Recepción: 28 de enero de 2008
Aceptación: 4 de diciembre de 2008

*Organización de los Países Bajos para la investigación Científica Aplicada (TNO).
** Facultad de Ingeniería Eléctrica Matemáticas y Ciencias de la Computación, Universidad Tecnológica de Delft.
Correo electrónico: oswaldo.moralesnapoles@tno.nl; cooke@rff.org

Introducción al modelo clásico de juicio estructurado de expertos: breve recuento del pasado y una aplicación reciente

Oswaldo Morales Nápoles* y Roger M. Cooke**

Resumen. El uso de las opiniones de expertos en las ciencias y en la toma de decisiones ha sido una práctica común especialmente después del periodo de posguerra de la Segunda Guerra Mundial. El uso de estas opiniones con métodos que permitan su evaluación y validación con bases científicas es más reciente. Este artículo presenta una introducción al juicio estructurado de expertos. Este método ha sido diseñado con el propósito de alcanzar consenso racional. Dos conceptos claves del método (*calibración* e *información*) son discutidos. Una aplicación reciente que hace uso del método es brevemente presentada.

Palabras clave: análisis de incertidumbre, juicio estructurado de expertos, redes bayesianas, calibración, teoría de la información, teoría de decisiones.

Introduction to the Classical Model of Structured Expert Judgment: Brief Overview of the Past and a Recent Application

Abstract. The use of experts' opinions in science and decision making has been a common practice especially after the postwar period of the Second World War. The use of these opinions with methods that allow evaluation and validation on a scientific basis is more recent. This article presents an introduction to the classical model of structured expert judgment. This method has been designed to reach a rational consensus. A brief historical overview is presented. Two main concepts of the method (calibration and information) are discussed. A recent application that makes use of the method is briefly presented.

Key words: uncertainty analysis, structured expert judgment, bayesian networks, calibration, information theory, decision theory.

Introducción

La incertidumbre está implícita en las ciencias y sus aplicaciones. La ingeniería, las ciencias sociales, las ciencias de la salud y las ciencias aplicadas en general, hacen uso de modelos matemáticos para explicar los fenómenos que son de su respectiva competencia. Estos modelos hacen uso de parámetros que vienen, cuando es posible, de la experimentación o de la recolección de observaciones repetidas de eventos. Frecuentemente, las observaciones son demasiado costosas en tiempo, en términos económicos o, en general, en términos de los recursos disponibles. Cuando éste es el caso, los parámetros usados en los modelos son simplemente obtenidos de las opiniones de los expertos construyendo el

modelo. Al proceso de obtener parámetros o mediciones inciertos de la experiencia, conocimientos y opiniones de expertos se le llama juicio de expertos.[1]

El *juicio estructurado de expertos* es un intento por hacer la actividad de solicitar opiniones de expertos transparente y sujeto a metodologías con el objeto de tratar éstas como datos científicos en un proceso formal de toma de decisiones. El juicio estructurado de expertos se emplea en problemas en los que existe experiencia científica. Esto implica que hay teorías y mediciones relevantes al problema bajo estudio, pero en las que los parámetros de interés no pueden ser medidos.

Por ejemplo, la toxicidad de algunas sustancias en humanos es, en principio, cuantificable. Cuantificarla significaría, haciendo una muy rápida descripción, someter a grupos de personas a diferentes concentraciones de estas sustancias tóxicas y después analizar los órganos de relevancia para el estudio. Por obvias razones, este tipo de experimentación no es posible en el quehacer científico. En lugar de esto, lo que se hace en ocasiones es experimentar con animales o usar observaciones de grupos que accidentalmente han sido sometidos a sustancias tóxicas. Manipulaciones matemáticas posteriores aproximan el resultado que se hubiera observado en un experimento como el descrito anteriormente con brevedad.

Para ampliar el ejemplo, la siguiente pregunta puede formularse a un grupo de expertos: ¿Cuál es su estimación del valor verdadero, pero desconocido en la tasa de mortalidad no-accidental de *la temporalidad deseada: corto, mediano o largo plazo;* en *la población de interés*, resultante de una *magnitud, dirección y duración del cambio de interés*; en $PM_{2.5}$[2], resultante de una población ponderada por la concentración base de *nivel básico* a lo largo de *la región de interés*?[3] Indique los percentiles 5, 25, 50, 75 y 95 de su distribución de incertidumbre (Evans *et al.,* 2007: 6599).

¿Qué podemos decir acerca de la incertidumbre en las opiniones de estos expertos? ¿Cuál es la mejor manera de combinar los juicios de expertos individuales para llegar a un consenso de grupo? ¿Qué podemos decir acerca de los expertos como asesores de incertidumbre? Éstas son algunas de las preguntas que el modelo clásico para consultas estructuradas de opiniones de expertos trata de responder.

En este artículo se trata de introducir al lector al modelo clásico de consulta estructurada de expertos. Los conceptos matemáticos básicos empleados en el modelo serán revisados. El presente documento también hace un brevísimo recuento de las aplicaciones que se han hecho en el pasado usando el modelo clásico y presenta una aplicación que actualmente está siendo desarrollada con la ayuda de este modelo.

La motivación para escribir este artículo encuentra su origen en la falta de documentación en lengua castellana en la materia. Se espera que este documento sea de utilidad para estudiantes, investigadores y el público en general interesado en la cuantificación y evaluación de incertidumbre en trabajos científicos. Al lector interesado en profundizar en los temas tratados en este documento se le refiere a la bibliografía al final del documento.

1. Juicios estructurados de expertos

Los juicios de expertos adquieren particular importancia cuando la incertidumbre científica impacta los procesos de decisiones. Porque existe incertidumbre en la comunidad científica, los expertos no tendrán los mismos puntos de vista con respecto al valor de algún(os) parámetro(s). Solicitar la estimación de parámetros de expertos no es una idea nueva en las ciencias aplicadas. Debe notarse que el alcanzar un acuerdo entre los expertos mismos no es el fin del juicio estructurado de expertos.

Una metodología para el juicio estructurado de expertos debe aspirar de acuerdo con Cooke y Goznes (2008) a tres diferentes objetivos que son:

a) *Censo*. En el que simplemente se hace un examen de la distribución de los distintos puntos de vista en la comunidad científica. Diferentes maneras de dar pesos específicos a los expertos pueden ser usadas cuando se busca censar la opinión científica.

b) *Consenso político*. Que es un procedimiento mediante el cual se otorga peso específico a las opiniones expertas de acuerdo con los intereses o grupos que representan. En este caso es común dar pesos específicos iguales a cada actor involucrado.

c) *Consenso racional*. Éste se refiere a un proceso de decisiones de grupo. El grupo acuerda previamente y se compromete con una metodología que será utilizada para generar una representación de la incertidumbre del propósito para el cual el panel ha sido convocado. No es necesario que cada miembro del panel adopte el resultado de la metodología como su creencia personal. Más bien se trata de una forma de acuerdo en cuanto a la distribución que representará al grupo.

El modelo clásico de juicio estructurado de expertos tiene como objetivo alcanzar el consenso racional. Para que el método cumpla con el requisito de racionalidad, éste

1. En inglés a este proceso se le conoce como expert judgment.
2. Partículas con diámetro aerodinámico menor o igual a 2.5 micrómetros.
3. Un ejemplo concreto puede ser ¿cuál es su estimación del valor verdadero pero desconocido en la tasa de mortalidad no-accidental de (una semana) en (el área metropolitana de la ciudad de Toluca) resultante de un (incremento en un día de 10μg/m3) en PM2.5 resultante de una población ponderada por la concentración base de la mezcla ambiental de aerosoles a lo largo del (Valle de Toluca)?

debe cumplir con las condiciones necesarias impuestas por el método científico. De acuerdo con Cooke (1991) estas condiciones son:

a) Capacidad de escrutinio y confiabilidad. Todos los datos, incluyendo los nombres de los expertos, sus predicciones y todas las herramientas de procesamiento deben estar abiertas para la posible evaluación de la comunidad y, de ser necesario, los resultados deben ser reproducibles por un grupo de revisores competentes.

b) Control empírico. Las predicciones cuantitativas hechas por los expertos son sometidas a controles de calidad empíricos.

c) Neutralidad. La metodología propuesta para evaluar y combinar los juicios expertos, debe alentar a los expertos a declarar sus opiniones verdaderas y no debe sesgar los resultados.

d) Equidad. Los expertos no deben ser evaluados antes de procesar los resultados de sus predicciones.

Una vez que los actores involucrados se han comprometido con una metodología que cumpla con estos requisitos, la metodología es aplicada y los resultados obtenidos. Si alguno de los actores involucrados decide retirarse del grupo después de conocer los resultados, éste incurre en una carga de demostración. Es decir, debe demostrar que el proceso ha violado alguna de las condiciones anteriores. Cualquier miembro del grupo puede retirarse del consenso porque los resultados son hostiles a sus intereses. Esto, desde luego no es una decisión racional y no pone en peligro el consenso racional.

Debe señalarse que el juicio estructurado de expertos no compete a todos los ámbitos de interés del género humano. Por ejemplo, el uso del juicio estructurado de expertos para cuantificar la velocidad de la luz en el vacío o la fuerza de gravedad en los polos es irrelevante. Estas cantidades son físicamente cuantificables y se ha logrado medir a satisfacción del público en general. Igualmente, el juicio estructurado de expertos no puede usarse para determinar las predilecciones de dios, ya que por principio no hay expertos en la materia en el sentido operativo de la materia ni son éstas cuantificables en términos físicos.

2. Breve recuento histórico

Como se mencionaba anteriormente, el uso de expertos en las ciencias aplicadas no es nuevo; sin embargo, la noción de que sus ideas, pronósticos, especulaciones y en general su quehacer puede ser de importancia en un proceso estructurado de toma de decisiones el relativamente nuevo. Este fenómeno se puede fechar con el establecimiento de la corporación RAND[4] en 1946 (Cooke, 1991).

El periodo de posguerra después de la Segunda Guerra Mundial trajo consigo un intervalo de tiempo caracterizado por el auge en la investigación y el desarrollo. La corporación RAND fue creada en respuesta a este fenómeno como un proyecto conjunto entre la Fuerza Aérea estadounidense y Aeronaves Douglas en 1946 llamado "Proyecto RAND". Más tarde, esta organización que contaba originalmente con 300 hombres, se independizó para convertirse un uno de los primeros "tanques del pensamiento" usados principalmente por el gobierno estadounidense. Dos metodologías desarrolladas en RAND se convirtieron en el estándar para juicios estructurados de expertos: el análisis de escenarios y el método Delphi.

En el análisis de escenarios, el analista identifica lo que considera tendencias de largo plazo. Estas tendencias son extrapoladas al futuro tomando en cuenta conocimientos teóricos y empíricos que sean relevantes para la extrapolación. El resultado es denominado el escenario *libre de sorpresas*, el cual sirve como marco para definir *alternativas canónicas* o *variaciones canónicas*. Éstas son generadas al variar algunos parámetros del escenario libre de sorpresas. Como se apunta en Cooke (1991), una de las principales desventajas del análisis de escenarios es su falta de rigurosidad para hacer predicciones. Observar tendencias y hacer unas cuantas variaciones canónicas con respecto a la tendencia observada dista mucho de hacer predicciones y sobre todo de intentar validarlas.

El enfoque básico del modelo Delphi se describe a continuación: un grupo de analistas selecciona un conjunto de asuntos de interés y expertos en éstos. Un cuestionario preliminar es enviado a los expertos para obtener comentarios y después definir el cuestionario final. Este último es enviado a los expertos y después recolectado nuevamente para post procesamiento. El conjunto de respuestas es reenviado a cada uno de los participantes junto con las medianas de cada variable y *el rango interpercentil*. Éste era definido por el método como el rango que contiene 50% de las respuestas. El 50% restante, es decir, las respuestas excluidas, esta conformado por el 25% de las respuestas con los valores más pequeños y 25% de las respuestas con los valores más grandes del total de las respuestas. A los expertos se les pregunta si desean cambiar sus opiniones de acuerdo con los resultados observados y para aquellos cuyas respuestas permanecen fuera del rango interpercentil se les pide que den argumentos para sustentar su respuesta para esa variable en particular. Las respuestas son procesadas nuevamente; esta vez incluyendo los argumentos de aquellos expertos que están fuera del intervalo interpercentil y el proceso es iterado hasta alcanzar una dispersión aceptable.

4. Research and Development.

Muchas críticas se han vertido sobre este enfoque algunas de las cuales son: que los expertos no son tratados equitativamente ya que se "castiga" a los que están fuera del intervalo interpercentil con una mayor carga de trabajo al tener que justificar sus respuestas; existen argumentos para pensar que el método Delphi es terminado más debido al aburrimiento (por el largo proceso iterativo) que por alcanzar consenso. Finalmente, hay algunos argumentos que muestran que expertos y no expertos producen resultados comparables en un ejercicio Delphi.[5]

El modelo clásico fue introducido formalmente en Cooke (1991). Como se mencionó anteriormente, este modelo es un intento por hacer el juicio estructurado de expertos más formal y sujeto a reglas científicas de lo que era hasta el momento. Éste se ha desarrollado en la Universidad Tecnológica de Delft (TUD) desde hace aproximadamente 17 años (Cooke y Goznes, 2008). En el cuadro 1 se presenta un resumen de las aplicaciones del modelo clásico por sector. Actualmente se cuenta en la TUD con estudios en los que más de 67 000 distribuciones de probabilidad subjetivas han sido extraídas de expertos mediante el modelo clásico.

3. El modelo clásico para juicio estructurado de expertos

Un concepto clave en el modelo clásico es el de variables semilla o de calibración. Adicionalmente a las variables de interés, los expertos son evaluados con las anteriormente mencionadas. Las variables de calibración son aquellas cuyo verdadero valor es conocido para el/los analista(s), pero no para el experto al momento de la consulta. Las variables de calibración cumplen con tres objetivos:

a) Cuantificar el desempeño de los expertos como asesores de probabilidades subjetivas. Las medidas que se obtienen para medir el desempeño individual de los expertos son la calibración y la información.

b) Permitir la combinación optimizada (ecuaciones 6 y 9 en la sección 3.3.) con base en el desempeño de las distribuciones de probabilidad individuales de los expertos.

c) Evaluar y, siendo optimistas, validar la combinación de los juicios de expertos.

El modelo clásico para la combinación de juicios de expertos es un modelo de agregación lineal, es decir, mediante promedios ponderados basados en el desempeño de expertos medido por las preguntas de calibración. El nombre modelo clásico viene por una analogía entre la medición de calibración y las pruebas de hipótesis en la estadística clásica. A grandes rasgos, la calibración mide la probabilidad de que un conjunto de resultados experimentales correspondan en un sentido estadístico con las respuestas de los expertos en las preguntas de calibración. La información mide el grado en el que una distribución está concentrada con respecto a una medida previamente escogida.[6]

En el modelo clásico, los expertos son confrontados con variables que toman valores inciertos en un rango continuo y se les pregunta, típicamente, el percentil 5, el 50 y el 95 de su distribución subjetiva de incertidumbre. En algunos estudios los percentiles 25 y el 75 también han sido usados. Otros percentiles también pueden ser usados de acuerdo con la elección de los analistas. En lo que sigue, para facilitar la exposición, asumiremos que se trabaja con los percentiles 5, 50 y 95.

Cuadro 1. **Resumen de aplicaciones del modelo clásico por sector. (Cooke y Goznes, 2008: 658)**

Sector	No. de expertos[1]	No. de variables[2]	No. de estimaciones[3]
Nuclear applications	98	2 203	20 461
Chemical and gas industries	32	217	3 386
Chemical toxicity to humans	24	186	1 105
Groundwater, and water pollution	18	59	497
Moveable barriers and Dike ring failures	31	153	3 217
Volcano eruptions, and reliability of dams	231	673	29 079
Space shuttles, space debris, and aviation	51	161	1 149
Health items: bovine respiratory diseases, Campylobacter on chickens, and SARS	46	240	2 979
Banking issues: options, rents, and operational risks	24	119	4 324
Occupational issues: falls from ladders, and thermal physics of buildings	13	70	800
Rest group	19	56	762
En total	587	4 137	67 759

[1] Experto es cada uno de los participantes en un ejercicio del modelo clásico (ver sección 5).
[2] Variable es cada una de las preguntas (incluyendo preguntas de calibración) hechas a cada uno de los expertos participantes en un determinado estudio.
[3] Este es el número de estimaciones hechas en total en todos los estudios realizados en el sector de aplicación correspondiente. Por ejemplo en el sector nuclear cada experto ha cuantificado en promedio aproximadamente 20 461/98 ≈ 208.8 variables. Naturalmente el número de expertos y variables estimadas por experto varía de estudio a estudio en cada sector incluyendo el nuclear.

5. Ver por ejemplo Sackman, H. (1975) Delphi critique. Santa Monica, CA: RAND Corporation. Citado en Cooke, 1991.

6. Comúnmente se mide qué tan concentradas están las estimaciones de los expertos asumiendo que la distribución de los valores en cada intervalo interpercentil es uniforme o log-uniforme.

3.1. Calibración

Bajo el supuesto anterior cada experto dividirá el rango de cada variable en 4 intervalos interpercentiles. Los expertos serán denotados e_j para $j = 1,\ldots, E$. Cada intervalo interpercentil tiene una probabilidad conocida de acuerdo con la definición de percentil. El vector que contiene cada una de las cuatro probabilidades de los intervalos interpercentiles será denotado por $\boldsymbol{p}$. Sus elementos son $p_1 = 0.05$ (la probabilidad de observar un valor menor o igual al valor del percentil 5), $p_2 = 0.45$ (la probabilidad de observar un valor menor o igual al valor del percentil 50 y mayor que el percentil 5), etcétera.

$$p = (0.05, 0.45, 0.45, 0.05)$$

Si se tienen N variables de calibración cuyos valores verdaderos son denotados por $x_1,\ldots, x_N$ podemos formar una distribución muestral de los intervalos interpercentiles de cada experto. El vector que contiene cada una de las cuatro probabilidades empíricas de los intervalos interpercentiles de cada experto será denotado $s(e_j)$. Los elementos de este nuevo vector son: $s_1(e_j) = \text{card}(\{i|\, x_i \leq \text{percentil } 5\})/N$, $s_2(e_j) = \text{card}(\{i|\, \text{percentil } 5 < x_i \leq \text{percentil } 50\})/N$, etcétera.

$$s(e_j) = (s_1(e_j), s_2(e_j), s_3(e_j), s_4(e_j))$$

Si las observaciones son realmente extraídas independientemente de una distribución con percentiles, como los expresados por los expertos, entonces la expresión en (1) está asintóticamente distribuida como chi-cuadrada (χ^2_{k-1}) con grados de libertad igual al número de intervalos interpercentiles k menos uno (3 en este caso).

$$2 * N * I(s(e_j) \,|\, p) = 2 * N * \sum_{i=1}^{4} s_i(e_j) \ln(s_i(e_j)/p_i) \qquad (1)$$

El estadístico de prueba (1) es usualmente conocido como la razón de verosimilitud (Cooke y Goosens, 2008). $I(s(e_j) \,|\, p)$ es la información relativa de $s(e_j)$ con respecto a p. El método le otorga una puntuación al experto e_j como la verosimilitud estadística de la hipótesis He_j: el intervalo interpercentil que contiene al valor verdadero de cada variable es extraído independientemente del vector de probabilidades p. La *puntuación* o *score* de *calibración*, denotado $PC(e_j)$ es el valor-p de la hipótesis anterior.

$$PC(e_j) = P\{2 * N * I(s(e_j) \,|\, p) \geq r\} \qquad (2)$$

En (2) r es el valor de (1) basado en las observaciones $x_1,\ldots, x_N$. Esta puntuación es la probabilidad, bajo la hipótesis He_j de observar una desviación al menos tan grande como r en N muestras si He_j es verdadera. La puntuación de calibración es absoluta y puede ser comparada entre diferentes estudios teniendo cuidado de ajustar por el mismo número efectivo de preguntas de calibración.

Es importante enfatizar que aunque se usa el lenguaje estadístico de prueba de hipótesis, el método no rechaza expertos como hipótesis; sin embargo, sí utiliza este lenguaje para medir el grado en el que los datos soportan la hipótesis de que las probabilidades obtenidas de cada experto son precisas. Observe que (2) alcanza su valor máximo cuando $s(e_j) = p$. Valores de $PC(e_j)$ cercanos a cero indican que es poco probable que las probabilidades obtenidas del experto j sean correctas. Valores altos de $PC(e_j)$ cercanos a 1 pero mayores a 0.05 por ejemplo indicarían que las probabilidades de cada experto son respaldadas estadísticamente por el conjunto de variables de calibración. Valores de $PC(e_j)$ del orden de 0.001 fallarían en conferir los niveles de confianza requeridos por el estudio.

3.2. Información

La segunda medida de desempeño de los expertos es la *información*. Intuitivamente se puede decir que ésta mide el grado en el que una distribución está concentrada. La información no puede ser medida absolutamente, sino con respecto a una medida de fondo, es decir, se mide si la distribución de los expertos está concentrada o extendida con respecto a otra distribución. Comúnmente las distribuciones uniforme y log-uniforme son utilizadas como medidas de fondo.

Para medir la información se requiere asociar una función de densidad con cada percentil pronosticado por cada experto. Para hacer esto el modelo clásico utiliza la densidad única que cumple con los percentiles de los expertos y que es mínimamente informativa con respecto a la medida de fondo. Las medidas de fondo uniforme y log-uniforme requieren un rango intrínseco en el que dichas medidas están concentradas. Para cada variable, este rango en el modelo clásico se elige como el intervalo más pequeño que contiene todos los percentiles pronosticados por los expertos y la observación en el caso de las variables de calibración.

La *puntuación o score de información promedio* para el experto j en las predicciones de variables inciertas 1, …, N se muestra en (3) y es denotado como PI(e_j)

$$\text{PI}(e_j) = \frac{1}{N} \sum_{i=1}^{N} I(f_{e_j,i} \,|\, g_i) \qquad (3)$$

En (3) $f_{e_j,i}$ es la densidad del experto *j* para la variable *i*, mientras que g_i es la densidad de fondo de la misma variable *i*. La información no depende de las observaciones en las preguntas de calibración así es que los expertos pueden darse puntuaciones grandes eligiendo percentiles muy cercanos entre ellos. La puntuación de información depende del rango intrínseco y de los pronósticos de otros expertos así es que no puede ser comparado a través de diferentes estudios. Para mayores detalles el lector es referido a Cooke (1991) y Cooke y Goznes (2008).

3.3. Tomador de decisiones

La combinación de los pronósticos individuales de expertos es llamada en el modelo clásico un *tomador de decisiones* y será denotado TD. Como se mencionaba anteriormente el modelo clásico implementa TD que son combinaciones lineales o promedios ponderados. Sobre las ventajas y desventajas de combinaciones lineales para distribuciones de probabilidad el lector es referido a Genest, C. y Zidek, J. V. (1986). El modelo clásico es esencialmente un método para derivar pesos específicos para el promedio ponderado. De acuerdo con el modelo clásico serán buenos asesores de incertidumbre aquellos expertos que estén altamente calibrados y que sean informativos. Los pesos específicos premiarán a los buenos asesores de incertidumbre y heredarán sus propiedades al TD.

El aspecto de premiar buenos asesores de incertidumbre a la hora de construir pesos específicos para el promedio ponderado es muy importante. Es deseable que los expertos no traten de engañar al sistema a través de sus respuestas para obtener un resultado por ellos deseado. Por ello es necesario imponer la restricción de una *puntuación estrictamente propia*. Brevemente, esto significa que un experto alcanza su peso específico máximo si y solo si sus predicciones son expresadas de conformidad con sus creencias verdaderas.

Los pesos específicos en la combinación lineal los denotaremos por *w*. El peso específico $w_\alpha(e_j)$ en (4) es una puntuación estrictamente propia[7]. La función indicadora es denotada por $1_\alpha(x)$ y $1_\alpha(x) = 0$ si $x < \alpha$ y $1_\alpha(x) = 1$ si $x \geq \alpha$.

$$w_\alpha(e_j) = 1_\alpha(PC(e_j)) \times PC(e_j) \times PI(e_j) \tag{4}$$

La restricción de puntuación estrictamente propia indica dar peso específico cero a los expertos por debajo del nivel de confianza α, sin embargo no dice cuál debe ser el valor de α. Por lo tanto, éste es elegido de manera que el producto de la puntuación de calibración e información del TD es máximo (6).

$$\text{TD}_\alpha(i) = \sum_{e_j=1}^{E} w_\alpha(e_j) f_{e_j,i} \Big/ \sum_{e_j=1}^{E} w_\alpha(e_j) \tag{5}$$

$$\text{PC}(\text{TD}_\alpha) \times \text{PI}(\text{TD}_\alpha) \tag{6}$$

Tal como se muestra en (5), se denota $\text{TD}_\alpha(i)$ a la combinación lineal de la variable i con pesos específicos proporcionales a (4). El TD de *pesos globales* (denotado TDPG) es $\text{TD}_\alpha^*(i)$ en donde α* maximiza (6). Este TD es llamado global porque la puntuación de información está basada en todas las variables de calibración pronosticadas por los expertos. El TDPG puede variarse para construir otro tipo de pesos específicos que puede ser usado alternativamente para cada variable de interés. Esto se logra usando una puntuación de información por cada variable en vez de la puntuación de información promedio mostrada en (3). En este caso para hacer la combinación lineal se usa el peso específico $w_\alpha(e_j, i)$ en (7).

$$w_\alpha(e_j, i) = 1_\alpha(PC(e_j)) \times PC(e_j) \times I(f_{e_j,i} | g_i) \tag{7}$$

Para cada α el tomador de decisiones por artículo TDA de la variable i se obtiene como en (8)

$$\text{TDA}_\alpha(i) = \sum_{e_j=1}^{E} w_\alpha(e_j,i) f_{e_j,i} \Big/ \sum_{e_j=1}^{E} w_\alpha(e_j,i) \tag{8}$$

$$\text{PC}(\text{TDA}_\alpha) \times \text{PI}(\text{TDA}_\alpha) \tag{9}$$

El TD de *pesos por artículo* (denotado TDPA) es $\text{TDA}_\alpha^*(i)$ en donde α* maximiza (9). Ambos pesos específicos por artículos y globales pueden ser llamados de manera concisa pesos específicos óptimos bajo una restricción de puntuación propiamente estricta. En ambos tipos de pesos específicos la calibración domina sobre la información y la información sirve para modular entre expertos más o menos igualmente calibrados.

El último tipo de TD a considerarse es el de pesos iguales. Este TD no toma en cuenta el desempeño de los expertos en las preguntas de *calibración* y les da el mismo peso específico como asesores de incertidumbre. El tomador de decisiones de pesos iguales se muestra en (10). Se puede observar que éste no depende de α.

7. Una puntuación *R* para una sola variable incierta que toma valores 1,…, *n* es una función que otorga estímulos $R(\boldsymbol{p}, i)$ para una predicción probabilística $\boldsymbol{p}$ cuando hay una realización *i*. El estímulo esperado para una probabilidad subjetiva $\boldsymbol{p}$ cuando un experto cree que la verdadera distribución es $\boldsymbol{q}$ es $EqR(\boldsymbol{p}|i) = \sum_1^n q_i R(p,i)$. La puntuación es propia si para todo $\boldsymbol{p}$ y $\boldsymbol{q}$ $EqR(\boldsymbol{p}|i)$ es maximizado cuando $\boldsymbol{p} = \boldsymbol{q}$.

$$\text{TDPI}(i) = \frac{1}{E} \sum_{e_j=1}^{E} f_{e_j,i} \qquad (10)$$

En el modelo clásico es deseable que los TD óptimos tengan un mejor desempeño que el TDPI y es también deseable que el TD propuesto no tenga un desempeño menor que el peor experto del panel. En la práctica, el TDPG es usado a menos que el TDPA tenga un desempeño marcadamente mejor.

Finalmente, se remarca que los nombres y credenciales de los expertos son parte de la documentación publicada de cada estudio de juicios estructurados de expertos; sin embargo, éstos no son asociados con sus pronósticos en la literatura abierta.

4. Una aplicación del presente

Actualmente, un modelo para la medición de riesgos en la industria de la aviación está siendo desarrollado por un consorcio para el ministerio de transporte y manejo de aguas holandés (*Rijkswaterstaat*). El consorcio está conformado por Det Norsk Veritas, White Queen y TUD; el modelo es llamado por sus siglas en inglés CATS[8] (Morales-Napoles *et al.,* 2007 y Ale *et al.,* 2007). El objetivo del modelo es cuantificar la probabilidad de accidentes en aeronaves comerciales con peso de despegue mayor o igual a 5 000 toneladas. El modelo inicialmente fue cuantificado con *árboles de fallos* (Vesely *et al.,* 1981). Adicionalmente, redes Bayesianas continuas-discretas no-paramétricas (Kurowicka *et al.,* 2005) y (Hanea *et al.,* 2006) fueron construidas y cuantificadas para modelar la probabilidad de errores humanos en la tripulación, controladores aéreos y personal de mantenimiento (Morales-Nápoles *et al.,* 2008). A las redes bayesianas continuas discretas no-paramétricas se les llamará RBCDNP.

El modelo genérico para el desempeño de la tripulación se muestra en la figura 1. Los tres modelos para desempeño humano incluido el mostrado en la figura 1 fueron parcialmente cuantificados haciendo uso del modelo clásico para juicio estructurado de expertos. No es el objetivo presentar al lector la teoría detrás de este tipo de modelos, sino simplemente dar una idea del tipo de aplicaciones que actualmente se apoyan del modelo clásico para el juicio estructurado de expertos. Para una mayor comprensión de este tipo de modelos se sugiere al lector consultar las fuentes originales.

En las RBCDNP los nodos representan distribuciones de probabilidad unidimensionales. Los arcos y la estructura gráfica en general representan dependencia entre los nodos. El grafo en general, representa una distribución de probabilidad multidimensional. Las distribuciones marginales (representadas en los nodos de la figura 1) se describen en el cuadro 2. Las correlaciones de rango y las correlaciones condicionales de rango son denotadas como $r_{X,\,Y}$ para la correlación de rango entre X y Y o $r_{X,\,Z|\,Y}$ para la correlación de rango de X y Z dado Y. Las correlaciones de rango condicionales de orden superior son representadas de manera similar. En la figura 1 la de orden mayor es $r_{14,13|10,\,12,\,8,\,9,\,11}$.

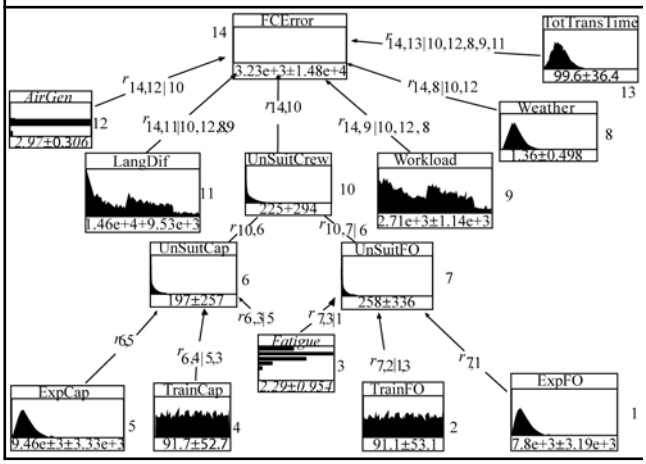
En la práctica, estas medidas de dependencia son difíciles de obtener de los datos ya que esto implicaría tener una muestra grande de la distribución de probabilidad conjunta completa. El recurso disponible es consultar la opinión de expertos. Aunque en menor grado, el juicio estructurado de expertos también ha sido usado para la consulta de dependencia (Clamen *et al.,* 1999; Clamen *et al.,* 2000; Kraan, 2002; Morales Nápoles *et al.,* 2007 y Morales Nápoles *et al.,* 2008).

En este caso, las preguntas que se hacen a los expertos son esencialmente: *¿Suponga que la variable X se encuentra por encima de su mediana, cuál es la probabilidad de que Y esté también por encima de su mediana*? La respuesta a esta pregunta de acuerdo con los métodos presentados en (Morales Nápoles *et al.,* 2007) y (Morales Nápoles *et al.,* 2008) puede traducirse en $r_{X,Y}$. Para preguntar correlaciones condicionales se puede preguntar:

a) ¿Suponga que no sólo la variable X se encuentra por encima de su mediana, sino también Z está por encima de su mediana, cuál es la probabilidad de que Y esté también por encima de su mediana? o

b) ¿Cuál es la razón entre la correlación de rango de X y Z y la correlación de rangos de X y Y?

Figura 1. **Red bayesiana para la medición del desempeño de tripulaciones de aeronaves.**



8. Causal Model for Air Transpors Safety que se puede traducir como Modelo Causal para la Seguridad en el Transporte Aéreo.

La respuesta a cualquiera de las preguntas anteriores permitiría calcular $r_{X, Z|Y}$. El método brevemente descrito hasta aquí para preguntar correlaciones de rango a expertos puede ser extendido para correlaciones condicionales de cualquier orden y ha sido usado exitosamente en la práctica (Morales Nápoles *et al.,* 2008).

En total 5 pilotos fueron consultados para la cuantificación del modelo presentado en la figura 1. En total 4 distribuciones marginales, 11 preguntas para obtener dependencia similares a las presentadas anteriormente[9] y 8 preguntas de calibración fueron realizadas. Un ejemplo de las preguntas usadas para obtener distribuciones marginales (ver el ejemplo en la sección 1) es:

> Considere la población de aeronaves con peso de despegue mayor a 5 700 kg, construidas por empresas occidentales efectuando vuelos actualmente a nivel mundial. Considere 10 000 vuelos escogidos aleatoriamente de la población total ¿En cuántos de estos vuelos el capitán y el primer oficial tendrán una lengua madre distinta? (indique los percentiles 5, 50 y 95 de su distribución de probabilidad).

Las preguntas de calibración son similares al ejemplo anterior en el sentido de que conciernen al campo de conocimiento del experto. Como se mencionó anteriormente para éstas, los analistas (y no los expertos) conocen el valor verdadero *a priori*. Un resumen de los resultados es presentado en el cuadro 3. Todos los cálculos fueron hechos en EXCALIBUR que es una herramienta de software desarrollada en el Departamento de Matemáticas aplicadas de TU Delft para el post procesamiento de resultados del modelo clásico.

El cuadro 3 presenta los resultados de evaluar a los expertos participantes en el estudio bajo el modelo clásico. En total 5 expertos y dos tomadores de decisiones se muestran en el cuadro 3.[10] La primera columna presente en el cuadro 3 muestra la clave de cada experto y la segunda columna su puntuación de calibración (ver ecuación 2). La razón de mayor a menor puntuación es aproximadamente 1.3E+4. Nótese que los expertos B y D tienen un valor-*p* por encima del 5%. Las puntuaciones de los expertos E y C son marginales y la del experto A es bastante baja. Valores de calibración del orden de 0.001 fallarían para conferir los niveles de confianza requeridos por el estudio. La puntuación de calibración para todas las variables (de interés y de calibración) se muestra en la columna tres. La puntuación de información para las

Cuadro 2. **Variables utilizadas en el modelo de desempeño de la tripulación de la figura 1.**

Nombre	Descripción	Recurso
FOExp	Número total de horas voladas desde la obtención de licencia. (Primer oficial)	Datos
FONotTraining	Número de días transcurridos desde el último entrenamiento recursivo (Primer oficial).	Datos
Fatigue	Stanford Sleepiness Scale. 1 significa "Sintiéndose vital y activo; ampliamente despierto" 7 significa "casi en ensueño; comienzo del sueño pronto; dificultad para mantenerse descuento".	Datos
CapNotTraining	Número de días transcurridos desde el último entrenamiento recursivo (Capitán).	Datos
CapExp	Número total de horas voladas desde la obtención de licencia. (Capitán)	Datos
CapUnSuit	Número de capitanes que no aprueban el examen de habilidades por cada 10,000 capitanes	JEE*
FOUnSuit	Número de primeros oficiales que no aprueban el examen de habilidades por cada 10 000 capitanes	JEE
Weather	Tasa de precipitación (mm/hr) traducida al radar de clima de la cabina.	Datos
DLanguage	Número de vuelos en el que el piloto al mando y el primer oficial tendrán diferente lengua madre por cada 100,000 vuelos.	JEE
CrewUnSuit	Número de tripulaciones (capitanes y/o primeros oficiales) que fallan su examen de competencia por cada 10 000.	JEE
AGeneration	Escala de 1 a 4 en donde 4 representa a la generación más reciente de aeronaves.	Datos
Workload	Número de situaciones que refieren a la tripulación al manual de operaciones anormales o de emergencia por cada 100,000 vuelos.	JEE
HError	Número de errores por cada 16 vuelos. Una distribución es ajustada a datos obtenidos por DNV acerca del número de errores por millar de vuelos. Un ejemplo de error puede ser: frenos no aplicados correctamente por despegue rechazado por fallo en sistema de la aeronave.	Datos Árboles Decisiones DNV

* Juicio estructurado de expertos.

Cuadro 3. **Resumen de resultados del modelo clásico de JEE en el modelo de desempeño de tripulación.**

Results of scoring experts and Relative Information to the DM
Bayesian Updates: no Weights: global DM Optimization: yes
Significance Level: 0.6638 Calibration Power: 1

Id	Calibr.	Mean relati total	Mean relati realizatioo	Numb real	UnNormalize weight	Normaliz.we without DM	Normaliz.we with DM
A	0.0265	0.7119	0.499	8	0.000	0	0.0
B	0.6638	0.95	0.574	8	0.381	1	0.5
C	0.0015	1.016	0.968	8	0.000	0	0.0
D	0.1850	1.317	1.029	8	0.000	0	0.0
E	5.115E-005	1.049	1.060	8	0.000	0	0.0
TDPG	0.6638	0.95	0.574	8	0.381	-	0.5
TDPI	0.2224	0.1046	0.099	8	0.2212	-	-

(c) 1989-2005 TU Delft

9. En total, de acuerdo con la figura 1, 13 correlaciones de rango son necesarias en el modelo. Sin embargo, $r_{10,6}$ y $r_{10,7|6}$ no fueron preguntadas a expertos, sino escogidas tal que $r_{10,6} = r_{10,7,}$ ambas son positivas y tan grandes como sea posible.

10. En este estudio, el TDPA y el TDPG son idénticos y por esta razón se discute únicamente el último.

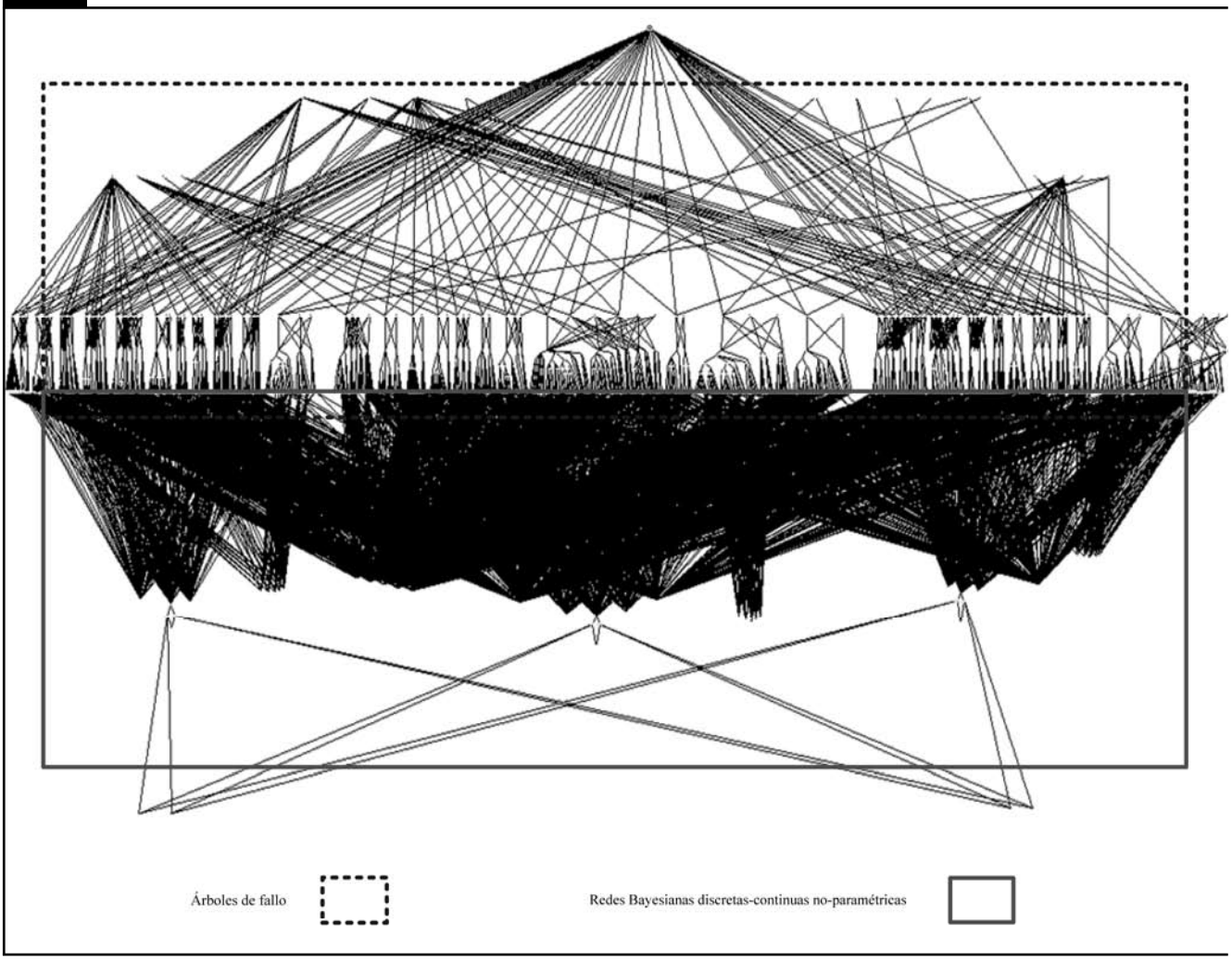
variables de calibración únicamente es mostrada en la cuarta columna del cuadro 3. Se observa que las puntuaciones de información son bastante similares entre expertos alrededor de un factor de 2. En este caso el experto mejor calibrado (B) también tiene una de las puntuaciones de información más bajas. Esto corresponde a un patrón recurrente en estudios anteriores.

Los pesos específicos para cada experto corresponden al producto de las columnas 2 y 4 (ecuaciones 4 y 7). Si estos pesos fueran normalizados y usados para formar promedios ponderados, los expertos A, D y B serían de influencia con 2.25, 32.49 y 64.98% respectivamente. El cuadro 3 muestra también que el TDPI está mejor calibrado que cada experto individualmente excepto el experto B. Sin embargo, su puntuación de información es pobre. Ésta es más pequeña que la de cualquiera de los 5 expertos participantes en todas las variables así como en las variables de calibración.

El cuadro 3 también muestra que el tomador de decisiones optimizado (TDPG) otorga todo el peso al experto B[11] (columnas 6-8). La puntuación de calibración del TDPG es aproximadamente 3 veces mayor que la del TDPI y la puntuación de información aproximadamente 9 veces mayor en todas las variables y 5.7 en las de calibración. Si no se realizara ninguna optimización en el TDPG entonces, después de normalizar los pesos específicos, los expertos A, D, B y el TDPG (no optimizado) serían de influencia con 2.06, 29.66, 59.32 y 8.71 % respectivamente. Obsérvese que aunque más expertos entran en la combinación, la puntuación de calibración del TDPG (no optimizado) sería 4.58 veces menor que la del TDPG optimizado. De manera similar, las puntuaciones de información en todas las variables y en las de calibración únicamente son 2.57 y 1.48 veces mayores respectivamente. En este estudio el TDPG es la elección recomendada ya que alcanza mejor desempeño que el TDPI o el TDPG-no optimizado.

11. Es decir que la rutina de optimización encuentra como valor de significancia la puntuación de calibración del experto B.

Figura 2. **Red bayesiana representando el modelo CATS para riesgos en el transporte aéreo.**



Una de las ventajas de las redes bayesianas es que cuando hay observaciones disponibles, la distribución conjunta puede ser actualizada. Este software es capaz de actualizar un modelo de las dimensiones del que se muestra en la figura 2 en minutos mientras que software comercial para redes bayesianas discretas no puede manejar un modelo como tal (Morales Nápoles *et al.,* 2007).

Conclusiones

El modelo clásico para juicios estructurados de expertos es una poderosa herramienta para epidemiólogos, científicos sociales, ingenieros y otros científicos aplicados. Su uso ha venido creciendo desde finales del siglo pasado hasta llegar a más de 67 000 pronósticos en total. Mas aún, su uso se está extendiendo al pronóstico de medidas de dependencia que son generalmente menos intuitivas tales como las correlaciones de rango y las correlaciones de rango condicionales. Este trabajo demuestra su importancia en modelos que son complejos no sólo en cuanto a la teoría matemática que emplean, sino por su tamaño tal como lo es el presentado en la figura 2.

Las redes Bayesianas para desempeño humano y los árboles de fallos mencionados anteriormente fueron posteriormente representados como una sola RBDNP. Éste modelo cuenta actualmente con 5,054 arcos y 1,503 nodos de los cuales 918 son probabilísticas y 585 funcionales que son los que representan los árboles de fallos. El modelo hasta la fecha de publicación de este documento es presentado en la figura 2.

La figura 2 presenta el modelo tal y como es mostrado en el software UniNet. Este software maneja redes Bayesianas continuas-discretas no-paramétricas y esta siendo desarrollado en el departamento de matemáticas aplicadas de la Universidad Tecnológica de Delft."

La caracterización de la incertidumbre que rodea a la comunidad científica con respecto a parámetros cuyo valor es desconocido es posible mediante el modelo clásico. En la literatura en lengua castellana son pocos los trabajos que usan esta herramienta. Se espera que este trabajo contribuya a la aceptación de esta herramienta tan completa para el avance de las ciencias aplicadas en los países de habla hispana.

ergo

Bibliografía

Ale, B.J.M. ; L.J. Bellamy; R. van der Boom; J. Cooper; R.M.Cooke; L. H. J. Goossens; A.R. Hale; D. Kurowicka; O. Morales; A.L.C. Roelen and J. Spouge (2007). *Further Developments of a Causal Model for Air Transport Safety (CATS); Building the Mathematical Heart*. Proceedings of ESREL.

Clemen, G. W. *et al.* (2000). "Assessing Dependencies: Some Experimental Results". *Management Science*. Vol. 46, Núm 8.

Clemen, G.W. *et al.*(1999). "Correlations and Copulas for Decision and Risk Analysis". *Management Science*. Vol. 45.

Cooke, R.M. (1991). *Experts in Uncertainty: Opinion ands Subjective Probability in Science*. Oxford University Press, USA.

Cooke, R.M. y L.H.J. Goossens (2008). "TU Delft Expert Judgment Data Base". *Reliability Engineering & System Safety*. Vol. 93.

Evans, J.S.; R.M. Cooke, A.M. Wilson; J.T. Tuomisto; O. Morales y M. Tainio (2007). "A probabilistic characterization of the relationship between fine particulate matter an mortality: Elicitation of european experts", *Environ. Sci. Technol.*, 41.

Genest, C. y J.V. Zidek (1986). "Combining Probability Distributions: a Critique and an Annotated Bibliography", Statistical Science. Vol 1., Núm. 1.

Hanea, A.; D. Kurowicka y R.M. Cooke (2006). "Hybrid Method for Quantifying and Analyzing Bayesian Belief Nets". *Quality and Reliability Engineering International*. Vol. 22.

Kraan, B.C.P. (2002). *Probabilistic Inversion in Uncertainty Analysis and Related Topics*. Delft University of Technology.

Kurowicka, D. y R.M. Cooke (2005). "Distribution-Free Continuous Bayesian Belief Nets", en Wilson, A.; N. Limnios; S. Keller-McNulty and Y. Armijo (eds.). *Modern Statistical and Mathematical Methods in Reliability*.

Morales-Napoles, O.; D. Kurowicka; R.M. Cooke and D. Ababei (2007). "Continuous-Discrete Distribution Free Bayesian Belief Nets in Aviation Safety with UniNet", Technical report DIAM. Elsevier.

Morales, O.; D. Kurowicka; A. Roelen (2008). "Eliciting Conditional and Unconditional Rank Correlations from Conditional Probabilities". *Reliability Engineering and System Safety. Vol.* 93.

Morales-Napoles O.; D. Kurowicka; R.M. Cooke and G.B. van Baren (2008). "Expert Elicitation Methods of Rank and Conditional Rank Correlations: An Example with Human Reliability Models in the Aviation Industry". *Working paper* DIAM. (In revision).

Vesely, W.E.; F.F. Goldberg; N.H. Roberts and D.F. Haasl (1981). *Fault Tree Handbook*. US. Nuclear regulatory Commission. NUREG-0492.

Goossens L.H.J.; R.M. Cooke; A.R. Hale (2008). *Rodic´-Wiersma Lj Fifteen years of expert judgement at TUDelft Safety Science*. 46.