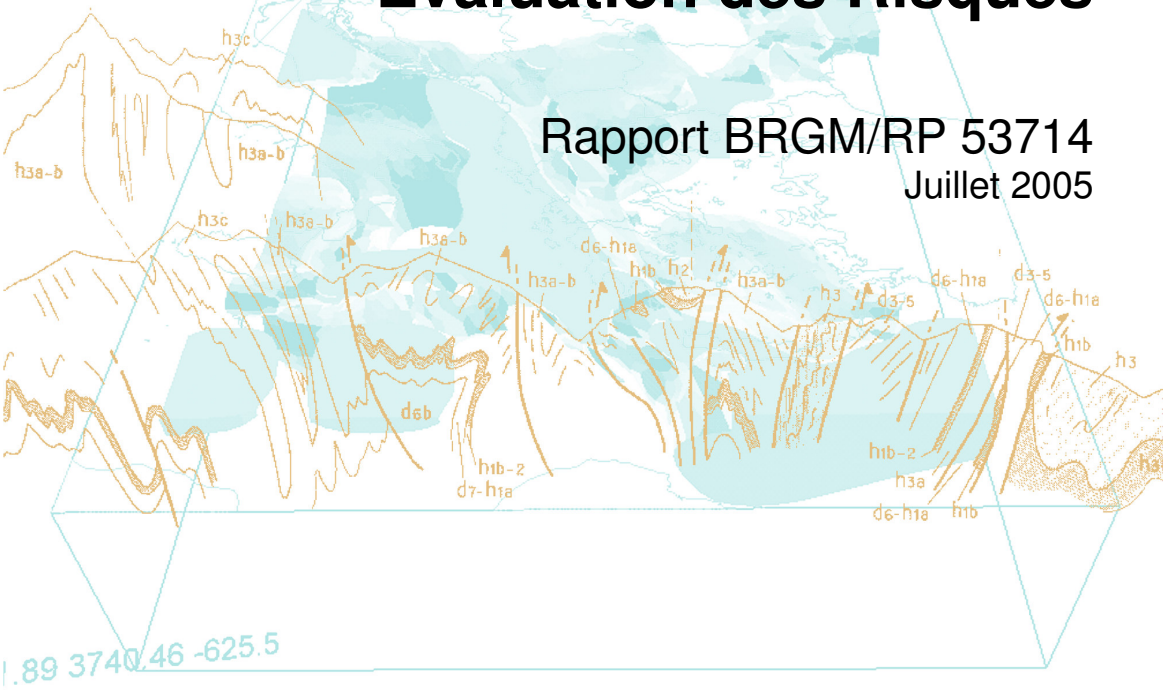
Document public

- HyRisk – Traitement Hybride des Incertitudes en Evaluation des Risques



Rapport BRGM/RP 53714

Juillet 2005

Géosciences pour une Terre durable

brgm

Document public

- HyRisk – Traitement Hybride des Incertitudes en Evaluation des Risques

Rapport BRGM/RP 53714
Juillet 2005

D. Guyonnet, Y. Ménard, C. Baudrit, D. Dubois

Vérificateur :	**Approbateur :**
Original signé par :	Original signé par :
D. CAZAUX	Ph. FREYSSINET

Le système de management de la qualité du BRGM est certifié AFAQ ISO 9001:2000

Géosciences pour une Terre durable
brgm

Mots clés : Risques, Incertitudes, Probabilités, Possibilités, Hybride.

En bibliographie, ce rapport sera cité de la façon suivante :
Guyonnet, D., Ménard, Y., Baudrit, C., Dubois, D. (2005) : HyRisk – Traitement Hybride des Incertitudes en Evaluation des Risques . Rapport BRGM/RP 53714.

Synthèse

L'incertitude est un aspect incontournable de l'évaluation des risques qui devrait être pris en compte dans l'évaluation en raison des conséquences possibles pour le processus décisionnel. L'incertitude dépend notamment de l'information dont dispose l'évaluateur par rapport aux facteurs qui influencent le risque.

Il est proposé dans ce rapport que différents types d'information relative aux variables intervenant dans un modèle d'évaluation des risques, se prêtent à différents modes de représentation de l'incertitude. En présence d'une information « riche » (p. ex. des mesures en nombre significatif), témoignant de variabilité, une représentation de type probabiliste s'applique. Par contre, face à une information plus pauvre (p. ex. du jugement d'expert), de nature incomplète ou imprécise (situation d'ignorance partielle), d'autres modes de représentation peuvent être jugés préférables. Lorsque l'information disponible ne permet pas de justifier une distribution de probabilité unique, une alternative consiste à utiliser des *familles* de distributions de probabilité.

Ce rapport illustre le lien entre des familles de distributions de probabilité et des distributions dites de « possibilité » (aussi appelés « intervalles flous »). Il présente la méthode dite « hybride », qui permet de propager conjointement une information de type probabiliste et possibiliste dans l'estimation de l'incertitude sur le risque calculé.

La théorie sous-jacente à la méthode hybride est présentée de manière didactique. Un didacticiel appelé HyRisk (téléchargeable gratuitement sur http://www.brgm.fr/hyrisk), qui met en œuvre cette méthode, est décrit et illustré par un exemple qui se réfère plus particulièrement au domaine du risque sanitaire.

Avertissement

Ni le BRGM, ni aucun autre organisme, n'apportent la moindre garantie quant au fonctionnement du didacticiel HyRisk, et n'assument aucune responsabilité quant à :

- son usage correct ou incorrect,
- l'interprétation exacte ou erronée des résultats calculés,
- d'éventuels dommages occasionnés par son utilisation.

Sommaire

Liste des illustrations

1. Introduction

1.1. OBJECTIFS

L'évaluation des risques (pour l'eau, la santé, l'environnement, les biens matériels, etc.) est un outil d'aide à la gestion des sites potentiellement pollués (MEDD, 2000). Ces risques sont évalués à partir d'études de terrain qui permettent de caractériser la qualité des différents milieux d'exposition, mais aussi de prédictions (de transferts, de transformations, d'expositions, d'effets sanitaires, etc.) établies à l'aide de modèles de calcul.

Par définition, toute prédiction effectuée à partir d'un modèle de calcul est affectée d'incertitude. Les sources d'incertitude sont multiples et concernent notamment le modèle lui-même (modèle conceptuel) ou encore les valeurs des variables qui interviennent dans ce modèle. Le guide MEDD préconise que les incertitudes soient prises en compte dans les évaluations.

De nombreuses méthodes existent pour traiter le problème des incertitudes. Ces méthodes vont du simple calcul d'intervalle aux simulations Monte Carlo (1D, 2D, …) ou aux méthodes faisant appel aux probabilités imprécises.

A l'origine de l'élaboration des méthodes mises en œuvre par le didacticiel HyRisk, il y a la recherche de deux principaux objectifs :

1) représenter les incertitudes relatives aux variables d'un modèle d'une manière cohérente par rapport aux informations dont on dispose vraiment concernant ces variables,
2) propager ces incertitudes en vue d'estimer l'incertitude relative au résultat du modèle.

Ce rapport présente, de manière didactique, la théorie sous-jacente à HyRisk ainsi que l'utilisation de l'outil. On notera qu'il n'est pas indispensable de maîtriser tous les détails théoriques inhérents à cette méthodologie pour pouvoir utiliser HyRisk. La méthode est néanmoins présentée en détail et l'utilisateur intéressé pourra se référer aux publications qui sont citées dans la bibliographie.

On notera que l'outil est appelé « didacticiel » car le principal objectif est de sensibiliser le lecteur par rapport à certains aspects du traitement des incertitudes. L'outil est construit dans un environnement très convivial, mais en contrepartie la complexité des modèles de calcul qui peuvent être exploités dans HyRisk est très limitée (formules mathématiques incluant des opérations simples). Une application de la méthodologie proposée à des modèles de calcul de risques plus complexes nécessiterait un autre environnement de développement.

1.2. VARIABILITE ET IMPRECISION : DEUX FACETTES DISTINCTES DE L'INCERTITUDE

Un aspect fondamental du traitement de l'incertitude en prédiction du risque a trait à l'utilisation de la notion de probabilité. Développée à partir du 17ème siècle avec une référence toute particulière aux jeux de hasard, la théorie des probabilités permet d'appréhender des processus dont la réalisation présente une certaine variabilité qui est le fruit du hasard (processus aléatoires). Un exemple simple de processus aléatoire est le résultat découlant de jets successifs d'un dé à 6 faces non pipé. On se trouve là dans le cas d'un système qui peut être assimilé à un système fermé (Oreskes et al., 1994) : la probabilité de chaque événement possible (résultat d'un jet) est connue (1/6) et cette probabilité ne changera pas en cours de jeu en raison d'un éventuel événement extérieur (cas d'un système ouvert).

Mais une des difficultés liées à l'utilisation de la théorie des probabilités classique dans le domaine des risques environnementaux a trait, d'une part, au fait que les systèmes considérés ne sont pas fermés mais ouverts, mais aussi au choix non ambigu de distributions de probabilité lorsque l'information disponible est de nature incomplète ou imprécise. Supposons que l'évaluateur du risque puisse « fermer », par la pensée, le système étudié en y accolant un « modèle ». Il se peut que pour une certaine variable (notée *x*) intervenant dans ce modèle, il ne sache rien d'autre que le fait que la valeur de cette variable est nécessairement supérieure à une certaine valeur *min* et nécessairement inférieure à une certaine valeur *max*. Une telle information n'exprime pas une variabilité de type aléatoire mais de l'imprécision (ou ignorance partielle). La variable est peut-être une variable aléatoire, pouvant être décrite par une distribution de probabilité unique, mais l'information dont dispose l'observateur ne lui permet pas de spécifier cette distribution. A noter qu'on parle aussi (Ferson et Ginzburg, 1996) d'incertitude objective dans le cas de la variabilité aléatoire (car liée à la nature aléatoire d'un phénomène réel) et d'incertitude subjective dans le cas de l'imprécision (car liée à la méconnaissance qu'a le sujet de ce phénomène réel).

Une convention en statistique qui date du début du développement de cette méthode consiste dans ce cas à répartir de manière uniforme la probabilité sur toutes les valeurs situées entre les bornes *min* et *max* (distribution uniforme ; Figure 1). Cette convention, connue sous le nom du principe de Laplace de raison insuffisante et qui a été généralisée dans le cadre de la méthode de l'entropie maximale (Levine et Tribus, 1978, Gzyl, 1995) est justifiée par ses utilisateurs par le fait que si on ne sait rien sur les probabilités respectives des valeurs situées dans l'intervalle, alors supposer toute autre distribution que la distribution uniforme reviendrait à supposer une information qu'on ne possède pas. Mais cette approche a été critiquée, notamment par d'éminents mathématiciens et statisticiens (Boole, 1929 ; Fisher, 1973), qui soulignent que l'hypothèse d'une distribution de probabilité unique en présence d'une information de type *min-max* revient à ne choisir qu'une seule parmi toutes les distributions possibles ayant le même support (Figure 2).

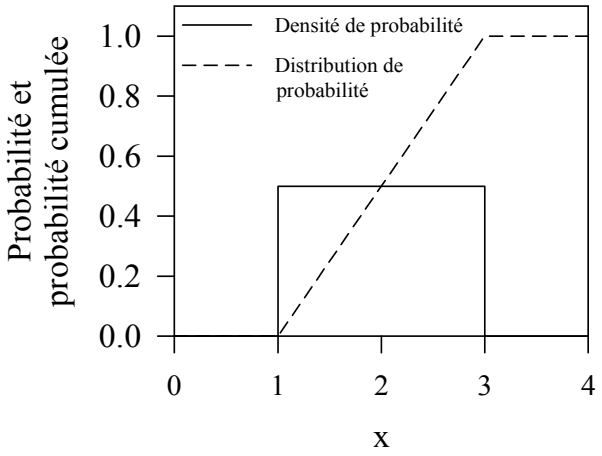


Figure 1 – Distribution de probabilité uniforme de support [1,3] : fonction de densité de probabilité et fonction de distribution de probabilité associée

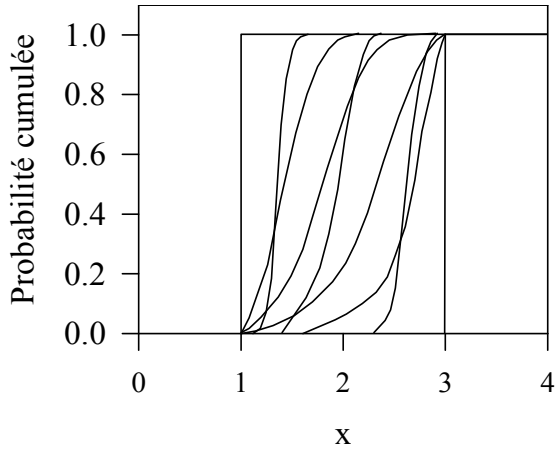


Figure 2 – Quelques représentants de la famille des distributions de probabilité dont le support est compris dans l'intervalle [1,3]

Une illustration très simple des éventuelles conséquences d'un amalgame entre variabilité aléatoire et imprécision, sous hypothèse d'indépendance entre les paramètres, est obtenue en considérant le problème suivant (voir Ferson, 1996). Supposons que l'on ait deux paramètres non reliés, A et B, dont on ne sait que peu de choses si ce n'est que A est situé entre 0 et 1, tandis que B est situé entre 1 et 2. On s'intéresse à la somme de A et B et plus particulièrement à la valeur moyenne de cette somme. Un simple calcul d'intervalle montre que la somme de A et B appartient à l'intervalle [1-3]. La valeur moyenne appartient également à cet intervalle : sur la base de l'information disponible on ne peut pas exprimer de préférence à l'intérieur de cet intervalle. En adoptant maintenant le principe de Laplace, on supposera que les paramètres A et B sont représentés par des fonctions de densité de probabilité

uniformes (on suppose l'équirépartition des probabilités sur toutes les valeurs des intervalles). Dans ce cas la somme de A et B est une fonction de densité de probabilité triangulaire de valeur moyenne 2 (ici égale au mode) et de valeurs minimum et maximum 1 et 3 respectivement. On constate donc que le choix *a priori* de distributions de probabilité uniques aboutit à un résultat beaucoup plus « précis » sur la valeur moyenne qu'avec le calcul d'intervalle. Mais cette précision est illusoire car on a supposé l'équirépartition des probabilités, alors que rien dans l'information disponible ne permettait d'étayer cette hypothèse.

En situation d'ignorance partielle, certaines approches dites « Bayésiennes » utilisent des distributions de probabilité *a priori* (ou subjectives) comme points de départ, puis corrigent ces distributions au fur et à mesure que de nouvelles informations deviennent disponibles (en appliquant le théorème de Bayes des probabilités conditionnelles ; « Bayesian updating »). Ces méthodes ont notamment été appliquées avec succès dans les domaines de l'exploration minière ou pétrolière où, durant la phase d'exploration d'un gisement, des données provenant de nouveaux forages permettent de corriger les simulations sur des teneurs en minerai ou des réserves d'hydrocarbures. Mais dans le domaine de l'évaluation des risques liés aux sources de pollution, il est rare que de nouvelles informations permettent de corriger les hypothèses initiales. Les distributions *a priori* deviennent dès lors des distributions *a posteriori*, introduisant une précision illusoire que ne justifie pas l'information réellement disponible.

En résumé, variabilité aléatoire et imprécision (ignorance partielle) sont deux facettes de l'incertitude en évaluation des risques qu'il convient de distinguer pour un traitement cohérent de l'incertitude (Casti, 1990, Ferson et Ginzburg, 1996). Cela est vrai tout particulièrement dans le domaine de l'évaluation des risques environnementaux ou sanitaires où l'on a fréquemment affaire à ces deux types d'information. Plutôt que de supposer a priori des distributions de probabilité uniques en présence d'imprécision (les « tirer d'un chapeau »), comment rester fidèle à l'information d'origine et en tenir compte dans la prédiction ?

Le didacticiel HyRisk met en oeuvre une méthode qui permet de distinguer entre ces deux formes d'incertitude dans l'évaluation du risque et de les propager dans le calcul de risque. Si on dispose d'une information « riche » (mesures en nombre significatif, informations statistiques, …) on s'orientera plutôt vers une approche statistique voire géostatistique (Chilès et Delfiner, 1999) de l'incertitude. Par contre, dans le cas d'une information « pauvre » (mesures éparses, variables mal connues, jugement d'expert, …), l'approche proposée dans ce manuel, qui utilise la notion de probabilités imprécises, peut être jugée utile.

2. Methodologie utilisée par HyRisk

2.1. INTRODUCTION

Le didacticiel HyRisk permet de représenter les incertitudes relatives aux paramètres d'un modèle et de les propager dans l'incertitude relative au résultat du modèle. Il est donc fait l'hypothèse que l'évaluateur dispose d'un modèle. HyRisk n'aborde pas le problème de l'incertitude relative au modèle conceptuel. Une manière d'aborder cet aspect pourrait consister à utiliser plusieurs modèles « envisageables », puis d'effectuer une fusion des résultats provenant de ces différents modèles. Mais à défaut de modèle « crédible », il est peut-être préférable que l'investigateur renonce à faire de la prédiction quantitative pour s'orienter plutôt vers une approche qualitative de type hiérarchisation multicritère, ou alors qu'il retourne à l'étude du système réel pour tenter d'identifier un modèle applicable.

Pour représenter et propager les incertitudes relatives aux paramètres d'un modèle, le didacticiel HyRisk fait appel à trois « outils » :

- l'échantillonnage aléatoire de fonctions de distribution de probabilité (PDF),
- le calcul d'intervalle flou,
- les fonctions de croyance de Dempster-Shäfer.

Ce chapitre présente ces différents outils de manière simple en les illustrant graphiquement.

2.2. ECHANTILLONNAGE DE DISTRIBUTIONS DE PROBABILITE

On considère un paramètre Y qui est fonction de plusieurs variables X_i dont les valeurs sont le fruit du hasard (variables aléatoires) :

$Y = f(X_1, X_2, \ldots X_n)$

Les fonctions de distribution de probabilité définissant les variabilités respectives de X_1, .., X_n sont supposées connues.

Si les distributions sont simples (distribution uniforme, normale, etc.) et que la fonction f est simple (combinaison d'additions, de soustractions, etc.), la distribution de probabilité décrivant la variabilité de Y peut être calculée directement.

Si ces distributions et/ou la fonction sont plus complexes, une approche robuste consiste à effectuer un échantillonnage aléatoire des distributions, à calculer Y et à recommencer l'opération un grand nombre de fois (méthode dite Monte Carlo ; Saporta, 1990, Vose, 1996). On obtient ainsi une distribution sur le paramètre Y.

Pour effectuer un échantillonnage aléatoire, on tire de manière aléatoire dans une distribution uniforme (à l'aide d'un algorithme spécifique, par exemple la fonction Alea d'Excel), un nombre (qu'on notera ν ; Figure 3) compris entre 0 et 1. On assimile ce nombre à une probabilité cumulée et on va chercher la valeur du paramètre correspondant à cette valeur de probabilité cumulée (Figure 3).

Si on fait l'hypothèse de l'indépendance entre les variables aléatoires X_1, …, X_n, alors on tire un nombre aléatoire pour chaque variable. Si par contre on a connaissance d'éventuelles corrélations entre les variables, celles-ci peuvent être prises en compte (Connover et Iman, 1982). On notera que la version 1 du didacticiel HyRisk ne permet pas cette possibilité (le didacticiel suppose l'indépendance entre les variables aléatoires).

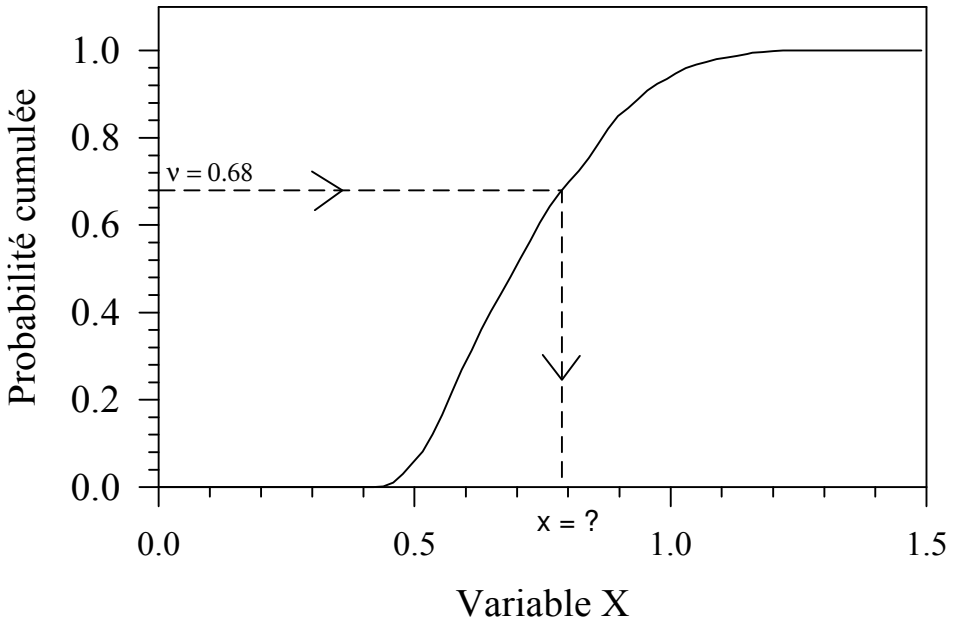


Figure 3 – Schéma représentant le tirage aléatoire d'une valeur x de la variable X.

2.3. DISTRIBUTIONS DE POSSIBILITE

2.3.1. Qu'est-ce qu'une distribution de possibilité ?

De la même manière qu'une distribution de probabilité permet de représenter, en langage mathématique, une information relative à une grandeur dont la valeur dépend du hasard (variabilité aléatoire), une distribution de possibilité permet de représenter une information qui est incomplète ou imprécise (situation d'ignorance partielle) (Zadeh, 1978 ; Dubois et Prade, 1988).

Une distribution de possibilité permet d'affiner et d'enrichir la notion de simple intervalle min-max par l'expression de préférences au sein de cet intervalle. Comme on le verra plus loin, cette représentation se prête tout particulièrement à la représentation de jugement d'expert.

Supposons, par exemple, qu'on demande à un expert agronome de fournir une estimation du pH d'un sol pour lequel il ne dispose que de quelques mesures éparses (sans possibilité d'en effectuer d'autres). Sur la base de ces quelques mesures, mais aussi de son expérience, l'expert pourrait fournir l'information suivante :

- Il estime que les valeurs de pH de ce sol se situent vraisemblablement entre 6.5 et 7.5,
- Il n'exclut pas des valeurs aussi faibles que 6 ou aussi fortes que 8.

Pour représenter cette information, on normalise tout d'abord la vraisemblance maximale à 1. Les valeurs appartenant à l'intervalle [6.5-7.5] ont donc une vraisemblance de 1 (on appelle cet intervalle le *noyau* de la distribution de possibilité ; Figure 4). C'est l'intervalle le plus précis, mais aussi le plus « risqué » (le degré de « certitude » que les valeurs se trouvent bien dans cet intervalle est le plus faible)

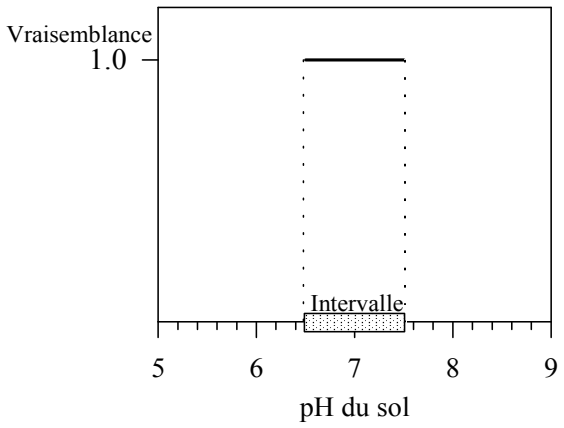


Figure 4 – Valeurs de pH d'un sol : intervalle de valeurs jugé le plus vraisemblable

Au fur et à mesure qu'on écarte les bornes gauche et droite de cet intervalle, on englobe des valeurs qui sont jugées de moins en moins vraisemblables. Jusqu'à arriver à l'intervalle [6-8], en dehors duquel l'expert estime que les valeurs sont impossibles (vraisemblance = 0 ; Figure 5). Cet intervalle, qui contient les valeurs dont la vraisemblance > 0, est appelé *support* de la distribution de possibilité. C'est l'intervalle le plus « sûr » mais le moins informatif 'le degré de certitude que les valeurs se trouvent bien dans cet intervalle est le plus élevé).

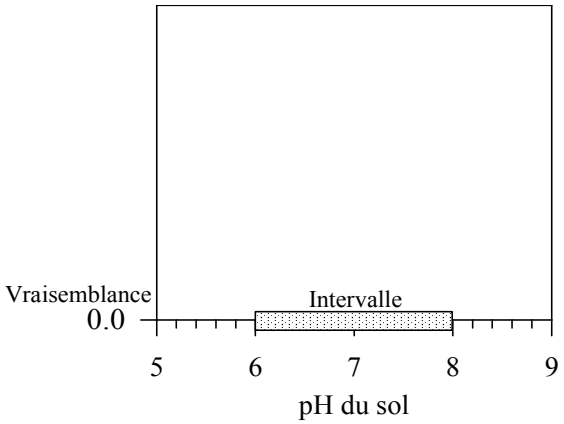


Figure 5 – Valeurs de pH d'un sol : intervalle de valeurs en-dehors duquel les valeurs sont jugées impossibles

A défaut d'information plus précise, on suppose typiquement une transition linéaire entre ces vraisemblances maximales et minimales (Figure 6). On définit ainsi une distribution de possibilité (la vraisemblance est appelée *possibilité*), qu'il faut voir comme un emboîtement d'intervalles de valeurs, chacun ayant son propre degré de vraisemblance.

On notera qu'une transition courbe peut être envisagée si l'expert possède des informations qui vont dans ce sens. Si par exemple les valeurs du support situées en dehors du noyau lui paraissent possibles, mais peu vraisemblables, il pourra choisir de donner une courbure concave aux branches latérales de la Figure 6.

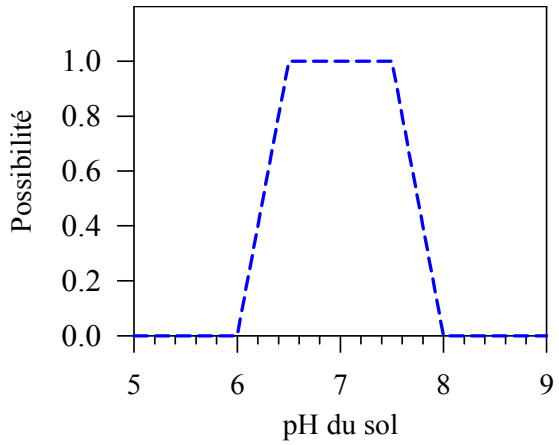


Figure 6 – Distribution de possibilité relative aux valeurs de pH d'un sol

Si au contraire elles sont jugées tout à fait possibles, la courbure sera plutôt convexe (jusqu'au cas limite où la distribution devient un simple intervalle min-max). Il est clair qu'une certaine part de subjectivité entrera dans la définition de ces courbures. Mais cette subjectivité aura beaucoup moins de conséquences sur les résultats de l'analyse, que celles découlant d'un choix *a priori* de distributions de probabilité uniques en présence d'ignorance partielle (information incomplète/imprécise).

On notera que cette notion de *possibilité* relative aux valeurs précises appartenant aux intervalles emboîtés est complémentaire de celle de *certitude* relative aux intervalles eux-mêmes. En effet, plus on élargit l'intervalle, plus on inclut des valeurs jugées peu vraisemblables, mais en contrepartie, plus on est certain que les vraies valeurs se trouvent bien à l'intérieur de l'intervalle (Figure 7). Ainsi, le degré de possibilité d'un intervalle peut être défini comme étant égal à 1 moins le degré de certitude que l'intervalle contient bien les vraies valeurs.

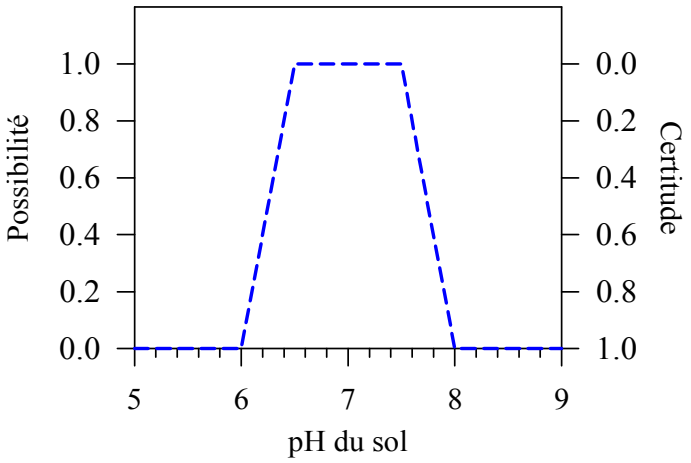


Figure 7 – Dualité possibilité - certitude

Ces notions de certitude et de vraisemblance peuvent être reliées à celle de probabilité imprécisément connue. Si on s'intéresse à la distribution de probabilité décrivant la probabilité que le pH soit inférieur à telle ou telle valeur X, tant que X reste supérieur à 8 (maximum du support), la probabilité est : P(pH < X) = 1 (on est « sûrs » que pH < X ; Figure 8).

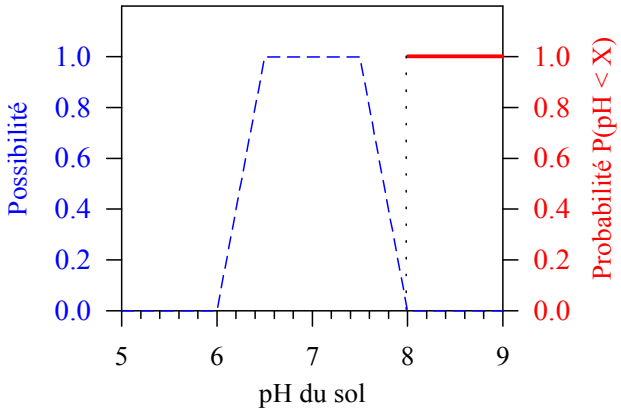


Figure 8 – Construction de la probabilité P(pH < X)

Lorsqu'on commence à entrer dans l'intervalle du support, la probabilité P(pH < X) n'est plus nécessairement égale à 1. Prenons par exemple la valeur 7.9 : cette valeur est la borne supérieure de l'intervalle dont le degré de certitude est estimé à 0.8 (Figure 7). On peut montrer que la probabilité P(pH < X) est comprise entre 0.8 et 1 (Figure 9).

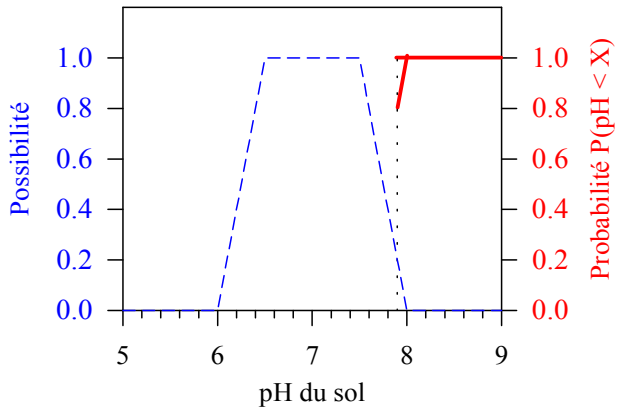


Figure 9 – Construction de la probabilité P(pH < X) (suite)

En procédant de la même manière jusqu'à la valeur X = 7.5, on obtient la Figure 10.

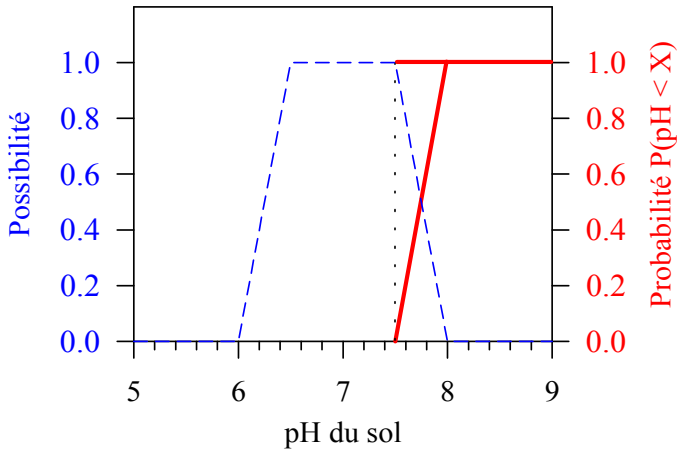


Figure 10 – Construction de la probabilité P(pH < X) (suite)

Dès lors qu'on est dans le noyau de la distribution ([6.5-7.5]) on a affaire au classique intervalle min-max. : sur la base de l'information disponible, on ne peut exprimer de préférences au sein de cet intervalle. La probabilité P(pH < X) pour X appartenant à cet intervalle est comprise entre 0 et 1 (Figure 11).

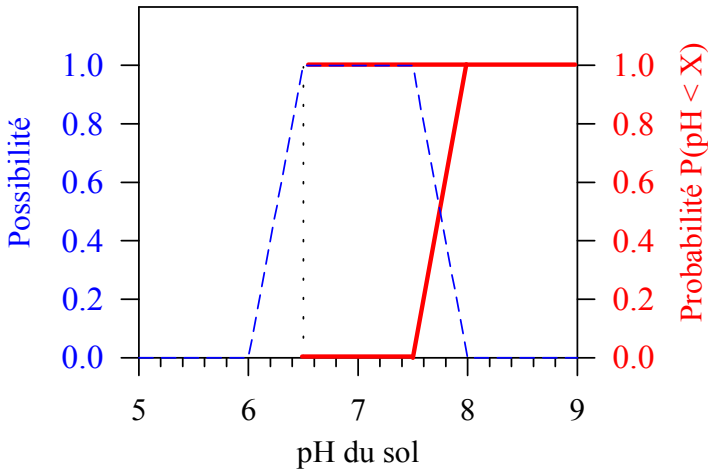


Figure 11 – Construction de la probabilité P(pH < X) (suite)

Considérons maintenant les valeurs situées à gauche du support. Pour toutes les valeurs X < 6, l'information fournie par l'expert implique que P(pH < X) = 0 (on est « sûr » de ne pas être inférieur à X ; Figure 12).

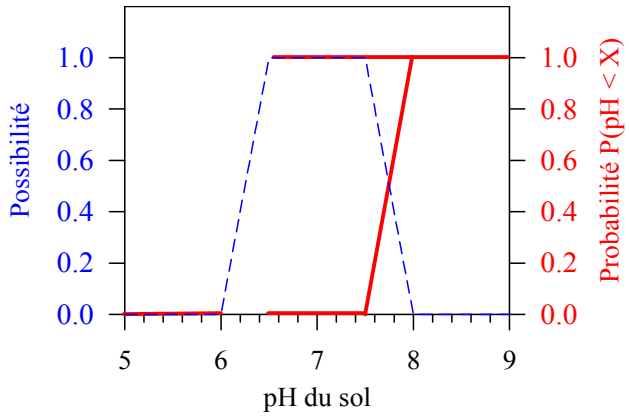


Figure 12 – Construction de la probabilité P(pH < X) (suite)

Pour 6 < X < 6.5, on peut encadrer P(pH < X) de la même manière que précédemment pour 7.5 < X < 8 (Figure 13).

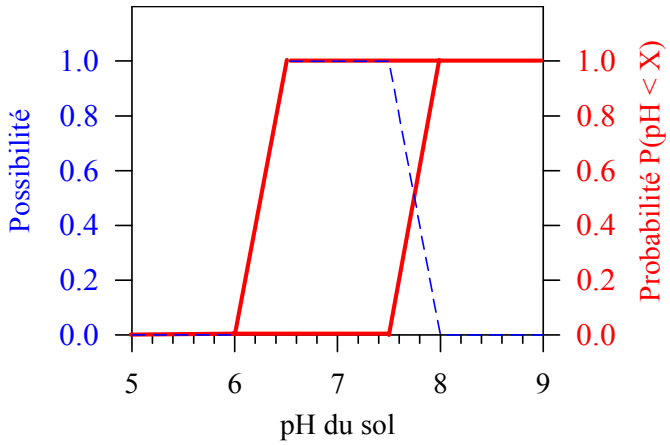


Figure 13 – Construction de la probabilité P(pH < X) (fin)

On voit donc qu'une distribution de possibilité peut être considérée comme représentant une *famille* de distributions de probabilité (voir aussi Dubois et Prade, 1992). La distribution « vraie » se situe quelque part entre les deux bornes de la Figure 13, mais la nature incomplète de l'information dont on dispose ne nous permet pas d'en sélectionner une seule. L'écart entre les distributions gauche (haute) et droite (basse) est une mesure de notre ignorance.

On notera qu'en théorie des possibilités (Dubois et Prade, 1988), ces deux distributions de probabilité limitantes sont appelées « Mesure de Possibilité » (notée Π) pour la distribution haute (celle de gauche) et « Mesure de Nécessité (notée N) pour la distribution basse (celle de droite).

2.3.2. Le calcul d'intervalle flou

Soit un paramètre Z qui est fonction d'un certain nombre de paramètres (notés a) représentés par des distributions de possibilité (intervalles flous) :

$Z = f(a_1, a_2, \ldots, a_n)$

Si la fonction f est simple (combinaison d'opérations simples), le calcul peut être effectué directement.

On considère par exemple deux paramètres, a_1 et a_2 de noyaux et supports respectifs (Figure 14a et b) :

- a_1 : support = [10 - 50], noyau = [12 - 20]
- a_2 : support = [30 - 80], noyau = [50 - 60]

On suppose : $Z = a_1 + a_2$. Dans ce cas, la valeur min du support de Z est simplement : 10 + 30 = 40 (somme des valeurs minimales) et la valeur max est : 50 + 80 = 130 (somme des valeurs max). Le support de Z est donc l'intervalle [40 - 130]. Pour le noyau on procède de la même manière pour obtenir l'intervalle [62 – 80]. L'intervalle flou de $Z = a_1 + a_2$ est représenté dans la Figure 14c.

A noter que les branches reliant les noyaux et les supports de a_1 et a_2 étant linéaires, et la fonction étant simple, on peut se contenter dans ce cas de faire le calcul uniquement sur les valeurs des noyaux et supports. Si, au contraire, on avait des courbes (ou une fonction complexe), il faudrait procéder de la même manière que précédemment pour des niveaux de possibilité intermédiaires entre 0 et 1 et construire tout l'intervalle flou de Z.

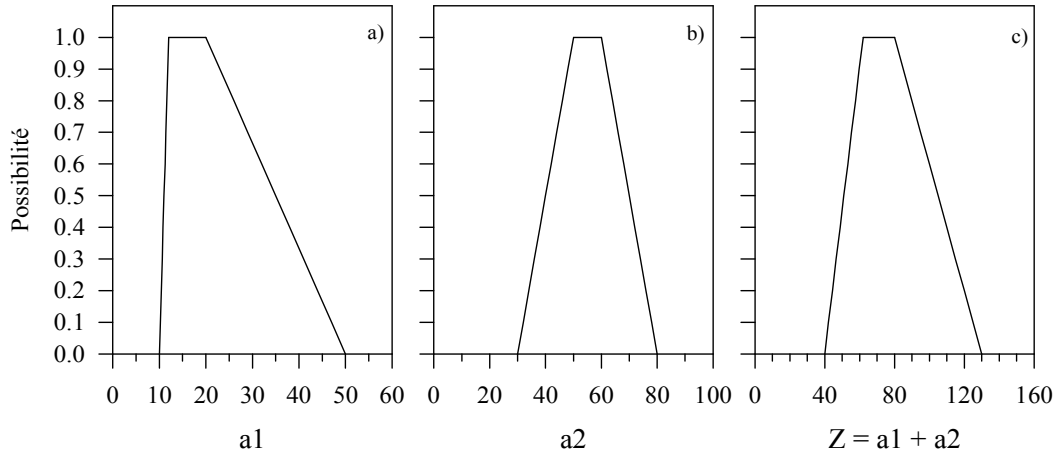


Figure 14 – Exemple simple de calcul d'intervalle flou : l'addition

Une autre conséquence de la simplicité de la fonction considérée dans le cas de la Figure 14 est qu'il est facile d'identifier les valeurs min et max sur le noyau et le support. Dans le cas de fonctions plus complexes, on utilise dans la pratique des

algorithmes de minimisation et de maximisation pour déterminer les valeurs min et max de Z en balayant sur toutes les valeurs du support, du noyau, et d'intervalles de valeurs correspondant à des niveaux de possibilité intermédiaires (en théorie des possibilités on appelle ces intervalles des alpha-coupes : les intervalles de valeurs dont le degré de possibilité est supérieur ou égal à une valeur alpha). Ce processus est illustré plus loin dans la section relative à la méthode hybride.

En conclusion de cette section on peut dire que le calcul d'intervalle flou est en fait un calcul d'intervalle classique mais effectué sur des intervalles correspondant à différents degrés de vraisemblance.

2.3.3. Représentation possibiliste et jugement d'expert

Comme il a été vu plus haut, la distribution de possibilité (intervalle flou) est simplement un intervalle qui a été enrichi d'une information supplémentaire : des plages de valeurs situées à l'intérieur de l'intervalle du support ont été qualifiées pour exprimer leur plus grande vraisemblance. Ces intervalles sont emboîtés : il s'agît là d'une particularité (et d'une limite ; voir section suivante) des distributions de possibilité. Mais cette caractéristique se prête tout particulièrement à une situation couramment rencontrée dans la pratique : le jugement d'expert. En effet, on peut s'attendre à ce qu'un expert soit cohérent avec lui-même : l'intervalle contenant les valeurs qu'il juge les plus vraisemblables est forcément inclus dans l'intervalle, plus large, des valeurs qu'il juge éventuellement possibles.

Dans la section suivante, on verra une situation où les intervalles ne sont pas nécessairement emboîtés mais peuvent se recouper.

2.4. LES FONCTIONS DE CROYANCE DE DEMPSTER-SHAFER

2.4.1. Introduction

La théorie des fonctions de croyance (Shafer, 1976), aussi appelée « théorie de l'évidence » ou « théorie de Dempster-Shafer », assigne des masses de probabilité à ce qu'elle appelle des « éléments focaux ». Ces éléments focaux peuvent êtres des valeurs ponctuelles ou des intervalles adjacents (cas d'une fonction de densité de probabilité classique) ou encore des intervalles emboîtés (cas de la théorie des possibilités illustrée précédemment). Mais la force de cette théorie réside dans le fait qu'elle s'applique également au cas où les intervalles se recoupent (ni adjacents ni emboîtés ; voir l'exemple de la section suivante). Elle fournit ainsi un mode de représentation de l'information qui s'applique à la fois au problème de la variabilité (théorie des probabilités classiques) et de l'imprécision ou ignorance partielle.

Les éléments focaux pour le cas d'une fonction de densité de probabilité classique sont illustrés par la (Figure 15). Les éléments focaux sont les intervalles de valeurs obtenus par un découpage vertical de la fonction. Dans le cas d'une distribution de possibilité,

par contre, les éléments focaux sont les intervalles de valeurs obtenus par découpage horizontal de la distribution (Figure 16). Le nombre de tranches utilisé dépend de la précision souhaitée (le degré de reproduction de la forme de la fonction initiale). Les masses de probabilité assignées à ces éléments focaux sont fonction de ce découpage. A noter que dans tous les cas, la somme des masses de probabilité de tous les éléments focaux est égale à 1.

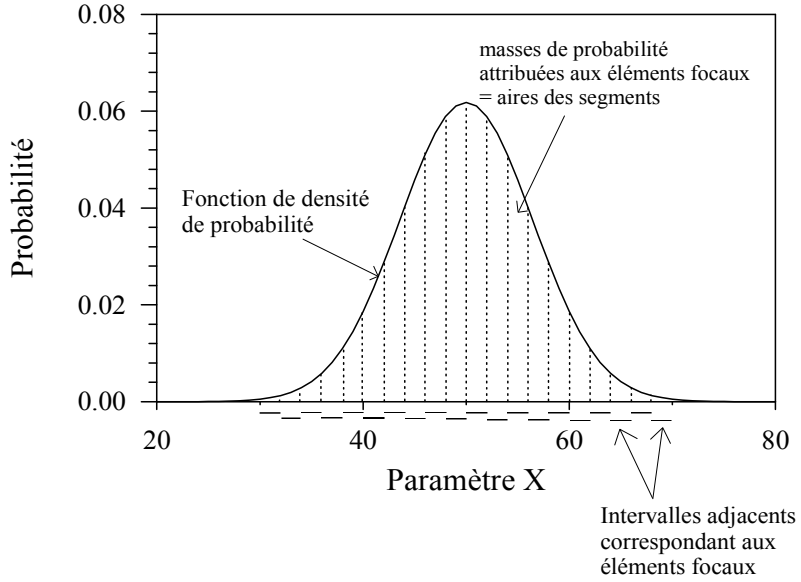


Figure 15 – Eléments focaux de Dempster-Shafer et masses de probabilité : cas d'une distribution de probabilité classique

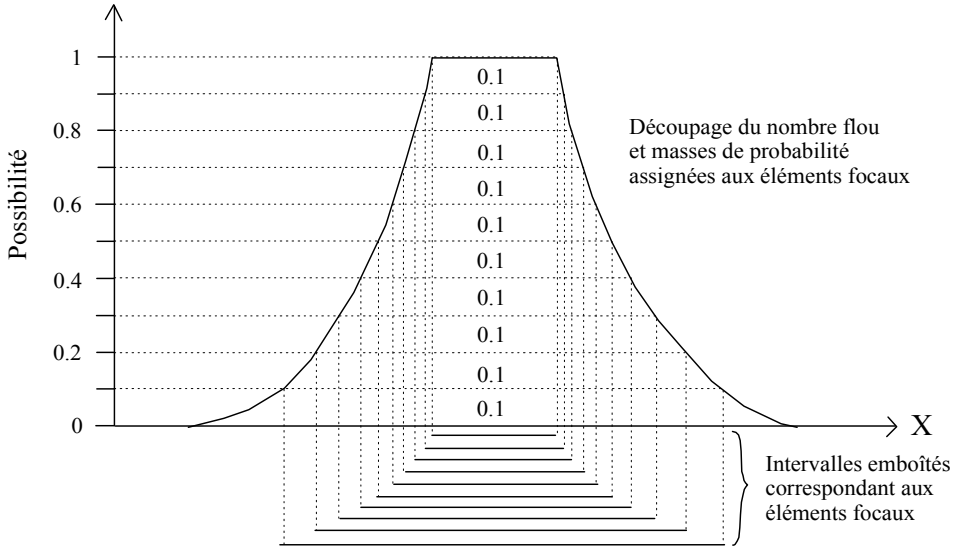


Figure 16 – Eléments focaux de Dempster-Shafer et masses de probabilité : cas d'une distribution de possibilité

Dans la théorie de Dempster-Shafer, la véracité d'une proposition est qualifiée à l'aide de deux indicateurs : la crédibilité (notée Cr) et la plausibilité (notée Pl). La crédibilité est la somme des probabilités de tous les éléments focaux qui impliquent forcément la proposition recherchée. Tandis que la plausibilité est la somme des probabilités de tous les éléments focaux qui ne contredisent pas nécessairement la proposition recherchée. On verra dans la section suivante, une illustration simple de ces deux indicateurs.

2.4.2. Illustration des fonctions de croyance de Dempster-Shafer

Pour cette illustration (à l'origine due à P. Smets), on considère le cas d'un sondage effectué dans le cadre d'une élection présidentielle. Il y a 5 candidats pour cette élection : 2 candidats de gauche (notés G_1 et G_2) et 3 candidats de droite (D_1, D_2 et D_3).

Le sondage donne les résultats suivants :

- 20% des personnes interrogées sont certaines de voter pour le candidat G_1,
- 5% pour G_2,
- 30% pour D_1,
- 10% savent qu'ils voteront pour un candidat de gauche, mais ne savent encore pas pour qui,
- 10% savent qu'ils voteront pour un candidat de droite, mais ne savent pas encore pour qui,
- 25% ne savent pas encore pour quel candidat ils vont voter.

Les masses de probabilité des éléments focaux (entre parenthèses) sont donc les suivantes :

- $m(G_1) = 0.2$
- $m(G_2) = 0.05$
- $m(D_1) = 0.3$
- $m(G_1 \text{ ou } G_2) = 0.1$
- $m(D_1 \text{ ou } D_2 \text{ ou } D_3) = 0.1$
- $m(G_1 \text{ ou } G_2 \text{ ou } D_1 \text{ ou } D_2 \text{ ou } D_3) = 0.25$

La somme des masses est bien égale à 1. La masse m(S) est une masse de probabilité qui devrait idéalement être distribuée parmi les éléments de S mais qui demeure « suspendue » en raison de l'absence d'information. Par exemple, dans le cas des personnes sans opinion, la masse de probabilité est assignée à l'ensemble des candidats ; l'hypothèse *a priori* d'une équirépartition de cette masse sur les différents candidats est ainsi évitée.

On s'intéresse maintenant à la proposition (notée A) selon laquelle les électeurs votent pour un candidat de gauche et on cherche à estimer la probabilité de cet événement.

En appliquant la théorie des fonctions de croyance, cette probabilité est comprise entre deux bornes : la crédibilité et la plausibilité. La crédibilité est la somme des probabilités des éléments focaux qui impliquent forcément l'événement recherché. D'après le sondage on a : $Cr(A) = m(G_1) + m(G_2) + m(G_1 \text{ ou } G_2) = 0.2 + 0.05 + 0.1 = 0.35$. C'est la plus petite probabilité de réalisation de l'événement. La plausibilité est la somme des probabilités des éléments focaux qui ne contredisent pas nécessairement l'événement recherché. On a donc : $Pl(A) = m(G_1) + m(G_2) + m(G_1 \text{ ou } G_2) + m(G_1 \text{ ou } G_2 \text{ ou } D_1 \text{ ou } D_2 \text{ ou } D_3) = 0.2 + 0.05 + 0.1 + 0.25 = 0.6$. En effet, l'élément focal (G_1 ou G_2 ou D_1 ou D_2 ou D_3) ne contredit pas nécessairement l'événement recherché, puisque deux candidats de gauche y figurent. On obtient donc pour la probabilité P(A) de l'événement A : $0.35 \leq P(A) \leq 0.6$.

A contrario, une approche qu'on pourrait qualifier de « Bayésienne », consisterait à faire une hypothèse sur la répartition des probabilités dans le cas des personnes indécises. Si on suppose (cas le plus fréquent) l'équirépartition de ces probabilités, on distribue alors uniformément la probabilité de 25% sur chacun des 5 candidats de l'élément focal (G_1 ou G_2 ou D_1 ou D_2 ou D_3) pour obtenir :

$P(A) = m(G_1) + m(G_2) + m(G_1 \text{ ou } G_2) + 0.05 + 0.05 = 0.45$.

Cette valeur se situe forcément entre les valeurs de Cr et Pl précédentes. Mais l'objection qu'on peut émettre par rapport à cette approche est que l'information dont on dispose ne nous permet pas de supposer une telle répartition uniforme. Les indicateurs Cr et Pl sont « fidèles » à l'information disponible : ils permettent d'éviter l'hypothèse *a priori* de la répartition des probabilités. L'écart entre Cr et Pl est une mesure de notre ignorance. Lorsqu'il n'y a pas d'imprécision (toutes les personnes interrogées savent pour qui elles vont voter), on a Cr = Pl et on retombe sur les probabilités classiques.

Ces deux indicateurs, Cr et Pl, seront utilisés plus loin pour restituer les résultats du traitement d'incertitude proposé.

2.5. LA METHODE HYBRIDE

La méthode hybride (Guyonnet et al., 2003 ; Baudrit et al., 2004) est une réponse simple, sur le plan intuitif, au problème de la combinaison d'informations de natures aléatoire et imprécise dans l'estimation du risque. Elle combine simplement l'échantillonnage aléatoire de type Monte Carlo avec le calcul d'intervalle flou.

Supposons qu'on ait un modèle de risque (noté R) qui soit une fonction de n variables représentées par des distributions de probabilité ; P_1, …, P_n et de m variables représentées par des distributions de possibilité (intervalles flous) ; F_1, .., F_m.

Le calcul hybride est un processus itératif qui consiste à (Figure 17) :

1. Tirer n nombres aléatoires compris entre 0 et 1 (ν_1, …, ν_n) d'une distribution uniforme et échantillonner les n distributions de probabilité pour obtenir une réalisation (p_1, …p_n) des variables aléatoires (Figure 17a)
2. Sélectionner un degré de possibilité α.
3. Chercher les valeurs *Inf* (plus petite valeur) et *Sup* (plus grande valeur) de $R(p_1, \ldots, p_n, F_1, \ldots, F_m)$, en considérant toutes les valeurs situées dans les intervalles des intervalles flous qui ont au moins ce degré de possibilité α (ces intervalles sont appelés des α-coupes ; Figure 17b).
4. Les valeurs *Inf* et *Sup* obtenues définissent les limites de l'intervalle de valeurs de $R(p_1, \ldots, p_n, F_1, \ldots, F_m)$ pour le degré de possibilité α.
5. Retourner à l'étape 2 et répéter les étapes 3 et 4 pour une autre valeur de possibilité (dans la pratique, on augmente α de 0 à 1 par pas de 0.1). La distribution de possibilité de $R(p_1, \ldots, p_n, F_1, \ldots, F_m)$ est obtenue à partir des valeurs *Inf* et *Sup* obtenues pour chaque valeur α.
6. Retourner à l'étape 1 pour générer une autre réalisation des variables aléatoires. On obtient ainsi une famille de ω distributions de possibilité pour $R(p_1, \ldots, p_n, F_1, \ldots, F_m)$ (Figure 17c) (ω étant le nombre de réalisations des variables aléatoires).

Le résultat du calcul hybride s'appelle un *intervalle flou aléatoire* (Gil, 2001). Il s'agît d'un résultat parfaitement analogue à celui *d'intervalle aléatoire* que l'on obtiendrait si à la place d'intervalles flous, on utilisait de simples intervalles *min-max* pour représenter l'information imprécise.

La question se pose ensuite de comment synthétiser cette information dans l'objectif de comparer le résultat à un critère ou seuil de risque. C'est là qu'interviennent les fonctions de croyance de Dempster-Shäfer.

2.6. POST-TRAITEMENT DU RESULTAT HYBRIDE

Dans Guyonnet et al. (2003) il était proposé d'extraire un résultat flou final de la Figure 17c, en sélectionnant les valeurs *Inf* et *Sup* (pour chaque α-coupe) de telle manière qu'on ait 5% de chances d'être respectivement en dessous ou au dessus. Ce résultat flou final pouvait ensuite être comparé à un critère de tolérance. Mais Baudrit et al. (2004) ont montré que cette méthode revenait à fusionner des intervalles qui étaient en fait indépendants les uns des autres et se traduisait par un résultat exagérément majorant en terme d'étendue de l'incertitude.

Baudrit et al. (2004, 2005) ont montré comment les fonctions de croyance de Dempster-Shafer s'appliquaient naturellement au problème de la comparaison du résultat hybride à un critère de tolérance.

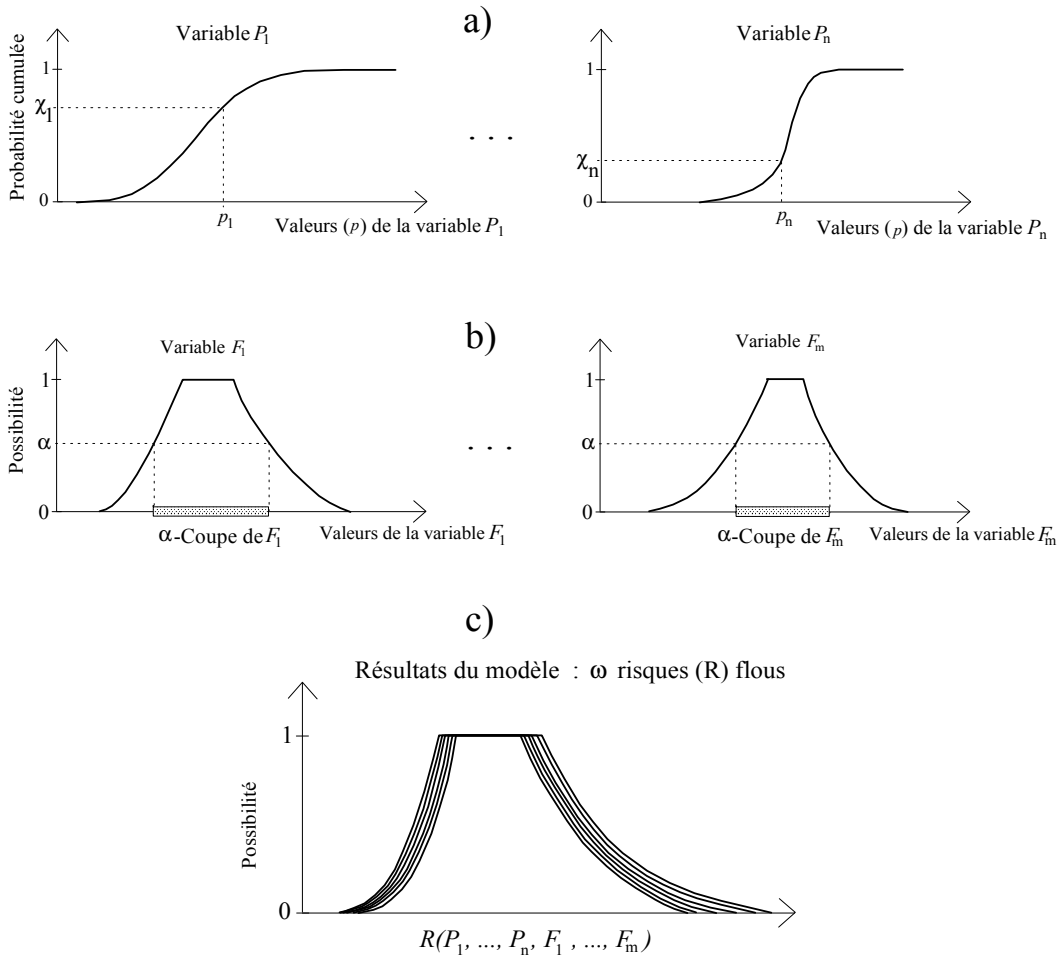


Figure 17 – Illustration schématique du calcul hybride

Considérons la proposition (notée A) dont on souhaite quantifier la probabilité :

A : le risque calculé *R* est inférieur au seuil *S*.

Cette probabilité va être qualifiée à l'aide des indicateurs que sont la crédibilité et la plausibilité (voir section 2.4.2). Il est rappelé que la crédibilité est la somme des probabilités de tous les éléments focaux qui impliquent forcément l'événement recherché tandis que la plausibilité est la somme des probabilités de tous éléments focaux qui ne contredisent pas nécessairement cet événement.

Tout d'abord il faut identifier les éléments focaux. Dans le cas du résultat du calcul hybride (Figure 17c), les éléments focaux sont les intervalles de valeurs obtenus par découpage de chaque résultat flou (voir la Figure 16 et la Figure 18).

Ensuite il faut définir les masses de probabilité à attribuer à chaque élément focal. Comme il a été illustré dans la Figure 16, pour un découpage en 10 tranches, la masse de probabilité attribuée à chaque intervalle est de 0.1. Mais dans le résultat hybride

(Figure 17c) ces intervalles sont également associés à une fréquence de réalisation qui est fonction du nombre d'itérations utilisé dans le processus Monte Carlo. Supposons que ce nombre soit 500. La fréquence de réalisation de chaque résultat flou est : 1/500. La masse de probabilité attribuée à chaque intervalle (élément focal) est donc : 1/500 * 0.1 = 1 / 5000 (à noter que 5000 est également le nombre d'intervalles générés lors du calcul hybride). On peut ensuite faire les sommations permettant de calculer les indicateurs Cr et Pl.

La Figure 18b présente deux (pour des raisons de clarté du graphique) familles d'éléments focaux (au lieu de dix pour le cas d'un découpage des intervalles flous en dix tranches) issues du calcul hybride

Figure 18a). On voit que les éléments focaux se chevauchent. Considérons un seuil arbitraire S. Tant que la valeur du seuil est inférieure à la plus petite borne inférieure des éléments focaux, on a : Pl = Cr = 0. Dès que le seuil dépasse la plus petite borne inférieure des éléments focaux on a : Pl = 1 / 5000 et Cr = 0. Au fur et à mesure que le Seuil dépasse une nouvelle borne inférieure d'élément focal, on ajoute les masses de probabilité (1 / 5000) à Pl. Pour que Cr devienne non nul, il faut que le seuil S soit supérieur à au moins une borne supérieure d'élément focal (l'événement A est complètement satisfait pour cet élément focal). Dès lors que le seuil est supérieur à la plus haute des bornes supérieure des éléments focaux, on a Pl = Cr = 1 (on est « sûrs » que le risque calculé est inférieur au seuil).

On notera au passage que si on suivait la même procédure pour le cas de la distribution de probabilité de la Figure 15, on obtiendrait deux distributions Pl et Cr qui seraient d'autant plus rapprochés que les éléments focaux sont nombreux et fins, jusqu'à ce qu'ils se confondent avec la fonction de probabilité cumulée classique (PDF) lorsque les éléments focaux tendent vers des points (singletons). Par ailleurs, l'application de la procédure au cas de la distribution de possibilité de la Figure 16, donnerait deux distributions Pl et Cr correspondant respectivement aux mesures de possibilité (Π) et de nécessité (N) de la théorie des possibilités de Dubois et Prade (1988).

Le calcul de Cr et Pl est réalisé de manière très simple dans HyRisk : le calcul hybride fournit les bornes inférieures et supérieures des intervalles sur le résultat du modèle *R*, pour chaque niveau de possibilité (0 à 1 par pas de 0.1). Il suffit de classer toutes les bornes inférieures en ordre croissant et de cumuler à chaque nouvelle valeur la masse de probabilité (1 / 5000) pour obtenir la distribution de plausibilité. Idem pour la crédibilité en classant cette fois toutes les bornes supérieures en ordre croissant.

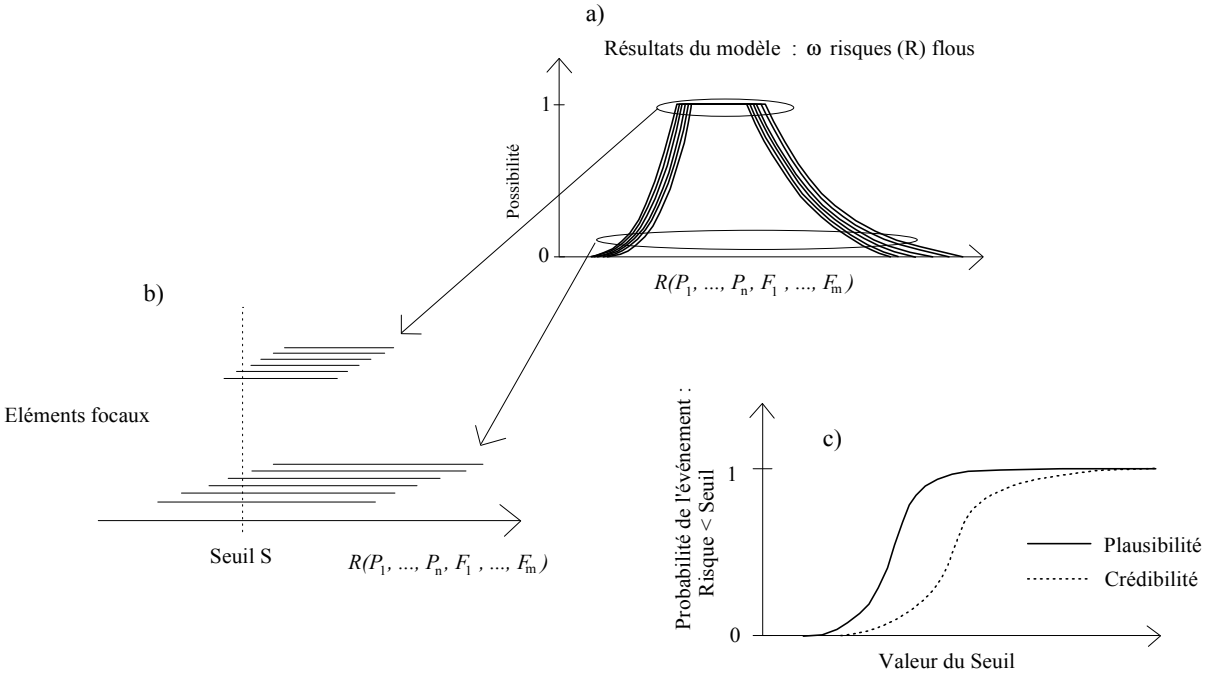


Figure 18 – Illustration du post-traitement du résultat hybride (a) : comparaison des éléments focaux et d'un seuil S (b) et indicateurs de la probabilité de l'événement : « Risque < Seuil » (c).

3. Le didacticiel HyRisk

3.1. PRESENTATION GENERALE

Le didacticiel HyRisk (téléchargeable sur http://www.brgm.fr/hyrisk) se présente sous la forme d'un classeur Microsoft Excel qui comprend 5 feuilles :

- La feuille 1 : fournit les principales instructions pour l'utilisation de HyRisk,
- La feuille 2 : permet d'entrer les paramètres du problème qu'on souhaite traiter,
- La feuille 3 : permet le calcul des distributions construites à partir des informations fournies dans la feuille 2,
- La feuille 4 : permet de lancer le calcul hybride et contient les résultats du calcul,
- La feuille 5 : présente les distributions de probabilité haute (Plausibilité) et basse (Crédibilité) pour la proposition : « le résultat du calcul est inférieur à une certaine valeur seuil ».

On peut par ailleurs spécifier une valeur de seuil spécifique pour obtenir les valeurs de crédibilité et de plausibilité pour la proposition « le résultat du modèle est inférieur à cette valeur de seuil spécifique ».

A noter qu'il est possible de sauver le classeur HyRisk.xls sous un autre nom correspondant à une application spécifique.

3.2. UTILISATION

On lance HyRisk comme n'importe quel document Excel. Lorsqu' apparaît le message concernant les macros, cliquer sur l'onglet « activer les macros ».

Pour définir un nouveau problème, aller à la feuille « Paramètres ».

Cliquer sur les onglets situés à gauche du triangle bleu (Figure 19) pour définir les nombres de variables probabilistes et possibilistes. Il est possible de définir jusqu'à 10 variables probabilistes et 10 variables possibilistes. Attention : si on souhaite qu'il n'y ait aucune variable d'un certain type alors que la feuille en comprend déjà en raison d'une utilisation antérieure, alors il faut entrer 0 pour le nombre de variables de ce type.

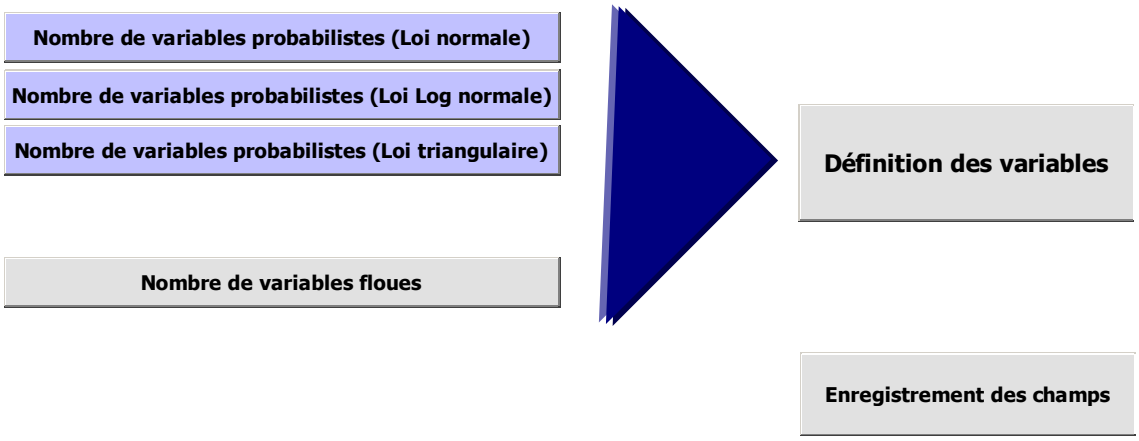


Figure 19 – Masque de saisie du nombre de variables et de leur type

Dès qu'on a entré les nombres de variables correspondant à chaque type, cliquer sur l'onglet « Définition des variables » : cela fera apparaître les champs permettant de décrire chaque variable (Figure 20 et Figure 22). Les indications dans la colonne « Descriptif » peuvent être modifiées pour entrer des informations utiles pour l'utilisateur (nom de variable, etc.). Dans l'exemple générique de la Figure 20, on a décrit 4 variables (notées A, B, C et D), dont deux sont des distributions normales de probabilité (A et C) tandis que les deux autres (B et D) sont des distributions de possibilité.

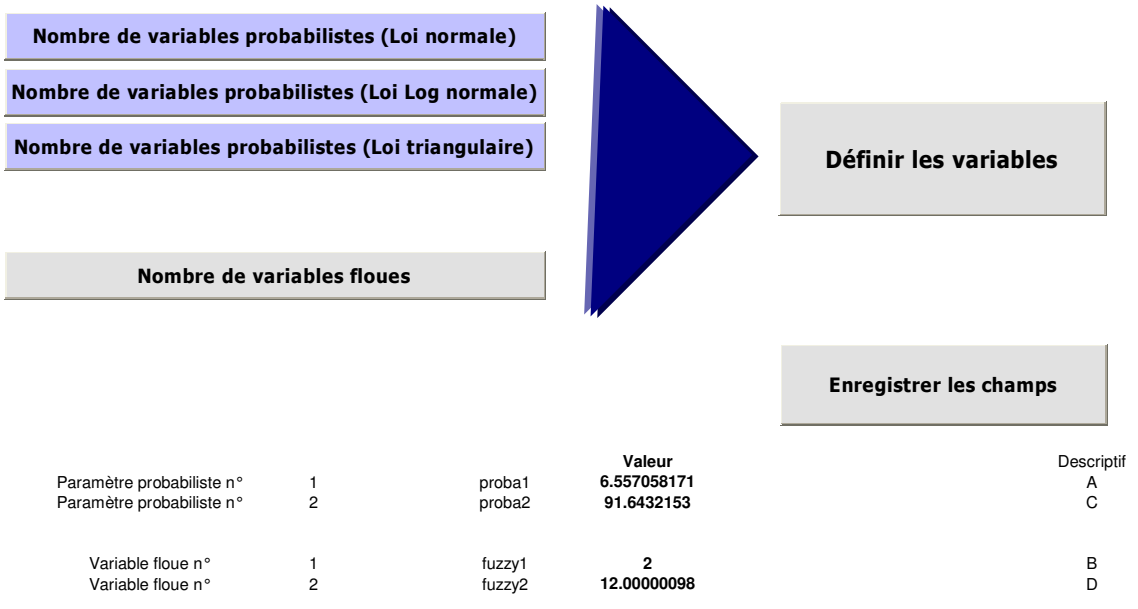


Figure 20 – Apparition automatique (en bas) de champs descripteurs

On va ensuite à droite de la feuille « Paramètres » vers les champs permettant de renseigner les variables (Figure 22) et on entre les valeurs souhaitées. Pour les distributions normales de probabilité, il faut fournir les moyennes et les écart-types, tandis que pour les distributions de possibilité, il faut fournir les limites du support (intervalle jugé « certain »), celles du noyau (intervalle jugé le plus vraisemblable) et les paramètres des exposants. Ces paramètres définissent les courbures des branches reliant les noyaux aux supports. Des valeurs de 1 (par défaut) définissent des droites. Des valeurs inférieures à 1 donnent des branches concaves (les valeurs situées en dehors du noyau sont jugées moins vraisemblables comparé au cas linéaire), tandis que des valeurs supérieures à 1 donnent des branches convexes (les valeurs situées en dehors du noyau sont jugées plus vraisemblables comparé au cas linéaire).

Une fois que ces valeurs ont été renseignées il faut cliquer sur l'onglet « Enregistrement des champs » (Figure 20).

Note : il arriver qu'apparaisse un message d'erreur (Figure 21). Ce dernier, dont l'apparition semble aléatoire, est sans conséquence. Cliquer sur OK et continuer.

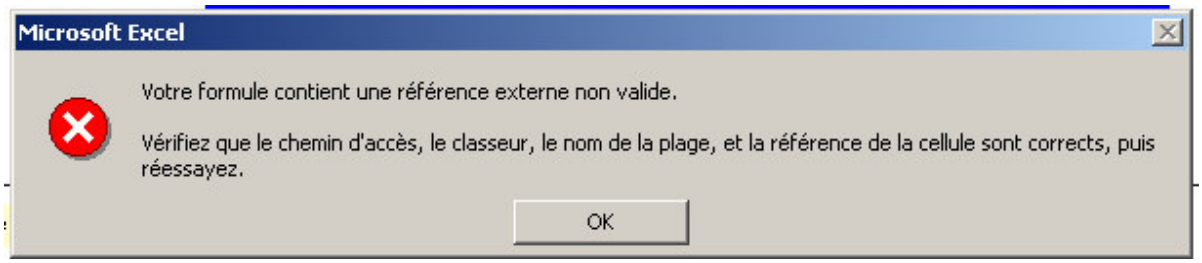


Figure 21 – Message d'erreur pouvant parfois apparaître : cliquer sur OK

On définit ensuite le « Modèle » qui est une combinaison des variables définies précédemment. Ce modèle est défini en entrant une formule dans la case du Modèle (Figure 22). On tape = et on va chercher les variables dans la colonne « Valeur » de la Figure 20. Les variables ont des noms standardisés qu'il n'est pas possible de changer (d'où l'intérêt d'insérer une description des variables dans la colonne « Descriptif »).

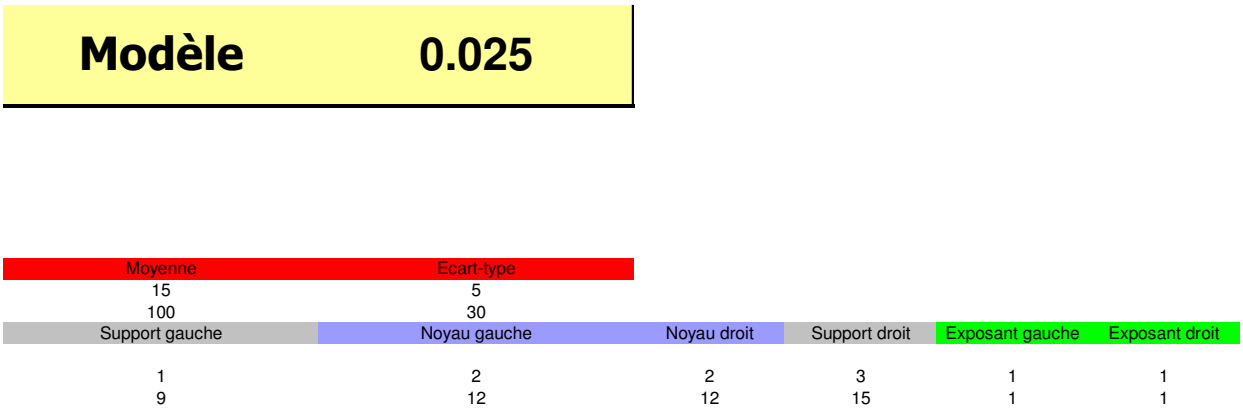


Figure 22 – Masque de saisie des valeurs décrivant les variables

Pour l'exemple de la Figure 22, le modèle a été défini par :
= Valprobaparam1*Valvarflou1/(Valprobaparam2*Valvarflou2)
c'est à dire (en reprenant les descriptifs des variables) : A * B / (C * D)

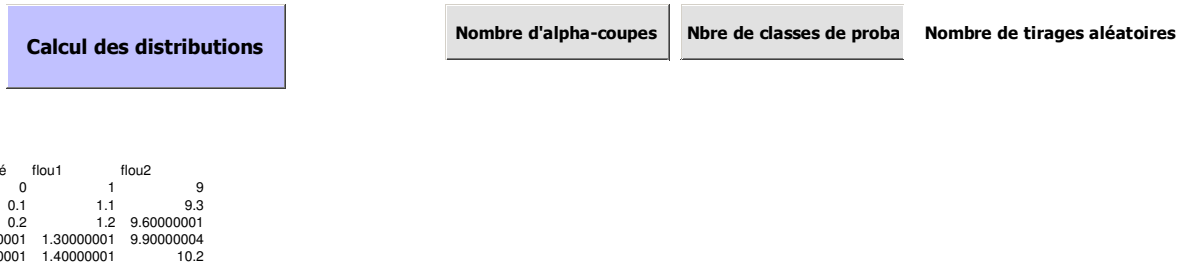
Note : si le résultat de « Modèle » est faible (inférieure à 10^{-5} environ), l'algorithme de recherche automatique des minimas et maximas d'Excel peut parfois rencontrer des difficultés. Il est conseillé d'inclure un facteur multiplicatif (par exemple 10^5) dans le « Modèle », ce qui permet d'avoir des valeurs proches de 1 et un calcul satisfaisant. Ce facteur est ensuite retranché des résultats du calcul (dans la feuille « Calculs et Résultats ») pour retrouver les valeurs correctes du « Modèle ».

On passe ensuite à la feuille « Calcul_distributions » qui permet de quantifier les distributions (Figure 23).

On clique sur l'onglet « Nombre d'alpha-coupes » pour entrer le nombre de valeurs discrètes de possibilité qui vont définir les intervalles flous. Un nombre égal à 10 permet en général de représenter ces nombres de manière suffisamment détaillée (dans le cas d'intervalles flous à branches très courbées on pourra choisir un nombre supérieur à 10 afin de bien reproduire les courbures).

Ensuite on passe à l'onglet « Nbre de classes de proba » qui permet de discrétiser les distributions probabilité. Ce nombre de points (typiquement situé entre 30 et 50) est sélectionné de manière à bien reproduire la forme des distributions. A noter qu'un nombre supérieur à 50 aura tendance à considérablement ralentir les calculs.

Enfin on précise le nombre d'itérations (tirages aléatoires) à utiliser dans le processus hybride (par exemple 100). A noter qu'en raison de la durée du calcul, il est préférable de commencer avec un nombre relativement restreint et éventuellement de l'augmenter si les distributions de plausibilité et de crédibilité obtenues ne sont pas suffisamment bien définies (trop hachées).



Possibilité	flou1	flou2
0	1	9
0.1	1.1	9.3
0.2	1.2	9.60000001
0.30000001	1.30000001	9.90000004
0.40000001	1.40000001	10.2

Figure 23 – Masque de saisie pour le calcul des distributions

On clique ensuite sur l'onglet « Calcul des distributions ». On peut visualiser les distributions dans la feuille « Paramètres » (Figure 24).

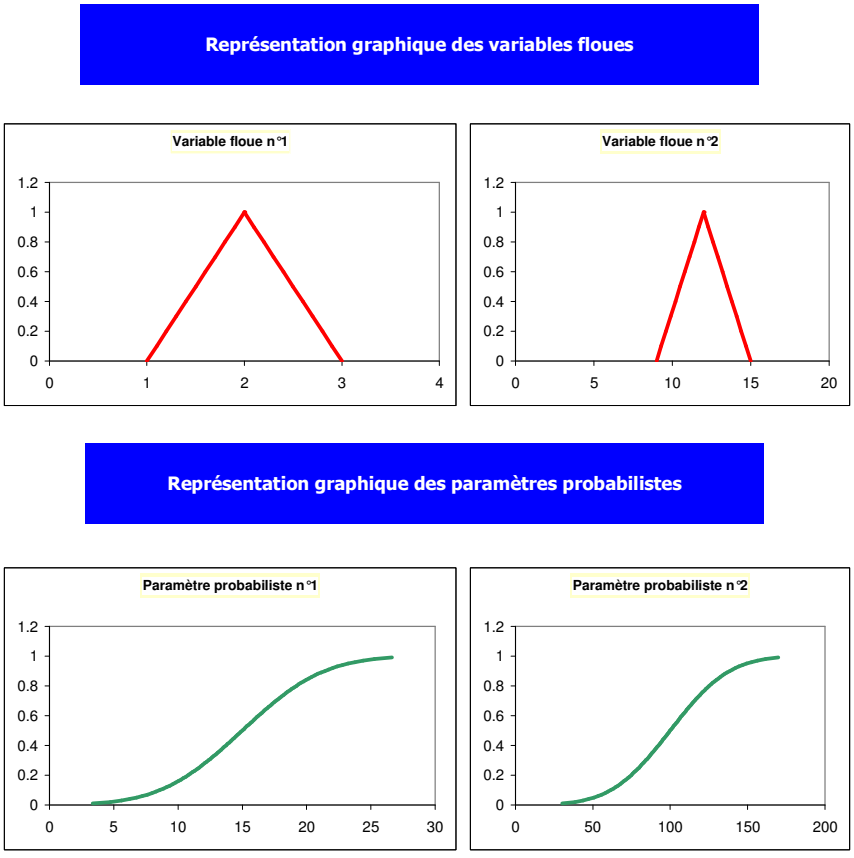


Figure 24 – Représentation graphique des distributions dans la feuille "Paramètres"

Tout est prêt pour le calcul hybride proprement dit ; on se rend donc dans la feuille « Calculs et résultats » (Figure 25). Avant de lancer le calcul, on peut préciser une

valeur de seuil pour laquelle HyRisk fournira les valeurs spécifiques de plausibilité (probabilité haute) et de crédibilité (probabilité basse) pour la proposition : « le résultat du modèle est inférieur au seuil ». On lance le calcul en appuyant sur l'onglet « Calcul ». Lorsque le calcul est terminé, on clique sur l'onglet « Graphique » pour visualiser les distributions de Plausibilité et de Crédibilité (Figure 26)

Calcul **Valeur seuil = 0.05** **Graphique**

Enveloppe externe	-0.001082002	-0.00121449	-0.001352502	-0.00149639	-0.00164652	-0.001803336	-0.00196728	-0.00213884	-0.00231857	-0.00250708	-0.002705
Enveloppe moyenne	0.011463648	0.01286736	0.01432956	0.01585398	0.01744468	0.01910608	0.020843	0.0226607	0.02456496	0.02656211	0.0286591
Enveloppe interne	0.024009298	0.02694921	0.030011622	0.03320435	0.03653589	0.040015496	0.04365327	0.04746024	0.0514485	0.0556313	0.06002321
Tirages\coupes	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
1	0.059872089	0.06720337	0.074840111	0.08280183	0.0911097	0.099786815	0.10885834	0.1183518	0.12829733	0.13872801	0.14968013
2	0.038507933	0.04322319	0.048134916	0.05325565	0.05859903	0.064179888	0.07001442	0.07612033	0.082517	0.0892257	0.09626978
3	0.034684695	0.0389318	0.043355869	0.0479682	0.05278106	0.057807825	0.06306308	0.06856277	0.07432435	0.08036697	0.08671169
4	0.024538012	0.02754267	0.030672515	0.03393555	0.03734045	0.040896687	0.04461457	0.04850537	0.05258145	0.05685637	0.06134499
5	0.021354443	0.02396927	0.026693054	0.02953274	0.03249589	0.035590739	0.03882626	0.04221227	0.04575952	0.04947981	0.05338608

Figure 25 –Feuille "Calcul et résultats"

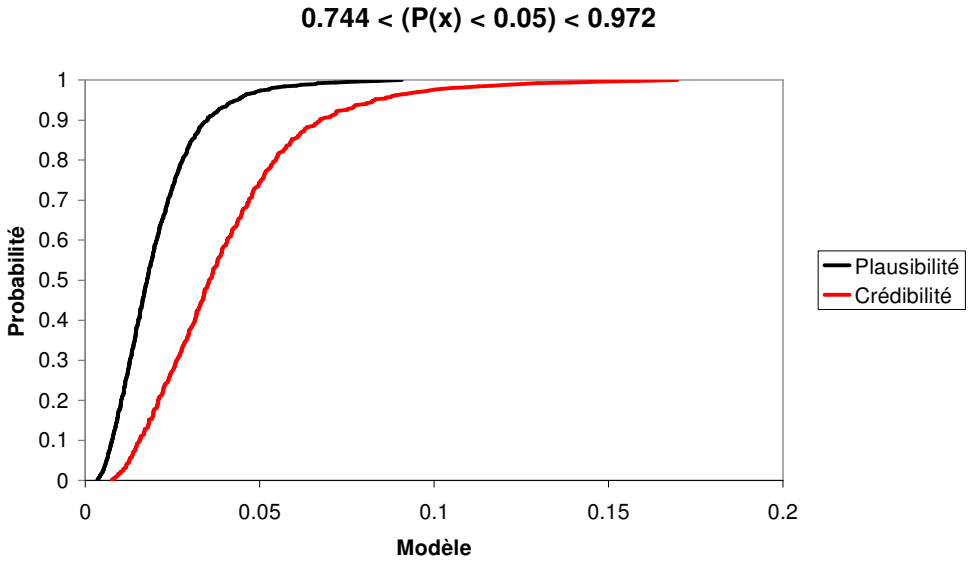


Figure 26 – Résultat du calcul hybride : Distributions de plausibilité et de crédibilité pour la proposition : "résultat du modèle < un certain seuil" et valeurs de Cr et Pl pour le cas "Seuil = 0.05"

3.3. EXEMPLE D'APPLICATION

3.3.1. Introduction

Pour cet exemple, on considère un cas simple d'exposition de personnes à un solvant organochloré (du 1,1,2 trichloroéthane), par le biais de la consommation d'eau polluée.

Pour une personne exposée, on estime un « excès de risque individuel » (ERI) qui est fonction de la dose absorbée par la personne et d'une valeur toxicologique de référence (VTR), qui est déterminée par les toxicologues et qui quantifie la relation dose-réponse pour la substance considérée.

Pour le cas du 1,1,2 trichloroéthane, les toxicologues considèrent qu'il s'agît d'une substance cancérogène dite « sans seuil » : l'exposition entraîne un risque quelque soit le degré d'exposition. La VTR pour cette substance est un excès de risque unitaire (ERU), c'est à dire une probabilité de cancer excédentaire par unité de dose journalière d'exposition.

Une dose d'exposition « vie-entière » (moyennée sur la durée de vie) est calculée à partir de :

$$D = \frac{I \cdot C \cdot FE \cdot DE}{PC \cdot DV} \qquad (1)$$

où :
D est la dose d'exposition (mg de polluant absorbé, par unité de poids corporel et par jour),
I est la quantité journalière d'eau ingérée (l/j),
C est la concentration dans l'eau ingérée (mg/l),
FE est la fréquence d'exposition (j/an),
DE est la durée d'exposition (an),
PC est le poids corporel (kg),
DV est « l'espérance de vie » (j)

L'excès de risque individuel (ERI) s'obtient de :

$$ERI = D \cdot ERU \qquad (2)$$

L'excès de risque individuel peut ensuite être comparé à un seuil de risque jugé tolérable.

3.3.2. Valeurs des paramètres

La représentation des valeurs des paramètres du problème fait appel aux deux modes de représentation décrites précédemment : les distributions de probabilité et de possibilité.

On suppose qu'on dispose d'un nombre significatif de mesures de la concentration (C) dans l'eau de boisson. Ces mesures permettent d'identifier une certaine variabilité aléatoire pour ce qui concerne cette concentration qui peut être décrite par une fonction de densité de probabilité triangulaire de mode 20 µg/l et de bornes inférieures et supérieures respectivement égales à 10 et 40 µg/l.

La durée d'exposition (DE) est également représentée par une distribution de probabilité : on supposera que des données statistiques existent concernant les temps de résidence de la population exposée. La durée d'exposition est représentée par une fonction de densité de probabilité triangulaire de mode 30 ans et de bornes inférieures et supérieures respectivement égales à 10 et 50 ans.

Le poids corporel et l'espérance de vie sont considérés comme constants (respectivement 70 kg et 70 ans) afin de conserver un caractère générique pour ce qui concerne les caractéristiques physiques de la cible, mais aussi par cohérence avec la VTR qui est déduite de données toxicologiques « vie entière ».

Tous les autres paramètres (ingestion, fréquence d'exposition, excès de risque unitaire) sont représentés par des distributions de possibilité triangulaires dont les supports et noyaux sont présentés dans le Tableau 1.

Paramètre	Unité	Mode de représentation	Borne inf	Mode ou noyau	Borne sup
Concentration dans l'eau	mg/l	Proba	0.005	0.01	0.02
Ingestion	l/j	Flou	1	1.5	2.5
Fréquence d'exposition	j/an	Flou	200	250	350
Durée d'exposition	an	Proba	10	30	50
ERU	$(mg/kg/j)^{-1}$	Flou	2×10^{-2}	5.7×10^{-2} (*)	0.1

(*) : (InVS, 2004)

Tableau 1 – Valeurs des paramètres utilisés pour l'exemple

On s'intéresse à la probabilité que l'excès de risque individuel soit inférieur à une valeur jugée tolérable par l'autorité sanitaire de 10^{-5} (probabilité d'apparition de cancer excédentaire chez une personne exposée).

3.3.3. Calcul hybride et résultats

Pour le calcul hybride, les intervalles flous ont été discrétisés en 10 alpha-coupes, les distributions de probabilité en 50 classes, et 100 itérations ont été utilisées pour l'échantillonnage Monte Carlo.

Le résultat du calcul est présenté en Figure 27. On note que pour la proposition « l'excès de risque individuel est inférieur à 10^{-5} », on obtient une probabilité comprise entre 0.7 (la valeur de crédibilité) et 1 (la valeur de plausibilité). L'écart entre ces deux valeurs est la conséquence de la nature imprécise de notre connaissance relative à certains facteurs de risque.

Se pose ensuite la question de l'acceptabilité de la comparaison entre le risque calculé et l'objectif de risque (10^{-5}). Une alternative consiste à utiliser la probabilité basse (la crédibilité) qui est la plus contraignante et donc la plus sécuritaire. Dans de cas il appartiendrait à l'autorité sanitaire de décider quel niveau de crédibilité il faut atteindre pour vraiment accepter le niveau de risque calculé. Une valeur de crédibilité de 70%, comme dans le cas de la Figure 27 pourrait paraître un peu élevée. Mais exiger une crédibilité de 100% serait une application trop stricte du principe de précaution.

On pourrait également imaginer élaborer un indicateur unique qui serait une combinaison (moyenne pondérée ?) de Pl et Cr (les décideurs préférant les indicateurs uniques). Mais il s'agît là de perspectives en cours de réflexion

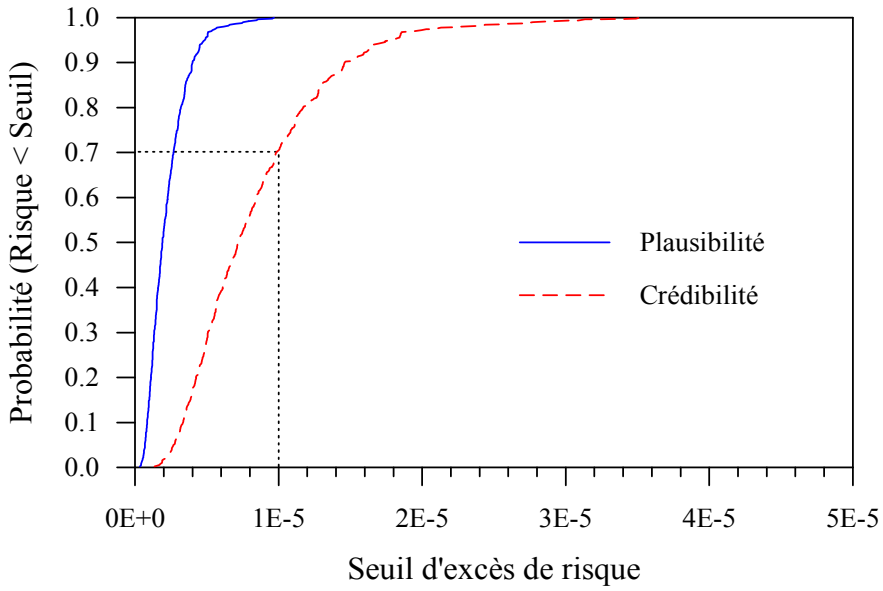


Figure 27 – Résultat du calcul hybride : distributions de plausibilité et de crédibilité pour la proposition « l'excès de risque individuel calculé est inférieur à une valeur seuil »

A noter que l'outil HyRisk permet de faire du calcul Monte Carlo classique ; il suffit pour cela de préciser un nombre nul pour les distributions de possibilité et définir les distributions de probabilité. A tirre d'exemple, la Figure 28 compare les résultats du calcul Monte Carlo obtenu en considérant que les intervalles flous du Tableau 1sont en fait des fonctions de densité de probabilité triangulaires. On constate que la distribution résultant de ce calcul se situe entre les distributions de plausibilité et de crédibilité calculées précédemment. Selon ce calcul Monte Carlo, la probabilité d'être en dessous

du seuil de 10^{-5} est égale à 93%, ce qui peut paraître suffisant pour accepter le risque. Mais cette valeur élevée est entièrement liée au fait qu'on a supposé des distributions de probabilité uniques en lieu et place des distributions de possibilité (qui, nous le rappelons, représentent des familles de dsitributions de probabilité). Encore une fois, le problème est d'arriver à justifier, au vu des informations disponibles, l'utilisation de dsitributions de probabilité uniques.

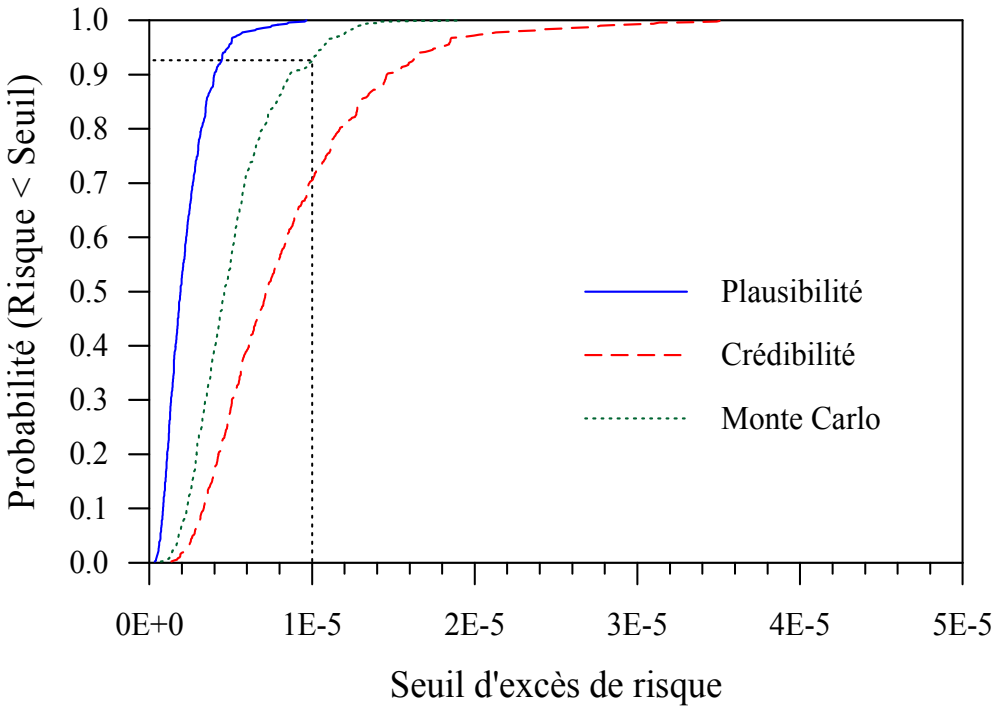


Figure 28 – Comparaison du calcul Monte Carlo classique avec celui en probabilités imprécises

4. Conclusions et perspectives

La méthode hybride présentée dans ce manuel est une alternative parmi d'autres visant à éviter que des choix a priori, non justifiés par l'information réellement disponible, n'aient une influence excessive sur les résultats des évaluations ainsi que sur le degré de confiance qu'on accorde à ces résultats.

Dans le cadre de son travail de thèse (en cours à la date de rédaction de ce rapport), Cédric Baudrit de l'Université Paul Sabatier de Toulouse explore d'autres alternatives et notamment certaines qui ont trait aux corrélations pouvant exister entre les variables du modèle. Dans la méthode hybride, telle qu'elle est mise en œuvre dans HyRisk, on fait une hypothèse d'indépendance entre les différentes variables probabilistes ainsi qu'entre le groupe de variables probabilistes et possibilistes. Par contre, dans le calcul d'intervalle flou, la même valeur de α sert à définir les α-coupes de chaque variable possibiliste, ce qui revient à faire une hypothèse de dépendance sur le degré de fiabilité des différentes α-coupes. Baudrit et al. (2004), proposent notamment de traiter à la fois les variables probabilistes et possibilistes directement dans le cadre commun des fonctions de croyance de Dempster-Shafer, en considérant toutes les corrélations possibles entre les variables. Une prochaine version du didacticiel HyRisk fera l'hypothèse d'une indépendance entre toutes les variables, qu'elles soient probabilistes ou possibilistes : le tirage aléatoire portera à la fois sur les distributions de probabilité et sur les intervalles des distributions de possibilité.

Une autre perspective de développement concerne l'utilisation de familles de distributions de probabilité définies à partir de moyennes et d'écart-types imprécisément connues. En effet, on peut imaginer que même face à une information incomplète/imprécise, un expert sache (de par son expérience) que si des mesures relatives à une variable pouvaient être effectuées en nombre suffisant, alors elles montreraient telle ou telle forme de variabilité (normale, log-normale, etc.). Pour le cas d'une distribution normale par exemple, les méthodes de simulation Monte Carlo 2D consistent alors à supposer des distributions uniformes pour la moyenne et l'écart-type puis à échantillonner ces distributions de manière aléatoire pour générer une distribution de probabilité sur la variable qui est échantillonnée dans un deuxième temps. L'objection que l'on peut émettre par rapport à cette approche a trait au choix arbitraire de distributions uniformes pour la moyenne et l'écart-type. Il paraît plus cohérent (plus conforme à l'information que détient l'expert) d'aborder l'incertitude relative à ces paramètres à l'aide d'intervalles. On définirait ainsi des familles de distributions de probabilité qui peuvent ensuite être exploitées en appliquant une approche analogue à celle décrite ici.

Enfin, si les méthodes abordées dans ce manuel peuvent présenter un intérêt sur le plan scientifique et méthodologique, la question se pose d'une exploitation dans un contexte de communication sur les risques et les incertitudes associées, vis à vis des différentes parties prenantes de la problématique « risque » (industriels, riverains, élus locaux, administration, …). Cet aspect fait actuellement l'objet d'une recherche dans le

cadre du projet (CREPS ; Cartographie du Risque – Exposition et Perception Sociale), réalisé avec le soutien de l'AFSSE (Agence Française de Sécurité Sanitaire Environnementale), en collaboration notamment avec des spécialistes des sciences sociales (voir Harpet et al., 2005).

Remerciements

Nous remercions M. Franck Marot de l'ADEME pour son soutien au projet IREA (Incertitudes – Risques d'Exposition – Ademe).

Bibliographie

Baudrit, C., Guyonnet, D., Dubois, D. (2005) – Post-processing the hybrid approach for addressing uncertainty in risk assessments. Journal of Environmental Engineering (sous presse).

Baudrit, C., Dubois, D., Guyonnet, D., Fargier, H. (2004) – Joint treatment of imprecision and randomness in uncertainty propagation. Dans : Information Processing and Management of Uncertainty in Knowledge-based Systems, Perugia, Italie, 4-9 juillet 2004.

Boole, G. (1854). An investigation of the laws of thought, on which are founded the mathematical theories of logic and probability. Walton and Meberly, London.

Casti, J. (1990) – Searching for certainty. William Morrow (Ed.), New York.

Chilès, J.-P., and Delfiner, P. (1999). – Geostatistics : Modeling Spatial Uncertainty. Wiley, New York.

Connover, W. Iman, R. (1982) - A distribution-free approach to inducing rank correlation among input variables. Technometric, 3, 311-334.

Dubois, D., Prade, H. (1992) - When upper probabilities are possibility measures. Fuzzy Sets and Systems, 49, 95-74.

Dubois, D., et Prade, H. (1988) - Possibility theory. New York Plenum Press, 263 pp.

Ferson, S., Ginzburg, L. (1996). Different methods are needed to propagate ignorance and variability. Reliability Engineering and System Safety, 54, 133-144.

Fisher, R.A. (1973) – Statistical methods of scientific inference. Hafner Press, New York.

Gil M.A. (2001). “Fuzzy Random Variables.” Special issue of *Information Sciences*, 133, nos. 1-2

Guyonnet, D., Dubois, D., Bourgine, B., Fargier, H., Côme, B., Chilès, J.-P. (2003a) – Hybrid method for addressing uncertainty in risk assessments. Journal of Environmental Engineering, 129, 68-78.

Gzyl, H. (1995). The Method of Maximum Entropy. Dans : Series on Advances in Mathematics for Applied Sciences, F. Bellomo and N. Brezzi (Eds.), Vol. 29, World Scientific Publishing Co.

Harpet, C., Guyonnet, D., Dor, F. (2005) - Risk perception and communication on risks: a field survey. Conférence ConSoil 2005, 3-7 Octobre 2005, Bordeaux, France.

InVS (2004) – Le stockage des déchets et la santé publique. Institut national de la Veille Sanitaire, version 3, Septembre 2003.

Levine, R ., Tribus, M. (1978) – The maximum entropy formalism. MIT Press, Cambridge.

MEDD (2000) - Guide pour la gestion des sites et sols pollués – l'évaluation détaillée des risques. Site www.fasp.info.

Oreskes, N., Shrader-Frechette, Belitz,K. (1994) – Verification, validation, and confirmation of numerical models in the earth sciences. Science, Vol. 263, pp.641-646.

Saporta, G., (1990) - Probabilités, Analyse des Données et Statistique – Editions Technip 1990.

Shafer, G. (1976). A mathematical theory of evidence. Princeton University Press.

Vose, D., (1996) - Quantitative risk analysis - A guide to Monte-Carlo simulation modelling. Wiley, New York.

Zadeh, L.. (1978) - Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems, 1, 3-28.

Centre scientifique et technique
Service EPI/DES
3, avenue Claude-Guillemin
BP 6009 – 45060 Orléans Cedex 2 – France – Tél. : 02 38 64 34 34