircraft Operating Manual, Basic Operating Manual, and/or (inter)national regulations. The states for this node are 'yes' or 'no'. This node is a deterministic node: an unfavorable condition of either one of its parents, alone or in combination will result in a condition for missed approach.
- *Missed approach execution* that describes whether the crew executes or does not execute a missed approach under certain circumstances (states 'yes' and 'no'). Compared to the *Condition for missed approach*, this node has an extra parent. The *In-flight crew alertness node* reflects the fact that the final decision to execute a missed approach is taken by the flight crew.

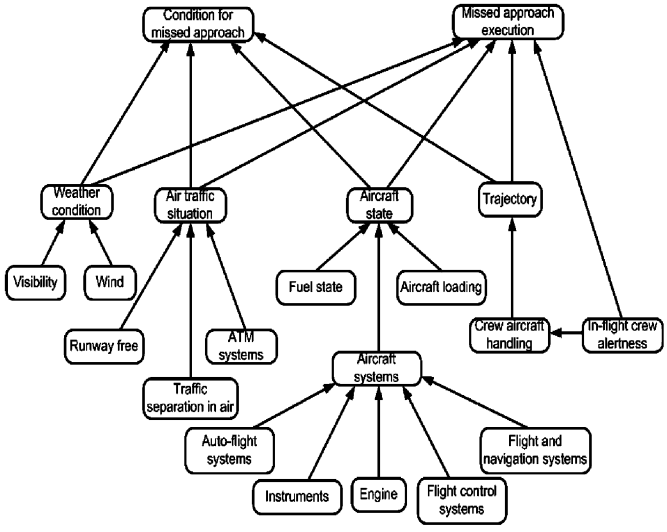

Fig. 7. Original BBN of the missed approach model.

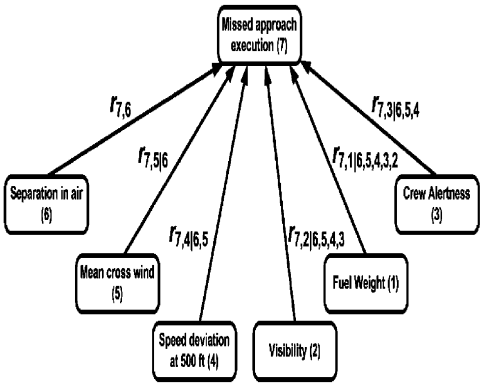

Fig. 8. Continuous version of the BBN for the missed approach model.

These two nodes are parents to the node *Failure to execute a missed approach when conditions are present* in further modelling which takes into account a possible accident situation. As stated before, some of the variables in Fig. 7 are more naturally modelled as continuous quantities for example: visibility, wind speed, fuel state, separation in air, etc. The variables are listed below according to their labelling in Fig. 8. The variables were quantified using field data.

1. *Fuel weight*: Measured in kilograms and is the remaining fuel at arrival based on data for 172 flights of a Boeing 737 at Schiphol airport.
2. *Visibility*: Measured in meters and is based on a sample of 27 million observations over Europe.
3. *Crew alertness*: Measured by the Stanford Sleepiness Scale in an increasing scale from 1 to 7, where 1 signifies "*feeling active and vital*; *wide awake*" and 7 stands for "*almost in reverie*; *sleep onset soon*; *struggle to remain awake*" the distribution used for this study comes from field studies by the Aviation Medicine Group of TNO Human Factors in 1295 flights.
4. *Speed deviation at* 500 ft: Deviation from bug speed[6] at 500 ft. The data comes from 13,753 approaches of a major European airline.
5. *Mean cross wind*: Usually expressed as a combination of speed (in knots) and direction (compass course) of the wind at any direction not favorable for the aircraft, the

[6]The bug speed is the target reference speed for the approach (calculated by the aircraft crew) plus allowance for conditions such as crosswind.

cross wind distribution comes from 380,000 takeoffs and landings conducted on three large European airports.

6. *Separation in air*: Longitudinal distance (in nautical miles) between the landing aircraft and the preceding aircraft in the approach path. The distribution was retrieved from a sample size of 2382 landings at Schiphol airport.
7. *Missed approach execution*: Number of missed approach executions per 100,000 flights at Schiphol airport. The expectation of this variable would be an estimate of the unconditional probability of executing a missed approach maneuver.

Information about the marginal distributions was available from different sources and the unconditional and conditional rank correlations were elicited with the procedure from Sections 4.1 and 4.2 from a single expert at the Dutch National Aerospace Laboratory (NLR) on December 20th, 2005 in a 2.5 h elicitation. The expert is a pilot for a major European airline and researcher at NLR, in total the expert answered seven questions.

One marginal distribution for *Missed Approach Execution per* 100,000 *Flights*, one unconditional rank correlation $r_{7,6}$ and the five conditional rank correlations from Fig. 8 were elicited. For the marginal distribution the expert was asked:

1. *Consider* 100,000 *thousand randomly chosen flights at Schiphol airport under the current conditions. On how many of these flights will a* missed approach *be executed? (To capture your uncertainty please provide the 5th, 25th, 50th, 75th and 95th percentiles of your uncertainty distribution.)*

A minimal informative distribution with respect to a log uniform background measure was fit with the data provided by the expert. Next, the dependence information was queried starting with the rank correlation $r_{7,6}$ as follows.[7]

2. *If* 50,000 *of the flights from the previous question were selected at random, then the number of flights that execute a missed approach should be approximately* $\frac{1}{2}$ *of your median estimate from previous question. Suppose that instead of selecting those* 50,000 *flights at random, you select those where* separation in air *is above its median value. What is your probability that, in this situation, the number of* missed approach *executions will be larger than* $\frac{1}{2}$ *of your 50th percentile estimate provided in the previous question?*

The assessment from question 2 is equivalent to an estimate of $P_1 = P(F_{X_7}(X_7) \geqslant 0.5 | F_{X_6}(X_6) \geqslant 0.5)$. The expert's assessment for this question was $P_1 = 0.15$ that from Fig. 3 corresponds to $r_{7,6} = -0.88$. The conditional rank correlation $r_{7,5|6}$ was elicited as follows.

3. *If* 50,000 *of the flights from question* 1 *were selected at random, then the number of flights that execute a missed approach should be approximately* $\frac{1}{2}$ *of your median estimate from question* 1. *Suppose that instead of selecting those* 50,000 *flights at random you select those where both* separation in air *and* mean cross wind *are both above their median values. What is your probability that, in this situation, the number of* missed approach *executions will be larger than* $\frac{1}{2}$ *of your 50th percentile estimate provided in question* 1*? (bearing in mind that your new assessment should be* $\in (0, 0.3)$*).*

The expert's assessment for question 3 is equivalent to an estimate of $P_2 = P(F_{X_7}(X_7) > 0.5 | F_{X_6}(X_6) > 0.5,\ F_{X_5}(X_5) > 0.5)$. The expert's answer to question 3 was $P_2 = 0.18$, and, with the methods described in 4.2 the corresponding value for $r_{7,5|6} = 0.20$ was found. The upper and lower bounds provided in question 3, i.e. the interval $(0, 0.3)$ were also computed on-line with the methods described in Section 4.2.

The rest of the conditional rank correlations were elicited in a similar way by sequentially adding information about the variables entering the conditioning set. The expert was provided with the upper and lower bounds for P_i $(i = 1, \ldots, 6)$ at each step in the elicitation only after he had provided his estimates to check for consistency. This way of assessing conditional rank correlations helped the expert understand the meaning of dependence and increased his "buy in" in the method. The results of the elicitation for the six arcs in the BBN for missed approach are summarized in Table 1.

Table 1
Results from expert's elicitation of conditional rank correlations

Conditional probability		Bounds for P_i[a]	Correlation	
P_1	0.15	$(0, 1)$	$r_{7,6}$	−0.88
P_2	0.18	$(0, 0.3)$	$r_{7,5\|6}$	0.20
P_3	0.20	$(0.01, 0.35)$	$r_{7,4\|6,5}$	0.12
P_4	0.24	$(0.02, 0.38)$	$r_{7,3\|6,5,4}$	0.23
P_5	0.22	$(0.04, 0.45)$	$r_{7,2\|6,5,4,3}$	−0.11
P_6	0.24	$(0.03, 0.40)$	$r_{7,1\|6,5,4,3,2}$	0.11

[a]Each P_i, $i = \{1, \ldots, 6\}$ sequentially adds variables to the model, for instance $P_1 = P(F_{X_7}(X_7) > 0.5 | F_{X_6}(X_6) > 0.5)$, $P_2 = P(F_{X_7}(X_7) > 0.5 | F_{X_6}(X_6) > 0.5, F_{X_5}(X_5) > 0.5)$, $P_3 = P(F_{X_7}(X_7) > 0.5 | F_{X_6}(X_6) > 0.5, F_{X_5}(X_5) > 0.5, F_{X_4}(X_4) > 0.5)$, and so on.

In [12] techniques to efficiently deal with the joint distribution when evidence becomes available (updating the BBN) are discussed. The two possibilities are:

- *The hybrid method*: To work with this method the information from Table 1 together with the marginal distributions for each variable were used to create a large sample file by means of the normal copula. A discrete version of the model can be built in order to take advantage of commercial software to perform fast updating each time a new policy is evaluated.
- *The normal copula-vine approach*: Since according to the methods described in Sections 3 and 4.2 all calculations

[7]The specification of the rank correlations required in the model presented in Fig. 8 is not unique (see Eq. (3.1)). For example instead of eliciting the (un)conditional rank correlations presented Fig. 8, one could also specify $\{r_{7,5}, r_{7,6|5}, \ldots\}$. In this case the order in which the variables entered the model was provided by the expert.

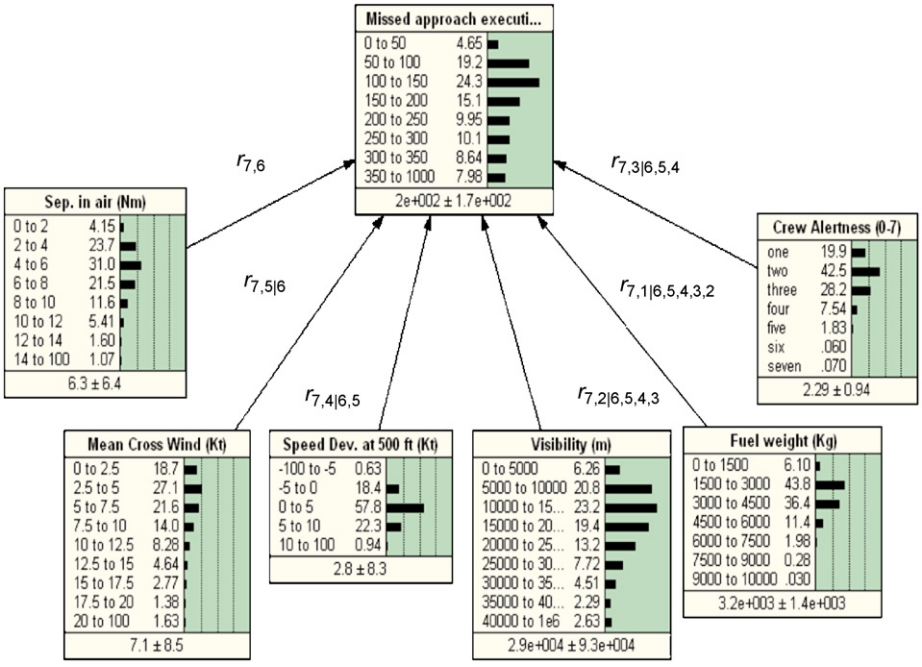

Fig. 9. Discretized BBN of the missed approach model with continuous quantities in Netica.

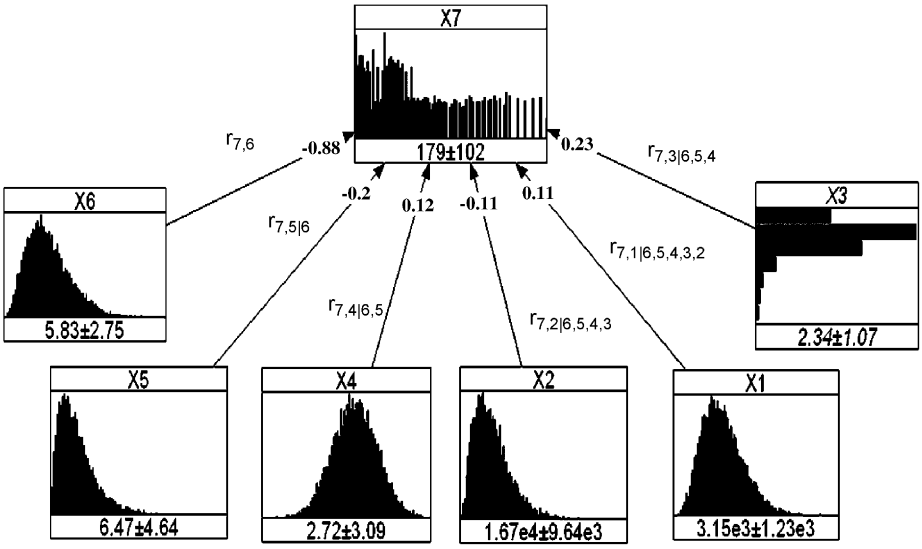

Fig. 10. Continuous BBN of the missed approach model with continuous quantities in UniNet.

are performed on a joint normal vine, the conditional distribution can be computed analytically.

To illustrate the hybrid method the professional software Netica© will be used. For the normal copula-vine approach the recently developed software application UniNet[8] will be used. Figs. 9 and 10 show the representation of the BBN for missed approach execution in Netica and UniNet, respectively. The rank correlations are included to stress the fact that both versions of the model introduced in Fig. 8 preserve the dependence structure elicited from the expert.

If instead of eliciting the six quantities in Table 1, the expert would have been asked to fill in the conditional probability table for X_7 missed approach execution per 100,000 flights with the discretization of its parent variables as in Fig. 9, then the expert would have had to provide over 1.2 million conditional probabilities that need to be consistent with the marginal distribution from Fig. 9 and still reflect the correct dependence information.

[8]UniNet has been developed for the Project commissioned by the Dutch Ministry of Transport. Currently UniNet supports both the hybrid method with the support of Netica and the analytical updating. The software is still under development.

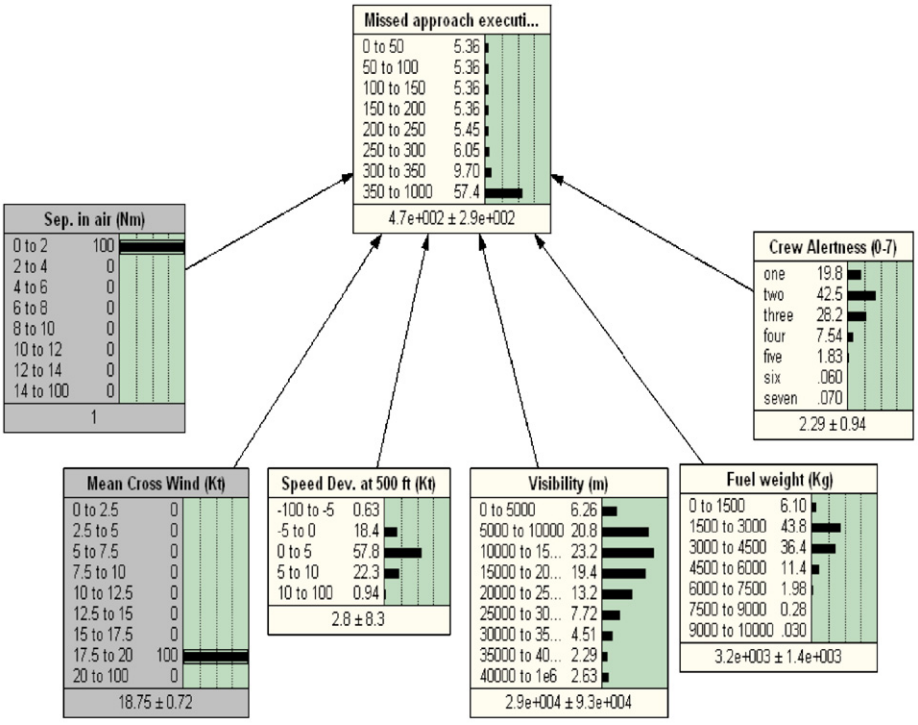

Fig. 11. Conditional distribution of missed approach executions per 100,000 flights given $X_6 \in (0,2)$ Nm and $X_5 \in (17.5,20)$ Kt.

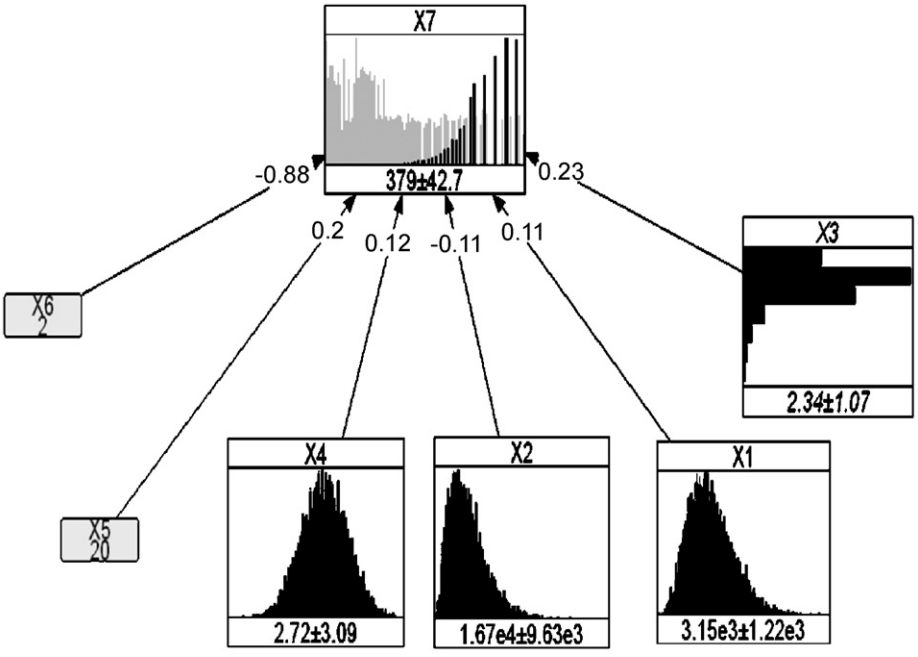

Fig. 12. Conditional distribution of missed approach executions per 100,000 flights given $X_6 = 2$ Nm and $X_5 = 20$ Kt.

Fig. 11 presents the distribution of missed approach executions per 100,000 flights given separation in air $\in (0,2)$ Nm and the mean cross wind $\in (17.5,20)$ Kt from Netica. The reader may compare this distribution with the unconditional distribution in Fig. 9. The unconditional mean is 200 missed approach executions per 100,000 flights (standard deviation of 170), while the mean of $(X_7|X_6 \in (0,2), X_5 \in (17.5,20))$ is 470 missed approach executions per 100,000 flights (standard deviation 290).

Fig. 12 presents the same conditional distribution as Fig. 11 computed analytically in UniNet. The unconditional distribution of X_7 is shown in grey behind the black histogram representing the conditional distribution of $X_7|X_6 = 2, X_5 = 20$. In this case the conditional mean is

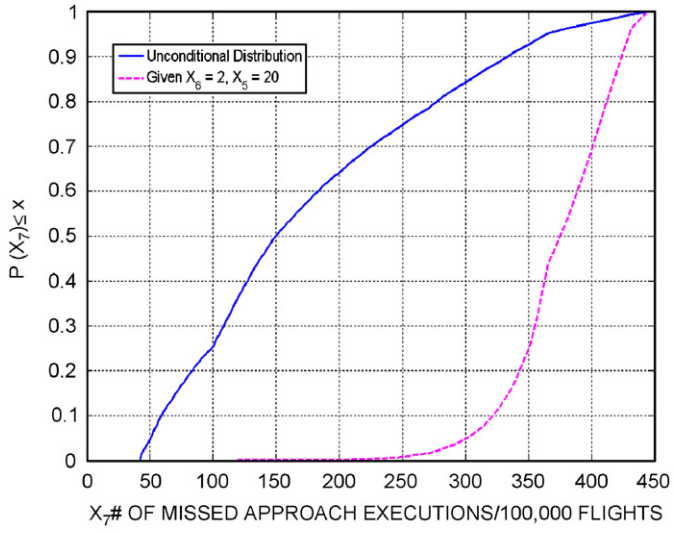

Fig. 13. Cumulative distribution function of X_7 and $X_7|X_6 = 2\,\text{Nm}$ and $X_5 = 20\,\text{Kt}$.

379 with standard deviation 47.7 missed approaches per 100,000 flights. While in Netica (Fig. 11) one can only condition in discretized states of each variable, UniNet allows for conditioning in point values. This is the usual way in which evidence becomes available in real situations.

The 500,000 samples from the joint distribution represented by Figs. 10 and 12 were obtained with UniNet. The cumulative distribution function of X_7 and $X_7|X_6 = 2, X_5 = 20$ were obtained and shown in Fig. 13. Observe that both Netica and UniNet show that $P(X_7 > 350) \approx 8\%$. In the conditional distribution computed with Netica this probability increases to $\approx 57\%$ while the analytical approach from UniNet shows that this value is as big as $\approx 75\%$. To finalize, in next section final comments and conclusions are presented.

6. Final comments and conclusions

This paper reviewed the elicitation of (conditional) rank correlations from conditional probabilities as inputs for continuous non-parametric BBNs. The conditional probability is a measure that has been elicited successfully in the past [5] and translated to rank correlations by assuming the minimum information copula realizing the joint distribution. This is the recommended way to elicit rank correlations from domain experts. The median value provides an intuitive choice for q and the normal copula presents computational advantages with respect to the minimum information copula. This motivates the choice of $P(F_X(X) > 0.5 | F_Y(Y) > 0.5)$ as the measure to be elicited and later translated to rank correlations.

The conditional probability technique for eliciting rank correlations is extended to allow the elicitation of conditional rank correlations. The nested constraints on successive conditional probability assessments give insight into the meaning of the dependence relations. To compute these constraints efficiently, so as to support the elicitation, the normal copula is a clear choice.

The application to missed approach demonstrates that it is possible to elicit unconditional and conditional rank correlations with intuitively meaningful conditional probabilities of exceedance. The results motivate the choice of the analytical updating (UniNet) vs. the hybrid method with Netica. Future research should be devoted to the issue of combining opinions of more than one expert and developing professional software tools for efficient implementation of the elicitation techniques herewith discussed.

Acknowledgment

The authors would like to thank Roger M. Cooke for his valuable ideas throughout the research on the topics of this paper.

Appendix

In this section a brief presentation of the techniques available to sample the joint distribution specified by the vine in Fig. 1 is presented. For a more detailed description of sampling techniques for regular vines the reader is referred to [6]. Assuming that random variables in the vine in Fig. 1 are uniform on (0,1) the density of the distribution satisfying the above dependence vine specification is [10]

$$\begin{aligned} f(x_1,x_2,x_3,x_4) = {} & c_{r_{1,2}}(x_1,x_2)c_{r_{2,3}}(x_2,x_3)c_{r_{3,4}}(x_3,x_4) \\ & \times c_{r_{4,5}}(x_4,x_5)c_{r_{1,3|2}}(F_{r_{1,2};x_2}(x_1), \\ & F_{r_{3,2};x_2}(x_3))c_{r_{2,4|3}}(F_{r_{2,3};x_3}(x_2), F_{r_{3,4};x_3}(x_4)) \\ & \times c_{r_{1,4|2,3}}(F_{r_{1,3|2};F_{r_{1,2};x_2}(x_1)} \\ & (F_{r_{2,3};x_2}(x_3)), F_{r_{2,4|3};F_{r_{2,3};x_3}(x_2)} \\ & (F_{r_{3,4};x_3}(x_4))), \end{aligned} \tag{A.1}$$

where $c_{r_{i,j}}$ denotes a copula density with correlation $r_{i,j}$ and $F_{r_{i,j};x_j}(x_i)$ denotes the conditional cumulative distribution function of X_i given X_j from the bivariate copula with rank correlation $r_{i,j}$.

The joint distribution specified by the (conditional) rank correlations on a vine with a given copula can be sampled on the fly. The algorithm involves sampling five independent uniform $(0,1)$ variables $U_1,\ldots,U_5$. We assume that the variables $X_1,\ldots,X_5$ in Fig. 1 are also uniform, then the sampling procedure can be stated as

$$x_1 = u_1,$$

$$x_2 = F^{-1}_{r_{1,2};x_1}(u_2),$$

$$x_3 = F^{-1}_{r_{2,3};x_2}(F^{-1}_{r_{1,3|2};F_{r_{1,2};x_2}(x_1)}(u_3)),$$

$$x_4 = F^{-1}_{r_{3,4};x_3}(F^{-1}_{r_{2,4|3};F_{r_{2,3};x_3}(x_2)}(F^{-1}_{r_{1,4|2,3};F_{r_{1,3|2};F_{r_{2,3};x_2}(x_3)}(F_{r_{1,2};x_2}(x_1))}(u_4))),$$

where $F_{r_{i,j};x_j}(x_i)$ is, as above, the conditional cumulative distribution function of X_i given X_j from the bivariate copula with correlation $r_{i,j}$ and F^{-1} denotes its inverse.

References

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

[2] Cooke RM, Goossens LHJ. TU Delft expert judgment database. Reliab Eng Syst Safety this issue, 2007; doi:10.1016/j.ress.2007.03.005.

[3] Clemenm GW, et al. Correlations and copulas for decision and risk analysis. Manage Sci 1999;45:208–24.

[4] Clemen GW, et al. Assesing dependencies: some experimental results. Manage Sci 2000;46(8):1100–15.

[5] Kraan BCP. Probabilistic inversion in uncertainty analysis and related topics. PhD thesis, Delft University of Technology, 2002.

[6] Kurowicka D, Cooke RM. Uncertainty analysis with high dimensional dependence modelling. New York: Wiley; 2006.

[7] Nelsen RB. An introduction to copulas. Lecture notes in statistics, vol. 139. Berlin: Springer; 1998.

[8] Yule G, Kendall M. An introduction to the theory of statistics, 14th ed. Belmont, CA: Charles Griffin & Co; 1965.

[9] Cooke RM. Markov and entropy properties of tree and vine-dependent variables. In: Proceedings of the ASA section on Bayesian statistical science, 1997.

[10] Bedford TJ, Cooke RM. Vines—a new graphical model for dependent random variables. Ann Stat 2002;30(4):1031–68.

[11] Kurowicka D, Cooke RM. Distribution-free continuous Bayesian belief nets. In: Proceedings of mathematical methods in reliability conference, 2004.

[12] Hanea AM, et al. Hybrid methods for quantifying and analyzing Bayesian belief nets. In: Proceedings of the 2005 ENBIS5 conference, 2005.

[13] Meeuwissen AMH, Bedford TJ. Minimally informative distributions with given rank correlation for use in uncertainty analysis. J Statist Comput Simul 1997; (57):143–75.

[14] Roelen ALC, et al. Causal modelling of air safety. demonstration model. Technical Report NLR-CR-2002-662, National Aerospace Laboratory, December 2002.

[15] Shachter RD, Kenley CR. Gaussian influence diagrams. Manage Sci 1998;35(5).