DOCUMENT PUBLIC

Programme GESSOL du MATE
Prise en compte de l'incertitude dans l'évaluation du risque d'exposition aux polluants du sol

Rapport d'avancement n° 3

Etude réalisée dans le cadre des opérations de Service public du BRGM 2001-POL-D04

Décembre 2001
BRGM/RP-51309-FR

BRGM
L'ENTREPRISE AU SERVICE DE LA TERRE

DOCUMENT PUBLIC

Programme GESSOL du MATE
Prise en compte de l'incertitude dans l'évaluation du risque d'exposition aux polluants du sol

Rapport d'avancement n° 3

Etude réalisée dans le cadre des opérations de Service public du BRGM 2001-POL-D04

D. Guyonnet, B. Côme, H. Fargier, D. Dubois

Décembre 2001
BRGM/RP-51309-FR

BRGM
L'ENTREPRISE AU SERVICE DE LA TERRE

Mots clés :

En bibliographie, ce rapport sera cité de la façon suivante :

Guyonnet, D., Côme, B., Fargier, H. Dubois, D. 2001 – Programme GESSOL du MATE. Prise en compte de l'incertitude dans l'évaluation du risque d'exposition aux polluants du sol. Rapport d'avancement no. 3. Rapport BRGM RP-51309-FR

Synthèse

Le 1er rapport d'avancement présentait une synthèse bibliographique sur les modèles d'absorption des métaux par les plantes, et une introduction au problème de la représentation de l'incertitude liée aux paramètres des modèles (approches probabiliste et possibiliste). Le 2ème rapport d'avancement proposait une méthodologie permettant de combiner deux types de représentations de l'incertitude (probabiliste et possibiliste) dans un même calcul du risque d'exposition de l'homme aux métaux du sol. Cette méthodologie était appliquée au cas d'un site industriel dont les sols sont pollués par du cadmium.

Tandis que le 2ème rapport d'avancement traitait le problème de l'évaluation du risque dans le sens direct (détermination du risque correspondant à un état de contamination du sol), le présent rapport d'avancement s'intéresse plus particulièrement au problème « inverse » : pour un risque jugé « tolérable », quelle concentration de polluant du sol peut être considérée admissible ? Ce rapport présente également, en annexe, une synthèse des différentes manières de combiner les modes de représentation de l'incertitude probabiliste et possibiliste dans le calcul du risque.

Sommaire

1. Introduction

Les objectifs du projet réalisé dans le cadre du Programme GESSOL du MATE (piloté par l'INRA) ont été présentés dans les rapports d'avancement 1 (Guyonnet, 2000) et 2 (Guyonnet et Bourgine, 2001). Dans ce deuxième rapport d'avancement, une méthodologie dite « hybride » était proposée pour combiner deux types de représentation de l'incertitude dans une même évaluation du risque d'exposition : une représentation dite probabiliste et une autre dite possibiliste. Cette méthodologie permet de représenter l'incertitude liée aux paramètres des modèles d'exposition d'une manière qui est plus cohérente avec la nature de l'information dont on dispose dans la pratique.

Dans le deuxième rapport d'avancement, cette méthodologie était appliquée au cas d'un site industriel réel dont les sols superficiels sont contaminés par du cadmium. Les résultats de cette application n'étaient que provisoires, notamment parce que la manière dont était traitée la question de l'absorption de cadmium par des légumes n'était pas satisfaisante. En effet, cette absorption était abordée par une équation de type « Facteur de Bioconcentration », qui suppose que la concentration dans la plante augmente de manière linéaire en fonction de la concentration dans le sol. Or cette représentation ne tient pas compte du fait que la capacité de la plante à absorber du métal est limitée, ce qui était d'ailleurs suggéré par les données mesurées présentées dans le rapport. Aussi une autre approche était proposée, mais pas traitée sur le plan numérique. Le traitement numérique complet sera documenté dans le rapport final de ce projet (à paraître en juin 2002), mais apparaît dans l'article (en anglais) présenté en annexe du présent rapport.

Tandis que le rapport précédent s'attachait à traiter le problème dans le sens direct (estimation d'un risque d'exposition en fonction d'une concentration mesurée dans le sol), on s'intéresse dans le présent rapport au problème dit « inverse » : pour un risque jugé tolérable, quelle peut être la concentration maximale dans le sol. Cette question présente un intérêt notamment pour la définition d'objectifs de qualité des sols (après traitement).

2. Traitement du problème « inverse »

2.1. RAPPEL DU MODELE D'EXPOSITION CONSIDERE

Le cas d'application présenté dans le 2ème rapport d'avancement concernait l'exposition de l'homme au Cd du sol par le biais de la consommation de légumes. L'équation permettant de calculer la dose absorbée est :

$$\text{Dose} = \frac{Q \cdot Cd_{pl} \cdot MS \cdot 1000}{BW} \qquad (1)$$

où :
Dose est la dose journalière de Cd absorbée par l'homme (µg/j kg^{-1}),
Q est la consommation journalière de légumes (kg légume / jour),
Cd_{pl} est la teneur en Cd dans le légume (mg Cd / kg sec de plante),
MS est la teneur en matière sèche de la plante (kg sec / kg total),
BW est le poids corporel de la cible humaine (kg).

La concentration de Cd dans le légume est lié à celle dans le sol par une équation de corrélation établie sur la base des données mesurées sur le site étudié (voir la Figure 1) :

$$Cd_{pl} = C_o + (Seuil - C_o) \cdot (1 - \exp[-kC_s]) \qquad (2)$$

où :
C_0 est la concentration à l'origine dans la plante (un « fond naturel ») (mg / kg sec de plante),
C_s est la concentration totale en Cd dans le sol (mg / kg sec de sol),
Seuil est la concentration dans la plante à l'asymptote (mg / kg sec de plante),
k est une constante contrôlant la courbure de la courbe de Cd_{pl} en fonction de C_s.

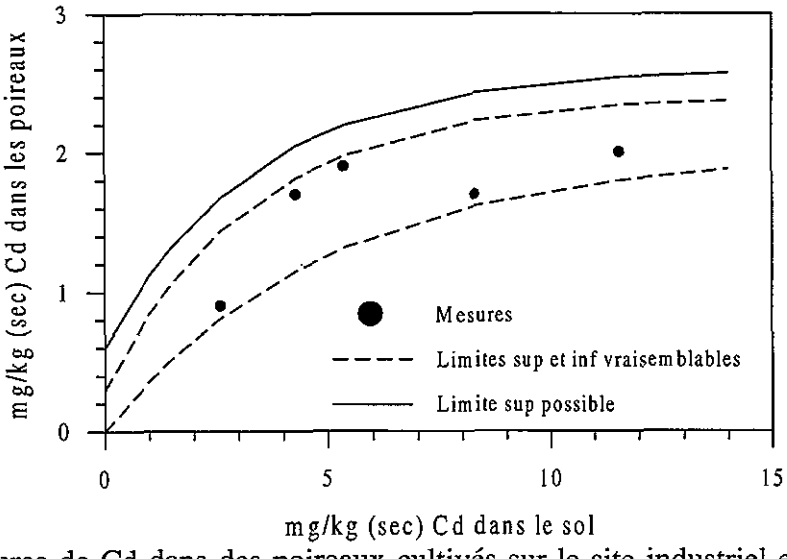


Fig. 1 – Mesures de Cd dans des poireaux cultivés sur le site industriel et équation de corrélation floue.

Dans le traitement direct d'estimation du risque en fonction de la teneur en Cd du sol (voir annexe 2), l'incertitude liée à la teneur en Cd du sol est traitée à l'aide d'une distribution de probabilité, tandis que celle liée aux paramètres C_o, *Seuil* et Q est traitée par des nombres flous. Les paramètres *MS* et *BW* sont supposés connus (nombres précis ou « crisp »).

2.2. EXPRESSION DU PROBLEME INVERSE

On s'intéresse ici à la question suivante :

« Comment définir la concentration du sol C_s telle que la dose calculée ne dépasse pas une dose jugée tolérable (notée D_o) ? »

Compte tenu de l'incertitude liée aux différents paramètres du modèle de calcul de la dose, cette question peut s'exprimer dans un cadre possibiliste de la manière suivante :

« Comment définir la concentration du sol C_s de manière à ce que la possibilité (notée Π ; voir les rapports no. 1 et 2) que la dose calculée dépasse la dose tolérable D_o ne soit pas supérieure à une certaine valeur ? ».

Cette formulation du problème est illustrée graphiquement dans la Figure 2 ci-dessous (voir Prade et Dubois, 1988). La dose calculée est un nombre flou représenté par sa fonction de répartition (μ). La possibilité de dépassement de la dose tolérable (D_o, considérée ici comme étant un nombre précis) est ici une valeur notée α.

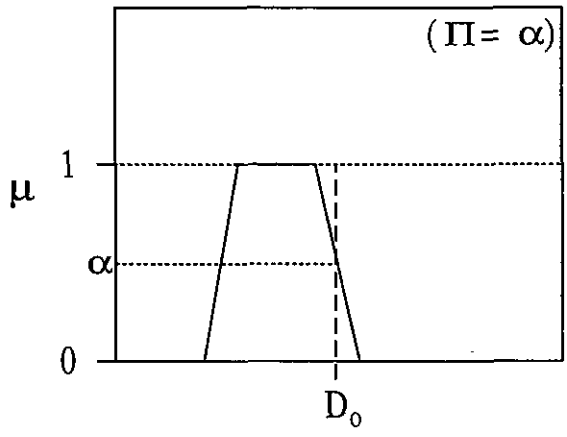


Fig. 2 – Comparaison possibiliste d'une dose calculée et d'un seuil D_o. La possibilité de dépassement du seuil est ici : $\Pi = \alpha$.

Il est important de noter qu'il appartiendra toujours à l'autorité sanitaire compétente de déterminer quel niveau de possibilité de dépassement est jugé acceptable (voir discussion dans le dernier chapitre).

2.3. APPLICATION AU CAS REEL CONSIDERE

2.3.1. Méthodologie générale

Le problème inverse peut être traité de manière générale à l'aide d'une approche itérative. L'approche hybride définie dans le rapport no. 2 (voir aussi en Annexe 1 du présent rapport) peut être mise en œuvre en faisant varier C_s jusqu'à ce que la possibilité de dépassement de la dose tolérable D_0 par la dose floue calculée corresponde à la valeur souhaitée. Mais cette approche peut être lourde en temps de calcul, aussi dans les cas simples est-il préférable d'adopter une approche spécifique

2.3.2. Méthodologie spécifique

Comme il a été dit précédemment, certains des paramètres entrant dans le calcul de la dose sont des nombres flous. Ils sont représentés dans la Figure 3.

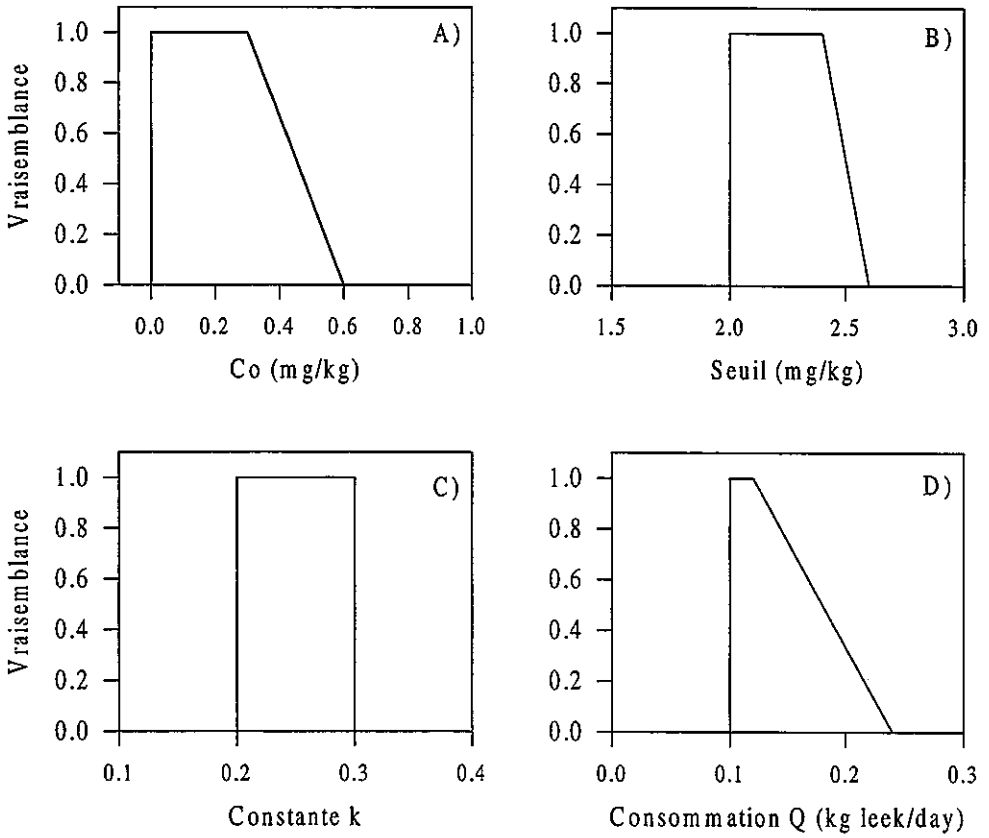


Fig. 3 – Nombres flous correspondant à certains des paramètres intervenant dans le calcul de la dose.

Les nombres flous des paramètres C_o, *Seuil* et k ont été déterminés sur la base de la Figure 1. Le paramètre Q (Consommation) a été déterminé sur la base d'INERIS (1999).

L'examen des Equations (1) et (2) montrent que la dose calculée atteint son maximum lorsque les paramètres C_o, *Seuil*, k et Q sont maximals. Ces valeurs maximals, en fonction du degré de vraisemblance α sont :

$C_{o\ max} = 0.6 - 0.3\ \alpha$
$Seuil_{\ max} = 2.6 - 0.2\ \alpha$
$k_{\ max} = 0.3$
$Q_{\ max} = 0.24 - 0.12\ \alpha$

A noter que pour un nombre flou de forme trapézoïdale, de support [a, b] et de noyau [c, d], la valeur maximale en fonction du degré de vraisemblance α est : d – (d – b) α.

La dose maximale est donc (3) :

$$Dose_{max} = MS \cdot 1000 \cdot (0.24 - 0.12\alpha) \cdot [0.6 - 0.3\alpha + (2 + 0.1\alpha)(1 - \exp[-0.3C_s])] / BW$$

A partir de l'équation (3), on peut exprimer *Cs* en fonction de α en fixant : Dose $_{max}$ = D_o, la limite jugée tolérable. On obtient :

$$C_s = \ln\left[1 + \frac{(0.6 - 0.3\alpha)}{(2 + 0.1\alpha)} - \frac{D_o}{(0.24 - 0.12\alpha)\lambda(2 + 0.1\alpha)}\right] / (-0.3)$$

où :
$\lambda = MS*1000/BW$

Pour une dose journalière maximale recommandée de 1 µg / j kg^{-1} (OMS, 1994), on obtient le graphique de la Figure 4.

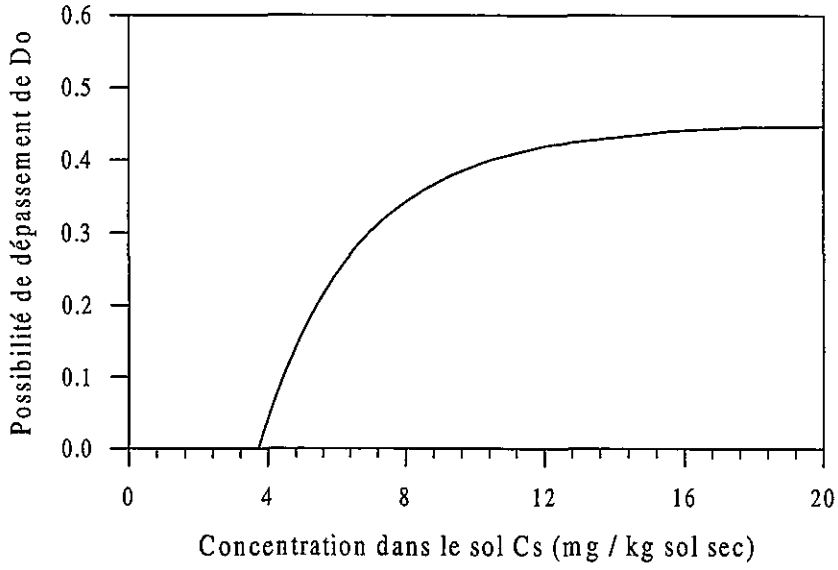


Fig. 4 – Concentrations en Cd dans le sol correspondant à différentes possibilités de dépassement de la dose jugée tolérable (1 µg / j kg^{-1}).

On note en Figure 4 qu'en dessous d'une concentration dans le sol d'environ 4 mg/kg, un dépassement de la dose tolérable est jugée impossible. Au fur et à mesure que la concentration dans le sol augmente, la possibilité de dépassement de la dose tolérable augmente également. Mais cette possibilité n'atteint jamais 1 en raison de l'effet de seuil pris en compte sur la base des données mesurées (Figure 1).

2.3.3. Discussion

La Figure 4 présente la possibilité de dépassement d'une dose jugée tolérable. On notera **qu'il appartient à l'Autorité Sanitaire compétente de définir la valeur réglementairement acceptable pour le niveau de possibilité d'excéder cette dose tolérable**.

Dans un cadre possibiliste, exiger une possibilité nulle d'excéder la dose tolérable aboutirait à des objectifs de dépollution extrêmement stricts, peut-être difficilement atteignables à un coût raisonnable. Vollmer et al. (1995) ont proposé une valeur de 0.3 pour établir un critère de concentration en Cd dans des sols superficiels. Comme l'approche adoptée par ces auteurs est très semblable à celle utilisée ici, elle est résumée ci-après.

Sur la base de considérations écotoxicologiques, Vollmer et al. (1995) établissent les propositions suivantes :

- Aucune nuisance aux plantes et animaux de ferme, ni à l'être humain, n'est constatée pour des concentrations en cadmium inférieures ou égales à 2 mg/kg de sol sec.

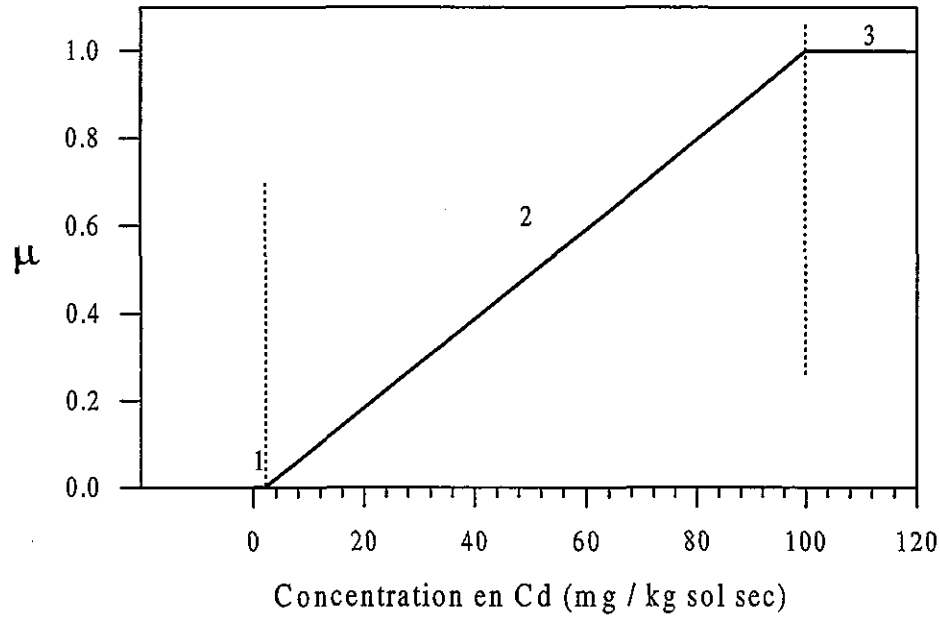
- Il y a quasi-certitude d'effets sur la santé des animaux d'élevage pour des concentrations supérieures ou égales à 150 mg/kg.

- Il y a quasi-certitude d'effets sur la santé humaine pour des concentrations de l'ordre de 100 mg/kg.

Une valeur de qualité du sol en cadmium, dite "d'intervention", protégeant le récepteur le plus sensible (l'homme) est donc située entre 2 et 100 mg/kg.

Vollmer et al. font l'hypothèse que la possibilité d'effets nuisibles croît linéairement entre la valeur de 2 mg/kg (possibilité nulle) et celle de 100 mg/kg (possibilité totale, fixée à 1).

Le critère de qualité du sol est alors représenté, en nombre flou, par la Figure 5.



1 : Zone d'acceptabilité totale
3 : Zone d'inacceptabilité totale
2 : Zone intermédiaire

Fig. 5 – Critère de qualité flou d'un sol (Vollmer et al., 1995).

L'ordonnée de cette figure permet de déterminer une valeur jugée pratiquement acceptable. Vollmer et al. (1995) indique que le niveau de possibilité d'effets nuisibles peut être fixé à 0,3. Ceci correspond à un seuil de teneur en Cd dans le sol de 30 mg/kg.

A noter qu'il s'agit d'un objectif générique, ne tenant pas compte des spécificités locales éventuelles.

Cette approche est elle-même dérivée d'un travail antérieur concernant la détermination, en nombres flous, de valeurs limites d'émissions de polluants atmosphériques dont le SO_2 (Pohl et al., 1995).

La démarche suivie comporte plusieurs étapes :

- Construire la fonction d'appartenance de la concentration en SO_2 dans la perspective de la santé humaine : pour cette dernière, il est admis qu'une concentration en SO_2 dans l'air inférieure ou au plus égale à 5 µg/m3 rend impossible les nuisances à la santé, et une valeur supérieure à 200 µg/m3 correspond à une quasi-certitude d'effets nuisibles.

- Prendre en compte de la même façon les aspects liés aux nuisances sur les plantes (ces dernières étant plus sensibles au SO_2 que les humains).

- A l'aide d'un "opérateur de pondération", combiner ces deux fonctions d'appartenance en une courbe unique, assurant le meilleur compromis entre ces deux aspects ; cet opérateur est choisi de manière à ce que la concentration limite actuelle réglementaire (en Suisse), soit 30 $\mu g/m^3$, soit obtenue pour une fonction d'appartenance (= possibilité d'effets indésirables) de 0,5 (il est reconnu que cette dernière valeur est **arbitraire**).

- Utiliser la fonction d'appartenance ainsi déterminée pour la concentration en SO_2 dans l'air pour calculer un "volume (d'air) critique flou", c'est-à-dire le volume d'air dans lequel il faut dissoudre 1 g de SO_2 pour garantir la concentration floue tolérable.

Dans ce travail, on utilise donc le fait que la fonction d'appartenance de la concentration dans l'air représente la mesure de possibilité qu'une valeur de concentration induise un effet sanitaire adverse.

3. Conclusion

Une application simple a montré comment le problème inverse pouvait être traité dans un cadre possibiliste. Parce que les équations en jeu (notamment les équations 1 et 2) sont simples, il a été possible de traiter le problème de manière spécifique, par résolution directe d'un système d'équations. Dans le cas où les équations en jeu seraient plus complexes, et qu'elles incluraient une ou plusieurs variables représentées par des fonctions de densité de probabilité, l'approche hybride définie dans le 2ème rapport d'avancement pourrait être appliquée de manière itérative : la concentration dans le sol peut être variée par pas successifs jusqu'à ce que la dose calculée floue dépasse la dose admissible pour la mesure de possibilité sélectionnée.

Le prochain rapport (le rapport final), synthétisera les différents rapports d'avancement réalisés dans le cadre de ce programme, et proposera des perspectives de recherche en matière de prise en compte de l'incertitude dans les évaluations des risques.

4. Références

Delgado, M. et S. Moral 1987 - On the concept of possibility-probability consistency. *Fuzzy Sets and Systems* 21, 311-318.

Dubois, D. et H. Prade 1982 - On Several Representations of an Uncertain Body of Evidence. *Fuzzy Information and Decision Processes* (M. M. Gupta, E. Sanchez editeurs),167-181.

Dubois, D. et H. Prade 1994 - Unfair Coins and Necessity Measures », *Fuzzy Sets and Systems* 10, 15-20.

Guyonnet, D. 2000 – Programme GESSOL du MATE. Prise en compte de l'incertitude dans l'évaluation du risque d'exposition aux polluants du sol. Rapport d'avancement no. 1. Rapport BRGM/RP-50347-FR

Guyonnet, D., Bourgine, B. 2001 - Programme GESSOL du MATE. Prise en compte de l'incertitude dans l'évaluation du risque d'exposition aux polluants du sol. Rapport d'avancement no. 2. Rapport BRGM/RP-50897-FR

INERIS, 1999 – Méthode de calcul des valeurs de constat d'impact dans les sols. INERIS, Verneuil-en-Halatte, France.

Klir, G. J. 1990 - A principle of uncertainty and information invariance. *International. Journal of General Systems*, 17(2-3), 249-275.

OMS (1994) – Directives de qualité pour l'eau de boisson. Volume 1 : Recommandations. 2ème édition. Organisation Mondiale de la Santé, Genève Suisse.

Pohl C., Ros M., Waldeck B., Dinkel F., 1995 - Fuzzy - Immissionsgrenzwerte. Eine Methodik zur unscharfen Modellierung von Immissionsgrenzwerten. Carbotech AG, Basel (CH).

Prade H., Dubois D., 1988 - Théorie des possibilités. Application à la représentation des connaissances en informatique. Collection "Méthode et Programmes", Masson Ed., Paris.

Shafer, G. 1976 - *A Mathematical Theory of Evidence*. Princeton University Press.

Vollmer M.K., Gupta S.K., Krebs R., 1995 - New standards on contaminated soils in Switzerland. Comparison with Dutch and German quality criteria. Proceedings of "Contaminated Soils", Third International Conference on the Biogeochemistry of Trace Elements, Paris, May 1995.

P. Smets et R. Kennes (1994) - « The transferable belief model », *Artificial Intelligence* 66 : 191-234.

Zadeh, L. A. 1978 - Fuzzy Sets as a Basis for a Theory of Possibility. *Fuzzy Sets and Systems* 1:3-28, 1978.

Annexe 1

Synthèse sur la combinaison des représentations probabiliste et possibiliste

A.1 INTRODUCTION

Ce chapitre est destiné à donner des pistes pour l'utilisation conjointe d'informations statistiques (probabilistes) et d'informations subjectives imprécises (possibilistes) pour mener des calculs d'analyse de risque lorsque les valeurs de parametres d'un modèle mathématique du phénomène étudié sont mal connues. L'aspect original de ce problème tient à cette presence conjointe de deux types d'incertitude, ce qui rend non seulement les calculs plus difficiles à mener, mais surtout, la méthodogie à adopter non triviale, voire non unique.

On s'intéresse plus particulièrement aux problèmes de la forme suivante : on calcule, à partir des données existantes mal connues, la valeur imprécise d'une certaine grandeur et on vérifie si cette valeur dépasse ou non un seuil fixé par des régulations (par exemple un seuil de pollution...). C'est le problème direct. On s'intéresse aussi au problème inverse, à savoir quel est le domaine de certains parametres du modèle qui assure qu'on ne dépassera pas les seuils fixés, lorsque la connaissance des autres paramètres est imprécise.

Pour permettre d'avancer sur cette question, qu'on peut appeler « propagation de l'incertain hétérogène», encore très peu abordée dans la littérature, il convient de faire le point sur les liens entre les principales théories de l'incertain : probabilités possibilités, probabilités imprécises, fonction de croyance. La représentation possibiliste utilise les fonctions d'appartenance d'intervalles flous, qu'on peut voir comme un empilement d'intervalles de confiance emboités. Il existe des méthodes de passage entre les représentations, dont les justifications commencent à être mieux comprises, notamment entre probabilité et possibilités, deux théories de l'incertain qui sont au confluent des autres représentations, et qui sont aussi les plus simples donc les plus applicables dans un premier temps. Les sous-chapitre A.2 et A.3 introduisent les principales théories de l'incertain et les méthodes de passages entre représentations.

On peut distinguer deux classes d'approches pour la propagation de l'incertain hétérogène (problème direct):

1) On exprime toutes les données mal connues dans un format unique et on se ramène au cas homogène.

2) On préserve la spécificité des incertitudes et on propage en deux étapes homogènes. Se pose la question de l'ordre dans lequel on procède (probabiliste puis possibiliste, ou l'inverse). Il semble que certains choix de cet ordre mènent à des procédures plus exploitables que d'autres choix.

La première approche est celle adoptée traditionnellement par les statisticiens : tout paramètre mal connu est représenté par une distribution de probabilité, même si les sources d'informations sont de fait hétérogènes (mesures, plus ou moins entâchées d'erreur, et opinions d'expert ; utilisation de probabilités uniformes ou Gaussiennes pour pallier l'absence de données). Les Bayésiens prétendent justifier cette approche sur

des bases « rationnelles » et en excluent toute autre. La technique de propagation la plus usuelle est la méthode par tirage aléatoire de Monte-Carlo. Une autre approche homogène existe pourtant : elle consiste à représenter toutes les données (numériques) par des intervalles, et à faire un calcul d'erreur. Cependant, le calcul d'erreurs ne repose pas sur les mêmes présupposés que la méthode de Monte Carlo. Dans cette dernière on fait souvent l'hypothèse d'indépendance entre les paramètres, ce qui crée des phénomènes de compensation des erreurs. En revanche, le calcul d'erreurs ne fait pas d'hypothèse d'indépendance (au sens stochastique), ni de dépendance, et s'avère être un calcul de pire et meilleur cas. C'est ce qu'on appelle souvent la non-interactivité entre les paramètres, l'idée étant qu'aucune configuration n'est exclue. La propagation possibiliste usuelle repose sur le calcul des intervalles flous, ce qui revient à effectuer le calcul d'erreurs sur les coupes de niveau des intervalles flous (qui sont des intervalles), et ce, pour tous les niveaux. Cette approche n'est pas beaucoup plus complexe que le calcul d'erreur usuel.

Quand on représente (via les transformations adéquates) toutes les données sous un format possibiliste, on a le choix entre les deux types d'hypothèses indépendance ou non-interactivité : cette dernière est admise dans le calcul d'intervalles flous usuel. Néanmoins il existe des calculs d'intervalles flous alternatifs (moins étudiés à ce jour, car plus complexes), dont un qui exploite l'idée d'indépendance (au sens de la compensation des erreurs). Si les données imprécises sont représentées par des fonctions de croyance, la méthode de propagation combinera calcul d'erreur et tirage aléatoire. Cette approche peut même simplifier la méthode de Monte Carlo usuelle dans certains cas, et fournir, à peu de frais, des encadrements des probabilités résultats. Ces questions sont abordées dans le sous-chapitre A.4. Le sous-chapitre A.5 discute de la comparaison d'un résultat incertain avec un seuil ou un objectif.

Pour ce qui est de la propagation hybride, on se restreint au cas où une partie des données est probabiliste, le reste étant possibiliste (représenté par des intervalles flous). Si on propage d'abord l'incertitude possibiliste, le résultat est un intervalle flou aléatoire (car dépendant des paramètres stochastiques). La question est alors d'extraire l'information pertinente et exploitable d'une statistique d'intervalles flous. Les variables aléatoires flous ont fait l'objet de nombreuses études (surtout théoriques) qui permettent de définir notamment une moyenne et une variance floues. Ces points sont abordés dans le sous-chapitre A.4.

Si on propage d'abord l'incertitude probabiliste, on obtient en théorie une distribution de probabilité sur la grandeur évaluée, indexée par les valeurs des paramètres possibilistes, ce qui semble très peu commode à représenter en pratique. On pourrait néanmoins faire un calcul d'intervalles aléatoires en effectuant pour chaque niveau de possibilité un calcul hybride d'intervalles et de Monte-Carlo. On obtiendrait alors une famille d'intervalles aléatoires indexé par des degrés de possibilité. Ce thème n'est pas abordé dans le rapport à ce stade, car il pose certainement des problèmes de complexité calculatoires et d'exploitabilité des résultats.

A.2. UN CADRE COMMUN POUR LES PROBABILITES ET LES POSSIBILITES

A.2.1. Mesures de Sugeno : définitions de base

On appelle capacité ou mesure de Sugeno (Sugeno, 1974) sur un référentiel Ω toute fonction telle que :

$\mu: 2^{\Omega} \rightarrow [0, 1]$

$A \subseteq B, \mu(A) \leq \mu(B)$ (monotonie dans l'inclusion)

Elle est dite normalisée si elle vérifie les propriétés :

$\mu(\varnothing) = 0 \qquad \mu(\Omega) = 1$

Ces propriétés sont minimales pour obtenir une fonction d'ensemble μ qui puisse permettre d'interpréter $\mu(A)$ comme l'évaluation d'un degré de confiance dans l'événement A.

Les mesures de probabilité, de possibilité, de nécessité, de plausibilité, de crédibilité sont des mesures de Sugeno. Elles sont respectivement définies par :

- Une capacité P est une mesure de probabilité ssi c'est une capacité additive c.-à-d. ssi : $\forall A, B \subseteq \Omega \; P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- Une capacité Π est une mesure de possibilité ssi :
  $\forall A, B \subseteq \Omega \; \Pi(A \cup B) = \max(\Pi(A), \Pi(B))$

- Une capacité N est une mesure de nécessité ssi :
  $\forall A, B \subseteq \Omega \; N(A \cap B) = \min(N(A), N(B))$

- Une capacité Pl est une mesure de plausibilité ssi elle est sous-additive à tout ordre n, c.-à-d. si,
  $\forall A_1, .., A_n \subseteq \Omega$,
  $Pl(A_1 \cup .. \cup A_n) \leq \sum_i Pl(A_i) - \sum_{i<j} Pl(A_i \cap A_j) + \ldots + (-1)^{n+1} Pl(A_1 \cap \cap A_n)$

- Une capacité Bel est une mesure de crédibilité (ou de croyance - « belief » en anglais) ssi elle est superadditive à tout ordre n, c.-à-d. si,
  $\forall A_1, .., A_n \subseteq \Omega$,
  $Bel(A_1 \cup .. \cup A_n) \geq \sum_i Bel(A_i) - \sum_{i<j} Bel(A_i \cap A_j) + \ldots + (-1)^{n+1} Bel(A_1 \cap \ldots \cap A_n)$

A l'ordre 2 par exemple, la sous-additivité et la super-additivité s'écrivent respectivement: $\mu(A \cup B) \leq \mu(A) + \mu(B) - \mu(A \cap B)$ et $\mu(A \cup B) \geq \mu(A) + \mu(B) - \mu(A \cap B)$. Il faut remarquer que la sous (resp. super)-additivité à l'ordre i n'implique pas la sous-(resp. super) additivité à l'ordre i + 1, contrairement à l'additivité.

La mesure duale d'une capacité μ est la capacité μ' telle que :

$$\forall A \subseteq \Omega, \mu'(A) = \mu(\Omega) - \mu(\Omega / A)$$

Possibilités et necessités, d'une part, et mesures de cédibilité et de plausibilité, d'autre part sont des fonctions duales l'une de l'autre : Si $\mu = \Pi$, alors $\mu' = N$, et si $\mu = Bel$, $\mu' = Pl$, et réciproquement.

Mesures de probabilité

Toute mesure de probabilité normalisée peut se définir à partir d'une distribution de probabilité $p : \Omega \rightarrow [0, 1]$ telle que $\sum_{\omega \in \Omega} p(\omega) = 1$ de la façon suivante :

$$P(A) = \sum_{\omega \in A} p(\omega)$$

p s'obtient évidement par : $\forall \omega \in \Omega, p(\omega) = P(\{\omega\})$

Les probabilités sont des mesures auto duales (la mesure duale d'une probabilité P est P elle même). On peut vérifier que toute probabilité est à la fois une plausibilité et une crédibilité.

Mesures de nécessité et de possibilité : théorie des possibilités

Toute mesure de possibilité (ou de nécessité) normalisée peut se définir, dans le cas fini, de façon équivalente à partir d'une distribution de possibilité $\pi : \Omega \rightarrow [0, 1]$ telle que $Max_{\omega \in \Omega} \pi(\omega) = 1$ de la façon suivante (Zadeh, 1978), (Dubois et Prade, 1987):

$$\Pi(A) = Max_{\omega \in A} \pi(\omega) \qquad N(A) = Min_{\omega \in \Omega / A} 1 - \pi(\omega)$$

π s'obtient évidement par : $\forall \omega \in \Omega, \pi(\omega) = \Pi(\{\omega\}) = 1 - N(\Omega \setminus \{\omega\})$.

Dans le cas infni ($\Omega = \mathbb{r}$), c'est faux. Il convient alors de partir de la distribution de possibilité π et de construire Π et N à l'aide des formules ci-dessus (remplacant max et min par sup et inf, respectivement). π est alors la fonction d'appartenance d'un ensemble flou de réels, usuellement d'un intervalle flou caractérisé par le fait que les coupes de niveau $\{\omega: \pi(\omega) \geq \alpha\}$ sont des intervalles fermés pour tout $\alpha \in (0, 1]$
La mesure duale d'une possibilité est en effet une nécessité et réciproquement:

$$\Pi(A) = 1 - N(\Omega / A) \quad N(A) = 1 - \Pi(\Omega / A)$$

En autres propriétés des possibilités et des nécessités, on peut également vérifier que :

- $\Pi(A) < 1 \Rightarrow N(A) = 0$,
- $A \cup B = \Omega \Rightarrow \max(\Pi(A), \Pi(B)) = 1, \min(N(A), N(B)) = 0$,
- toute mesure de possibilité est une mesure de plausibilité ,
- toute mesure de nécessité est une mesure de crédibilité.

Enfin, on dit qu'une distribution de possibilité π_1 (ou les mesures de possibilité et de nécessité correspondantes) est plus spécifique que π_2 si π_1 est incluse dans π_2 au sens des ensembles flous, c'est-à-dire si :

$$\forall\ \omega \in \Omega,\ \pi_1(\omega) \leq \pi_2(\omega)$$

Mesures de plausibilité et de crédibilité (Shafer, 1976 ; Smets, Kennes, 1994)

Les mesures de plausibilité et de crédibilité normalisée peut se définir de façon équivalente à partir d'une distribution de poids m : $2^{\Omega} \rightarrow [0, 1]$ telle que $\sum_{A \in 2\Omega} m(A) = 1$ de la façon suivante :

$$Pl(A) = \sum_{B,\ A \cap B \neq \varnothing} m(B) \qquad Bel(A) = \sum_{B,\ B \subseteq A} m(B)$$

m est appelée une fonction de masse et tout A tel que $m(A) > 0$ est un élément focal : m(A) est la part de la croyance totale attribuée à l'affirmation « $x \in A$ » exactement, et à aucune autre affirmation du type « $x \in B \subset A$ ». Autrement dit, m(A) est la probabilité pour que la seule information disponible est « $x \in A$ ». On note Foc(m) l'ensemble des élément focaux de m ($Foc(m) = \{F \subseteq \Omega, m(F) > 0\}$). La valeur $m(\Omega)$ est la probabilité qu'on ne sache rien (et vaut 1 quand on n'a pas d'information du tout). Dans ce cas $Pl(A) = 1$ et $Bel(A) = 0$ pour tout $A \neq \varnothing$, Ω. Entre autres propriétés des couples Bel, Pl on peut également vérifier que $Bel(A) \leq Pl(A)\ \forall\ A \subseteq \Omega$. Bel(A) évalue à quel point il est certain que l'information disponible, modélisée par la fonction m, implique « $x \in A$ ». Pl(A) évalue à quel point l'information disponible ne contredit pas l'affirmation « $x \in A$ ». Bel est plus exigeante que Pl : Bel (A) = 1 si « $x \in A$ » est certain. Pl (A) = 1 si « $x \in A$ » est seulement possible. En revanche, Pl (A) = 0 indique que « $x \in A$ » est absolument impossible.

La mesure duale d'une plausibilité est une crédibilité et réciproquement. En effet :

$$Pl(A) = 1 - Bel(\Omega / A) \qquad Bel(A) = 1 - Pl(\Omega / A)$$

La fonction de masse m s'obtient depuis Bel en appliquant la transformation dite de Moebius. Il s'agit tout simplement de la solution du système d'équations linéaires :

$$\{Bel(A) = \sum_{B,\ B \subseteq A} m(B),\ \forall\ A \subseteq \Omega\},$$

dont les inconnues sont les m(B). On peut prouver qu'il y a toujours une solution positive ($m(B) \geq 0,\ \forall\ B$), pourvu que Bel satisfasse bien les axiomes des crédibilités.

Lorsque les éléments focaux de m sont des singletons, Bel(A) = Pl(A) pour tout A : les deux mesures sont alors des probabilités.
Lorsque, à l'inverse, les élément focaux de m sont emboîtés, Bel est une mesure de nécessité et Pl une mesure de possibilité (éléments emboîtés signifie que, pour toute paire d'éléments focaux A_i , $A_j \in Foc(m)$, $A_i \subseteq A_j$ ou $A_j \subseteq A_i$, ou, en d'autres termes, que les k éléments focaux de m forment une chaîne $A_1 \subseteq A_2 \subseteq ... \subseteq A_{k-1} \subseteq A_k$).

Notons que les fonctions de croyance ont été étudiées essentiellement dans le cas fini. Une fonction de masse sur les réels affectera donc des poids à un nombre fini de sous-ensembles de r, en pratique des intervalles. La notion de fonction de répartition (cumulée) devient une paire de fonctions de répartition supérieures et inférieures F* et F_* définies par F*(x) = Pl((–∞, x]) et $F_*(x)$ = Bel((–∞, x]) respectivement. On peut définir des fonctions de croyance continues dans des cas particulier, en utilisant des densités de probabilité bidimensionnelles continues p(x, y) > 0 si et seulement si $x \leq y$. Alors p(x, y) est interprété comme une « densité » de masse m([x, y]), et les fonctions de crédibilité et de plausibilité se calculent à l'aide des intégrales appropriées

A.2.2 Familles de probabilités inférieures et supérieures

Soit $\boldsymbol{P}$ une famille de mesures de probabilité sur un référentiel Ω. Pour tout $A \subseteq \Omega$, on peut définir :

sa probabilité supérieure	:	$P^*(A) = \sup_{P \in \boldsymbol{P}} P(A)$
sa probabilité inférieure	:	$P_*(A) = \inf_{P \in \boldsymbol{P}} P(A)$

En d'autres termes la valeur de la probabilité P(A) est imprécise:

$$\forall P \in \boldsymbol{P}, P_*(A) \leq P(A) \leq P^*(A)$$

L'ensemble $\boldsymbol{P}^* = \{P, P_*(A) \leq P(A) \leq P^*(A)\}$, appelé « enveloppe probabiliste », induit par les bornes issues de la famille $\boldsymbol{P}$ contient $\boldsymbol{P}^*$ mais $\boldsymbol{P} \neq \boldsymbol{P}^*$ en général, puisque le calcul de P^* et P_* est une projection de $\boldsymbol{P}$. Par exemple si famille $\boldsymbol{P}$ est un ensemble fini, $\boldsymbol{P}^*$ est un sous-ensemble convexe de $[0, 1]^{card(\Omega)}$. C'est même un polyhèdre, car les contraintes qui définissent P sont linéaires $\boldsymbol{P}^* = \{P, P(A) \geq \alpha_A = P^*(A), \forall A\}$. La notion de fonction de répartition (cumulée) devient une paire de fonctions de répartition supérieures et inférieures F* et F_* définies par F*(x) = P*((–∞, x]) et $F_*(x) = P_*((-\infty, x])$ respectivement.

On peut comprendre tout couple de mesures [Bel, Pl], et donc tout couple [N, Π], comme étant les probabilités inférieures et supérieures induites par une famille de probabilités :

- Bel (ou, de manière équivalente, Pl) définit la famille $\boldsymbol{P} = \{P, \forall A \subseteq \Omega, P(A) \geq Bel(A)\} = \{P, \forall A \subseteq \Omega, P(A) \leq Pl(A)\}$. Dans ce cas en effet $P_* = Bel$ et $P^* = Pl$ et $\forall P \in \boldsymbol{P}$, $Bel \leq P \leq Pl$. Mais les probabilités inférieures (resp. supérieures) ne sont pas des crédibilités (resp. des plausibilités), en général[1].

[1] On peut toujours appliquer la transformée de Moebius pour obtenir les solutions en m du système d'équations $\{ \sum_{B,\, B \subseteq A_i} m(B) = P_*(A_i)\}$ – le système admet toujours

- On peut comprendre tout couple de mesures [N, Π] comme étant les probabilités inférieures et supérieures déduites d'une famille définie à partir de $\boldsymbol{P} = \{P, \forall A \subseteq \Omega, P(A) \geq N(A)\} = \{P, \forall A \subseteq \Omega, P(A) \leq \Pi(A)\}$; : dans ce cas en effet, $P^* = \Pi$ et $P_* = N$. Dans le cas fini, $\boldsymbol{P}$ peut se définir à partir de π par : $\boldsymbol{P} = \{P, \forall i=1, k\ P(A_i) \geq 1 - \alpha_{i-1}\}$. où $A_i = \{\omega: \pi(\omega) \geq \alpha_i\}$.

- Réciproquement, étant donné $A_1 \subseteq A_2 \subseteq ... \subseteq A_{k-1} \subseteq A_k$ et $\alpha_i > 0$, i=1, k tel que $\boldsymbol{P} = \{P, P(A_i) \geq \alpha_i\}$, alors on sait que P^* est une possibilité et P_* une nécessité. Si les A_i ne sont pas emboités, P^* n'est pas forcément une plausibilité.

  Comme les intervalles [Bel, Pl] dont ils sont des cas particuliers, les intervalles [N, Π] sont donc les encadrement de probabilités mal connues : $N \leq P \leq \Pi$. Mais, puisque $\Pi(A) < 1 \Rightarrow N(A) = 0$, ce sont des intervalles généralement assez larges, de la forme [0, a] ou [b, 1]. On peut donc aussi définir une contrepartie de la fonction de répartition (cumulée) pour les mesures de possibilité, et de necessité comme cas particulier des notions introduites plus haut : F* et F_* sont définies (cas continu) par

$$F^*(x) = \Pi((-\infty, x]) = \pi(x) \text{ pour } x \leq \omega^*$$
$$= 1 \text{ sinon.}$$

et
$$F_*(x) = N((-\infty, x]) = 0 \text{ pour } x \leq \omega^*$$
$$= 1 - \pi(x) \text{ sinon}$$

où est ω^* une valeur telle que $\pi(\omega^*) = 1$.

A.3 TRANSFORMATIONS POSSIBILITES ↔ PROBABILITES ↔ FONCTIONS DE CROYANCE

Comme précisé plus haut, l'importance de ces transformations tient à la possibilité qu 'elles offrent de representer toutes les données incertaines dans le même format.

A3.1. Probabilités → Fonctions de croyance

Aucune transformation n'est ici évidemment nécessaire, puisque toute probabilité est à la fois une plausibilité et une crédibilité : $Bel(A) = Pr(A) = Pl(A)\ \forall A \subseteq \Omega \Leftrightarrow$ les éléments focaux de m sont des singletons (soit : $m(F) > 0 \Rightarrow Card(F) = 1$)).
La fonction de masse correspondant à une probabilité coincide avec la distribution de cette probabilité:

$$m(\{\omega_i\}) = p(\omega_i)$$

(y compris dans le cas continu).

---

une solution - et on peut vérifier que $P_*(A) = \sum_{B, B \subseteq A} m(B)$. Mais la fonction m ainsi calculée ne sera pas forcément positive sur tous les événements.

A.3.2. Possibilités / Nécessités → Fonctions de croyance

Aucune transformation n'est nécessaire non plus ici, puisque qu'une mesure de possibilité (resp. de nécessité) est une mesure de plausibilité (resp. de crédibilité), plus particulièrement une mesure de plausibilité (resp. de crédibilité) dont les éléments focaux sont emboîtés.

Si l'on part d'une fonction de masse m à élément focaux emboîtés $A_1 \subseteq A_2 \subseteq ... \subseteq A_{k-1} \subseteq A_k$, $A_i \in$ Foc(m), la distribution de possibilité qui sous-tend la mesure est :

$$\forall\ \omega \in A_i / A_{i-1}, \pi(\omega) = \sum_{j \geq i,} m(A_j)$$

Ou, de manière équivalente : $\forall\ \omega \in A_i / A_{i-1}$, $\forall\ \omega' \in A_{i-1} / A_{i-2}$

$$\pi(\omega) = \pi(\omega') - m(A_{i-1})$$

Ainsi, les éléments de A_1 reçoivent une possibilité de 1, ceux de A_2 / A_1 une possibilité de $1 - m(A_1)$, ceux de A_i / A_{i-1} une possibilité de $1 - (m(A_1) + ... + m(A_{i-1}))$, etc. Dans le cas d'une fonction de croyance plus générale, on utilise la formule plus générale

$$\pi(\omega) = \sum_{\omega \in A} m(A).$$

Elle définit une fonction qui inclut distribution de possibilité et de probabilité.

Inversement, les éléments focaux de la distribution de masse codant une mesure de possibilité Π sont les couples de niveaux alpha de π ($1 = \alpha_1 > \alpha_2 > ... \alpha_k > \alpha_{k+1} = 0$):

$$A_i = \{\omega\ ,\ \pi(\omega) \geq \alpha_i\}, \text{ pour tout } \alpha_i \text{ t.q. } \exists\omega\ ,\ \pi(\omega) = \alpha_i$$

Et leurs masses sont données par :

$$m(A_i) = \alpha_i\ - \alpha_{i+1} \text{ (donc } m(A_k) = \alpha_k \text{ pour l'élément focal le plus extérieur).}$$

Notons qu'on ne peut reconstruire complètement m à partir de π que dans le cas possibiliste (emboité).

Dans le cas continu possibiliste, on peut définir la distribution de possibilité π à partir de deux fonctions f^- et $f^+ : [0, 1] \rightarrow r$, respectivement croissante et décroissante, telles que, $f^-(\alpha) \leq f^+(\alpha)\}$ et de la mesure Lebesgue λ (uniforme) sur [0, 1], à savoir $\pi(\omega) = \lambda(\{\alpha :\ f^-(\alpha) \leq \omega \leq f^+(\alpha)\})$. Dans ce cas, les coupes de niveaux $\{\omega: \pi(\omega) \geq \alpha\} = [f^-(\alpha), f^+(\alpha)]$.

A.3.3. Fonctions de croyances → probabilités

Le principe de la transformation dite « pignistique » par (Smets, 1990) , orignellement introduite par (Dubois et Prade, 1982) et Williams (1982) est de généraliser le principe

d'indifférence de Laplace aux fonctions de masse : prenons un m qui attribue toute la masse (= 1) à un seul élément focal F. Tous les éléments de F ont la même plausibilité.

D'après le principe de Laplace selon lequel ce qui est équiplausible est équiprobable, la distribution de probabilité correspondant à la fonction de masse m doit accorder à chaque élément de F la même probabilité 1 / Card(F). Appliquer ce principe à chacun des éléments focaux d'une fonction de masse quelconque revient à faire la somme convexe des probabilités obtenues pour chaque élément focal pris isolement, modulée bien sûr par les poids des éléments focaux m(F); ce qui justifie l'équation de la probabilité pignistique pp associée à une fonction de croyance.

$$pp(\omega) = \sum_{F,\, \omega \in F} (m(F) / Card(F))$$

La transformation pignistique a été formellement justifiée par Smets et Kennes (1990) dans le cadre des théories de représentation de l'incertitude, comme la façon rationnelle de parier à l'aide d'une probabilité subjective, pour un individu dont les connaissances sont fidèlement représentées par la fonction de masse m.

Les mesures de plausibilité et de crédibilité peuvent donc être comprises comme des enveloppes de probabilités chacune définissant un polyèdre $\boldsymbol{P} = \{P, P(A_i) \geq Bel(A_i)\}$ que l'on veut réduire à une probabilité unique. On doit choisir P dans ce polyhedre, afin de respecter la contrainte $P_*(A) \leq P(A) \leq P^*(A)$. La probabilité pignistique vérifie bien cette contrainte. Elle est même au centre de gravité de $\boldsymbol{P}$. C'est donc un choix respectant la symmétrie du problème et reflétant au mieux les tendances exprimées par les paires (Bel(A), Pl(A)).

Elle a été en fait imaginée d'abord en théorie des jeux : si l'on considère que Ω est un ensemble de joueurs, et que Bel(A) rend compte de la force relative d'une coalition de joueurs, la transformation pignistique de Bel est l'indice de Shapley (XX REF) correspondant à cette capacité. $pp(\omega)$ évalue l'influence globale du joueur ω au travers de ses alliances. On peut replacer « joueur » par « critère », en théorie de la décision.

A.3.4. Fonctions de croyances → possibilités / nécessités

Dubois et Prade (1991) se sont posé la question de l'approximation d'une paire (Bel, Pl) par une paire (N, Π). On peut définir des approximations intérieures et des approximations extérieures d'un couple (Bel, Pl).

Approximation intérieure

Un couple (N, Π) définit une approximation intérieure de (Bel, Pl) ssi pour tout $A \subseteq \Omega$, $\Pi(A) \leq Pl(A)$ (ou, de manière équivalente, $Bel(A) \leq N(A)$). On cherche une approximation optimale qui maximise la largeur des intervalles $[N(A), \Pi(A)]$. On peut montrer qu'une seule distribution π satisfait cette exigence et elle est maximale au sens

de l'inclusion d'ensemble flous. C'est la distribution la moins spécifique telle que $\forall A \subseteq \Omega$, $\Pi(A) \leq Pl(A)$ et elle coincide avec le π défini plus haut par :

$$\forall \omega \in \Omega, \pi(\omega) = \sum_{A, \omega \in A} m(A) = Pl(\{\omega\})$$

Approximation extérieure

Un couple (N, Π) définit une approximation extérieure de de (Bel, Pl) ssi pour tout $A \subseteq \Omega$, $\Pi(A) \geq Pl(A)$ (ou, de manière équivalente, $Bel(A) \geq N(A)$). On cherche ici une approximation optimale qui minimise la largeur des intervalles [N(A), Π(A)]. Il n'y a plus unicité de la solution. On peut par exemple obtenir une approximation extérieure en ordonnant arbitrairement les éléments focaux A_i de 1 à n (n étant le nombre d'éléments focaux) et en posant :

$\Pi(A_1) = Pl(A_1)$,
$\Pi(A_1 \cup A_2) = Pl(A_1 \cup A_2)$, ...,
$\Pi(A_1 \cup ... \cup A_n) = Pl(A_1 \cup ... \cup A_n) = 1$.

Le problème est que, sauf cas particulier, il n'existe pas en général d'approximation extérieure la plus spécifique de l'intervalle [Bel, Pl]. Ces approximations on un sens pour des probabilités imprécises (P^*, P_*).

Un cas particulier où cet encadrement possibiliste extérieur optimal est unique est celui où les éléments focaux de m sont disjoints (par exemple, le cas où Bel = Pl est une probabilité) et linéairement ordonnés par m. Dans ce cas, la mesure de possibilité la moins spécifique telle que $A \subseteq \Omega$, $\Pi(A) \geq Pl(A)$ est la mesure Π définie par :

$$\Pi(A_{\sigma(1)} \cup ... \cup A_{\sigma(i)}) = Pl(A_{\sigma(1)} \cup ... \cup A_{\sigma(i)}) \ \forall i = 1, n$$

où σ est la permutation des éléments focaux telle que $m(A_{\sigma(j)}) < m(A_{\sigma(i)}) \ \forall \ i < j$

A.3.5. Possibilités / Nécessités → Probabilités

Puisque toute mesure de nécessité est une mesure de crédibilité particulière on pourra transformer une mesure de nécessité/possibilité en une mesure de probabilité en appliquant la transformation pignistique. Rappelons que le principe de la transformation pignistique est le principe de Laplace, qui dit que tout ce qui est équiplausible doit être équiprobable. La transformation pignistique d'une distribution de possibilité a été étudiée dans le cas discret (Yager 1982 ; Dubois, Prade 1982; Dubois et al 1993).

Appliquer les principes de la transformation pignistique revient donc à construire la distribution de probabilité à partir des coupes de niveaux alpha de la distribution et de la fonction de masse m déduite de π. En pratique si $\Omega = \{\omega_1, \omega_2, ... \omega_n\}$ et $\pi(\omega_1) = 1 \geq \pi(\omega_2) \geq ... \geq \pi(\omega_n) \geq \pi(\omega_{n+1}) = 0$)

$$pp(\omega_i) = (\pi(\omega_i) - \pi(\omega_{i+1}))/ i \text{ , pour } i = 1, n$$

En effet, si $\pi(\omega_i) > \pi(\omega_{i+1})$, alors $\{\omega_1, \omega_2, \ldots \omega_i\}$ est focal et $m(\{\omega_1, \omega_2, \ldots \omega_i\}) = \pi(\omega_i) - \pi(\omega_{i+1})$. Cette transformation est bijective.

La transformation pignistique dans le cas continu (Sandri 1991)

Dans le cas continu, cette transformation s'écrit :

$$p(\omega) = \int_0^{\pi(\omega)} (1/ \text{Card}((\pi)_\alpha) d\alpha \text{ où } (\pi)_\alpha = \{\omega, \pi(\omega) \geq \alpha\}$$

Par exemple, si π est un triangle isocèle de base a autour de 0, on obtient :

$$p(\omega) = \int_0^{\pi(\omega)} 1 / (2.\, a\, .\, (1 - \alpha)) \; d\alpha = -.\, (\log(\pi(\omega))/2a$$

Dans le cas où π est un intervalle flou de support [a, d], de noyau [b, c] et de type L/R ($\pi(\omega) = L(\omega)$ si $\omega \in [a, b]$, $\pi(\omega) = 1$ si $\omega \in [b, c]$, $\pi(\omega) = R(\omega)$ si $\omega \in [c, d]$), la transformation s'écrit :

$$p(\omega) = \int_0^{\pi(\omega)} 1 / (R^{-1}(\alpha) - L^{-1}(\alpha)) \quad d\alpha$$

Par exemple, si π est trapezoidal :

$$L(\omega) = (\omega - a) / (b - a), \qquad R(\omega) = (\omega - d) / (c - d),$$
$$L^{-1}(\alpha) = (b - a)\, \alpha + a, \qquad R^{-1}(\alpha) = (c - d)\, \alpha + d$$

Et l'on obtient :

$$p(\omega) = \int_0^{\pi(\omega)} 1 / (p\, \alpha + q) \quad d\alpha \qquad \text{où } p = c - d - b + a \text{ et } q = a - d$$
$$p(\omega) = 1 / p\, .\, \ln ((p\, \pi(\omega) + q) / q)$$

Le maximum d'entropie (Klir, 1990)

Considérons une famille de probabilités ***P*** que l'on veut caractériser par l'un de ses éléments. L'objectif étant de biaiser au minimum l'information générée en réduisant ***P*** à une probabilité unique, Klir (1990) propose de caractériser cette famille par celui de ses éléments qui possède l'entropie maximale. Rappelons que l'entropie d'une mesure de probabilité est :

$$\text{Entropie}(P) = \Sigma_{\omega \in \Omega} p(\omega) \log(p(\omega))$$

Cette mesure est maximale pour les distributions uniformes et minimale en absence d'incertitude (i.e. quand $\exists\, \omega_0 \in \Omega$, $p(\omega_0) = 1$).

Le principe du maximum d'entropie va donc choisir dans ***P*** une probabilité maximisant $\Sigma_{\omega \in \Omega} p(\omega) \log(p(\omega))$. Rappelons que cette sélection revient à effectuer le maximum

d'hypothèses d'indépendance possibles entre événements ; ce type de transformation est plus adapté aux probabilités objectives (statistiques) que subjectives. Notamment si ***P*** contient la distribution uniforme, c'est celle-là qui sera choisie aussi proche qu'elle soit du bord du polyhèdre ***P***. Elle ne reflète dont pas du tout les tendances exprimées par les bornes de probabilités.

Quoiqu'il en soit, le principe de maximum d'entropie a été appliqué par Moral (1986) à la transformation d'une possibilité en probabilité. En considérant que les ω_i de Ω ont été ordonnées de manière à c e que $\pi(\omega_1) \leq \pi(\omega_2) \leq ... \leq \pi(\omega_n)$, la transformée de π est :

$$p(\omega_1) = \min_{i=1,n} \pi(\omega_i) / i$$
$$p(\omega_k) = \min_{i=k,n} (\pi(\omega_i) - \Sigma_{j=1,k-1}\, p(\omega_k)) / (i - k + 1)$$

A.3.6. Probabilités → Possibilités / Nécessités

Le principe ici est de trouver une approximation supérieure (extérieure) d'une probabilité P : on se retrouve donc dans un cas particulier de l'approximation supérieure de fonctions de croyances (cf. Section 3.3.2) : si on considère que P represente toute l'information disponible (issue d 'un histogramme par exemple), on cherche donc la possibilité Π la plus spécifique telle que $\forall A \subseteq \Omega$, $\Pi(A) \geq P(A) \geq N(A)$, afin de perdre le moins d'information possible. Le problème est qu'en général, n'y a pas unicité de la solution optimale, sauf si l'on ajoute que l'ordre donné sur Ω par les distributions doit être conservé, c'est à dire que :

$$p(a) \geq p(b) \Leftrightarrow \pi(a) \geq \pi(b)$$

ce qui semble une condition raisonnable.

Approximation possibiliste extérieure d'une probabilité : cas discret
(Dubois, Prade 1982 ; Delgado, Moral 1987)

Dans le cas discret, la distribution π la plus spécifique satisfaisant aux conditions citées ci-dessus est :

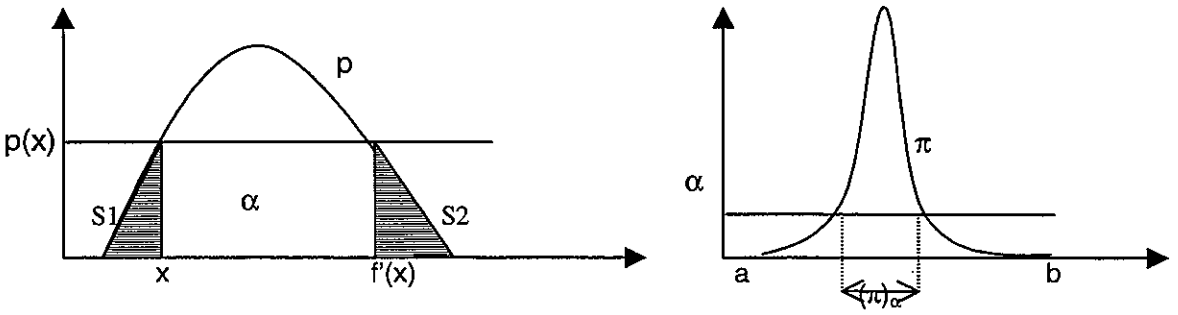
$$\pi(\omega) = \Sigma_{\omega',\, p(\omega') \leq p(\omega)}\, p(\omega')$$

Si p est linéaire ($p(\omega_1) > p(\omega_2) > ... > p(\omega_n)$), on a : $\pi(\omega_1) = 1$, $\pi(\omega_2) = 1 - p(\omega_1)$, etc. On vérifie facilement que $\Pi(\{\omega_i, ..., \omega_n\}) = P(\{\omega_i, ..., \omega_n\})$. Cette transformation est bijective.

Approximation possibiliste extérieure d'une probabilité : cas continu
(Dubois et al 1991)

Les mêmes principes s'appliquent si p est une distribution continue sur un sous ensemble de la droite réelle. Dans le cas d'une distribution continue unimodale (voir Figure ci-dessous) :

$\pi(x) = S_1 + S_2$ où $S_1 = P((-\infty\ x])$, $S_2 = P([f'(x), +\infty))$
avec $f'(x) = \max\{y, p(y) = p(x)\}$



Cela revient au calcul de l'intervalle de confiance [x, f'(x)] de probabilité α: tel que $\pi(x) = \pi(f'(x)) = 1 - \alpha = S_1 + S_2$ autour du mode de P. Dit autrement : $P((\pi)_\alpha) \geq 1 - \alpha$, où $(\pi)_\alpha$ est la coupe de niveau α de π.

Il faut noter que lorsque p est une distribution unimodale, π(x) est toujours une fonction concave de part et d'autre du mode. Si p est par exemple la distribution uniforme sur un intervalle [a, b], l'approximation extérieure symmétrique la plus spécifique de p sera la distribution de possibilité triangulaire (a, (a + b) / 2, b)[2]. Cette distribution de possibilité est moins spécifique que la transformée de n'importe laquelle des probabilités de même support de densité unimodale symmétrique. La distribution de possibilité triangulaire est donc un choix rationnel quand on ne connaît qu'un intervalle contenant le paramètre incertain, et qu'on suppose sa distribution symmétrique.

Transformation pignistique inverse (Dubois, Prade 1983 ; Dubois et al 2001)

Les principes de la transformation pignistique d'une fonction de croyance en probabilité permettent d'imaginer une autre transformation possibiliste : on considère que p est en fait une probabilité subjective, qui est l'expression sous forme probabiliste d'une connaissance plus imprécise d'un agent (laquelle est supposée avoir la forme d'une fonction de croyance Bel). Ce ne sont que les règles imposées par le pari qui forcent l'agent à fournir une probabilité unique. On suppose qu'il choisit la transformation pignistique de Bel. En particulier si l'agent ne possède aucune information il fournit une probabilité uniforme.

Afin de modéliser correctement l'information détenue par l'agent, il faut donc rechercher la mesure de crédibilité dont p est la transformation pignistique. Comme il y a beaucoup de solutions, on fait l'hypothèse prudente d'ignorance maximale de l'agent. On cherche donc la moins informative des fonctions Bel de transformée pignistique p

---

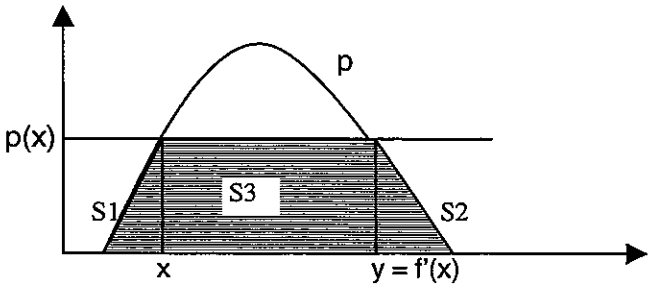
[2] elle ne respecte pas la condition $p(a) \geq p(b) \Leftrightarrow \pi(a) \geq \pi(b)$. Si on l'imposait ici, on obtiendrait la possibilité uniforme (fonction caractéristique du support de p).

fournie par l'agent. On suppose que l'imprécision de Bel est mesurée par la cardinalité moyenne pondérée des éléments focaux (I(Bel) = Σ m(A)·Card(A)). L'idée est que les grands ensembles sont plus imprécis que les petits.

On peut montrer que la mesure de crédibilité, de transformée pignistique p, qui maximise I(Bel) est toujours une mesure de nécessité ; plus précisément, quand p est une densité continue unimodale (voir figure ci-dessous), sa transformée pignistique inverse est définie par la distribution de possibilité π suivante :

$$\pi(x) = S_1 + S_2 + S_3$$

avec $S_1 = P((-\infty\ x])$, $S_2 = P([f'(x), +\infty))$ $S_2 = P([x, f'(x)])$,
où $f'(x) = \max(y, p(y) = p(x))$



Notons que si p est une distribution uniforme, la possibilité obtenue est elle aussi uniforme ($\pi(\omega) = 1$ pour tout $\omega \in [a, b]$).

Dans le cas discret, cette transformation s'écrit :

$$\pi(\omega) = \sum_{\omega' \in \Omega} \min(p(\omega), p(\omega))$$

A.4. SATISFACTION D'UN CRITERE PAR UNE FONCTION A PARAMETRES HETEROGENES : LE PROBLEME DIRECT

Soit une fonction f de k variables x_i dont certaines sont des variables aléatoires et d'autres des variables possibilistes et un critère sur f. On pourra par exemple prendre k = 2 et $f(x, y) = ax + y$, où x est contraint par une distribution de probabilité p et y par une distribution de possibilités π, et tester le critère $f(x, y) \leq e$. La question est de savoir dans quelle mesure le critère est satisfait.

A.4.1 Le cadre général de la présence simultanée de variables aléatoires et possibilistes

Lorsque toutes les variables de f sont des variables aléatoires, f est également une variable aléatoire dont on peut calculer la distribution de probabilité ou la fonction de répartition dans le cas discret:

$$P(f = u) = \Sigma_{u1, ..., uk,\ f(u1, ..., uk) = u}\ p(u_1, ..., u_k) = p_f(u).$$
$$P(f \leq e) = \Sigma_{u1, ..., uk,\ f(u1, ..., uk) \leq e}\ p(u_1, ..., u_k) = F_f(e)$$

où F_f est la fonction de répartition de $f(x_1, .., x_k)$. Cette dernière valeur est typiquement la probabilité que le critère soit satisfait .Si les variables aléatoires sont indépendantes :

$$P(f = u) = \Sigma_{u1, ..., uk,\ f(u1, ..., uk) = u}\ p_1(u_1) * ... * p_k(u_k).$$
$$P(f \leq e) = \Sigma_{u1, ..., uk,\ f(u1, ..., uk) \leq e}\ p_1(u_1) * ... * p_k(u_k).$$

On peut plus généralement calculer la probabilité de l'événement E :

$$P_f(E) = \Sigma_{u1, ..., uk,\ f(u1, ..., uk) \text{ satisfait } E}\ p(u_1, ..., u_k)$$
$$= \Sigma_{u1, ..., uk,\ f(u1, ..., uk) \text{ satisfait } E}\ p_1(u_1) * ... * p_k(u_k).$$

Dans le cas continu, on doit calculer la densité de probabilité ou la fonction de répartition avec les intégrales correspondantes aux formules ci-dessus, analytiquement si c'est possible, ou les approximer par différentes méthodes (par exemple une méthode de Monte Carlo).

Lorsque toutes les variables de f sont restreintes par des distributions de possibilité, les valeurs plus ou moins possibles de f seront également restreintes par une distribution de possibilité. Son calcul fera appel à des méthodes de calcul d'intervalles flous, analytiques si possible, ou à des approximations par alpha coupes.

$$\pi_f(u) = Sup_{u1, ..., uk,\ f(u1, ..., uk) = u}\ \pi(u_1, ..., u_k).$$

On peut également calculer l'équivalent des fonctions de répartition pour les mesures N et Π associées à la distribution, ou encore la possibilité et la nécessité que tel ou tel critère soit satisfait.

$$\Pi(f \leq e) = Sup_{u1, ..., uk,\ f(u1, ..., uk) \leq e}\ \pi(u_1, ..., u_k) = F_f^*(e).$$
$$N(f \leq e) = Inf_{u1, ..., uk,\ f(u1, ..., uk) > e}\ 1 - \pi(u_1, ..., u_k) = F_{f^*}(e),$$
$$\Pi(E) = Sup_{u1, ..., uk,\ f(u1, ..., uk) \text{ satisfait } E}\ \pi(u_1, ..., u_k),$$
$$N(E) = Inf_{u1, ..., uk \text{ ne satisfait pas } E}\ 1 - \pi(u_1, ..., u_k),$$

Si les variables possibilistes sont non-interactives (on ne fait alors pas d'hypothèse sur les liens entre les variables), on peut remplacer $\pi(u_1, ..., u_k)$ par $\min(\pi_1(u_1), ..., \pi_k(u_k))$ dans les formules précédentes, par exemple:

$$\pi(u) = Sup_{u1, ..., uk,\ f(u1, ..., uk) = u}\ \min(\pi(u_1), ..., \pi(u_k))$$

Si les variables sont renseignées par des sources indépendantes , on peut remplacer $\pi(u_1, ..., u_k)$ par $\pi_1(u_1) \cdot ... \cdot \pi_k(u_k)$ dans les formules précédentes. Par convention, si π_i est la fonction d'appartenance d'un intervalle flou M_i, on note $f(M_1, ..., M_k)$ l' intervalle flou de fonction d'appartenance π_f.

La question est donc de savoir quelle est la nature du résultat lorsque f met en jeu les deux types de variables. Soient $x_1, \ldots x_i$ les variables aléatoires, et $x_{i+1}, \ldots x_k$ les variables possibilistes. Deux réponses sont possibles, selon la façon d'aborder ce problème.

Approche Hétérogène

Cette approche repose sur la remarque suivante : si les variables aléatoires $(x_1, \ldots x_i)$ ont leur valeurs fixées à $(u_1, \ldots u_i)$, $f(u_1, \ldots u_i, x_{i+1}, \ldots x_k)$ est une une variable possibiliste dont les valeurs sont restreintes par un ensemble flou $f(u_1, \ldots u_i, M_{i+1}, \ldots M_k)$. A chaque combinaison $(u_1, \ldots u_i)$ de valeurs des variables aléatoires correspond donc un ensemble flou, et une probabilité qui lui est attachée (la probabilité de la combinaison $(u_1, \ldots u_i)$) : $f(x_1, \ldots x_i, x_{i+1}, \ldots x_k)$ est un donc un nombre flou aléatoire décrit par une distribution de probabilité sur ... des ensembles flous !

En faisant une hypothèse de non interactivité sur les variables possibilistes, et supposant que x_j soit restreinte par l'intervalle flou M_i de distribution π_i pour $j = i+1,\ldots,k$, la probabilité que la distribution de possibilité sur : $f(x_1, \ldots x_i, x_{i+1}, \ldots x_k)$ soit le sous ensemble flou M de Ω est :

$$p(f = M) = \Sigma_{u1, \ldots, ui,\, f(u1, \ldots, ui,\, Mi+1, \ldots, Mk) = M}\, p(u_1, \ldots, u_i)$$

On obtient donc une variable floue aléatoire.

Si tant est que le calcul de p est possible, on devrait alors calculer des distributions de probabilité sur les degrés de possibilité ou de nécessité de tel ou tel événement. Inversement, on devrait pouvoir calculer des intervalles de confiance à d% de ces mesures de possibilité.

En pratique, en appliquant une méthode de Monte-Carlo aux variables aléatoires on peut engendrer une « statistique floue » à partir d'un ensemble T de tirages de tuples $(u_1, \ldots u_i)$ c'est à dire un échantillon d'intervalles flous $\{f(u_1, \ldots u_i, M_{i+1}, \ldots M_k), (u_1, \ldots u_i) \in T\}$ qu'il convient d'exploiter.

Approche Homogène

Cette seconde approche considère que variables aléatoires comme variables possibilistes sont en fait des ensembles aléatoires décrits par des distributions de masse m_k, dont les éléments focaux U_k sont des singletons ou des intervalles consonants selon la nature de la variable concernée. f est donc également un ensemble aléatoire, dont la fonction de masse est :

$$m_f(A) = \Sigma_{U1, \ldots, Uk,\, f(U1, \ldots, Uk) = A}\, m_1(U_1) * \ldots * m_k(U_k)$$

sous l'hypothèse d'indépendance des variables.

Le calcul de $f(U_1, ..., U_k)$ est effectué par le calcul d'erreurs. Il faut noter que les élément focaux $f(U_1, ..., U_k)$ de m_f ne sont en général ni emboîtés, ni des singletons. On obtient donc une fonction de croyance.

On peut à partir de cette distribution de masse calculer les indices :

$$Pl(f \leq e) = \Sigma_{A,\, inf(A) \leq e}\, m_f(A) = F_f^*(e),$$
$$Bel(f \leq e) = \Sigma_{A,\, sup(A) \leq e}\, m_f(A) = F_{f*}(e),$$
$$Pl(E) = \Sigma_{\forall\, a \in A,\, a \text{ satisfait } A}\, m_f(A),$$
$$Bel(E) = \Sigma_{\exists\, a \in A,\, a \text{ satisfait } A}\, m_f(A).$$

Là encore, le calcul analytique est souvent loin d'être simple dans le cas continu, en particulier quand la fonction met en jeu de nombreuses densités de probabilité. Néanmoins on peut approximer ces densités par des ensembles aléatoires FINIS dissonants, comme on le verra plus loin.

A.4.2 Calcul approché et exploitation des résultats pour l'approche hétérogène par Monte-Carlo plus analyse de sensibilité

Cette approximation repose sur une compréhension de f comme générant un ensemble flou aléatoire. Il s'agit d'effectuer une Monte Carlo sur les variables aléatoires seules pour estimer f à partir d'un échantillon d'ensemble flous. A chaque tirage, on fixe des valeurs pour chacune des variables aléatoires. Puisque ces variables sont devenues des constantes, on applique f à des constantes et des intervalles flous. On obtient à l'issue du Monte Carlo une famille d'intervalles flous (un par tirage) $\Phi = \{F_1, ..., F_n\}$: Φ est une approximation du nombre aléatoire flou $f(x_1, .., x_k)$. On cherche ensuite à estimer dans quel mesure cette approximation satisfait une critère e.

Dans Guyonnet et al (2001) il a été proposé de synthétiser cette famille en un ensemble flou unique F_d, tel que, dans d % des F_i, $\mu_{Fi}(x) \geq \mu_{Fd}(x)$. Pour ce, on envisage de procéder par alpha coupes : pour chaque niveau de coupe α, on possède donc une famille d'intervalles $(\Phi)_\alpha = \{(F_1)_\alpha, ..., (F_n)_\alpha\}$, où chaque $(F_i)_\alpha$, est de la forme $[b\_inf_{i\alpha}, b\_sup_{i\alpha}]$. On synthétise cette famille d'intervalles en un « intervalle d% » noté $[b\_inf(\alpha, d), b\_sup(\alpha, d)]$ et défini par :

$$b\_inf(\alpha, d) = \sup \{ b\_inf_{i\alpha}, Card(\{j : b\_inf_{j\alpha} \geq b\_inf_{i\alpha}\} \geq n * d / 100\}$$
$$b\_sup(\alpha, d) = \inf \{ b\_sup_{i\alpha}, Card(\{j : b\_sup_{j\alpha} \leq b\_sup_{i\alpha}\} \leq n * d / 100\}$$

$b\_inf(\alpha, d)$ est la borne inf inférieure à d% des bornes infs et $b\_sup(\alpha, d)$ est la borne sup supérieure à d% des bornes sup.

F_d est ensuite reconstitué à partir de ses coupes $[b\_inf(\alpha, d), b\_sup(\alpha, d)]$. F_d est donc une estimation « à d% », moyennant les approximations par Monte Carlo et par α

coupes, des valeurs plus ou moins possibles de f. On peut donc calculer des estimations de la possibilité et de la nécessité que le critère soit satisfait. Par exemple :

$$\Pi_d(f \leq e) = \sup_{u \leq e} \mu_{Fd}(u) \qquad N_d(f \leq e)) = \inf_{u > e} 1 - \mu_{Fd}(u)$$

On peut aussi calculer des possibilité et des nécessité cumulées :

$\Pi_{d\,cumul}(e) = \Pi(f \leq e)$ est l'ensemble des point possiblement après f et
$N_{d\,cumul}(e) = N(f \leq e)$ est l'ensemble des point nécessairement après f

Cette méthode repose sur une hypothèse d'approximation à laquelle il faut faire attention : pour calculer les intervalles [b_inf(α, d), b_sup(α, d)], on traite des deux bornes indépendamment, alors qu'elles ne le sont pas : si d % des bornes inf sont supérieures à b_inf(α, d) et d % des bornes sup inférieures à b_sup(α, d), il est faux d'affirmer que d % des intervalles sont contenus dans les intervalles [b_inf(α, d), b_sup(α, d)]. Par exemple, considérons les intervalles suivants [1, 20], [2, 21], ..., [20, 39]. La borne inférieure à 95% est 2, la borne supérieure à 95% est 38. Or seuls 18 intervalles sur les 20 (i.e. 90%) sont compris dans [2,39]. Le problème est que les bornes inférieures et supérieures des intervalles sont traités comme des variables indépendantes alors qu'ils ne le sont pas. Cela implique que le F_d approximé par cette méthode ne peut pas être utilisé pour l'estimation de *tous* les types de critères. Heureusement, les critères ne mettant en jeu qu'une seule borne sont en général correctement estimés. En particulier, $N_d(f \leq e)$ estime correctement la nécessité que d% des intervalles flous satisfont le critère (ce calcul ne met en jeu que les bornes droites des intervalles) ; de même, $\Pi_d(f \leq e)$ estime correctement la possibilité que d% des intervalles flous satisfont le critère (ce calcul ne met en jeu que les bornes gauche des intervalles), et $[\Pi_d(f \leq e), N_d(f \leq e)]$ est bien un encadrement possibiliste de la probabilité que d% des intervalles flous tirés par Monte-Carlo satisfont le critère[3].

Autres méthodes possibles :

- Calculer la moyenne floue E(f) des $\{F_1, ..., F_n\}$:et un écart-type scalaire et(f) (voir Gil et al.). En déduire $N(E(F) \leq e)$ $\Pi(E(F) \leq e)$, $N(E(F) + et(f) \leq e)$ et $\Pi(E(F) + et(f) \leq e)$ $N(E(F) - et(f) \leq e)$ et $\Pi(E(F) + et(f) \leq e)$
- Calculer $N(f \leq e)$ et $\Pi(f \leq e)$ pour chaque $F \in \{F_1, ..., F_n\}$. On obtient des statistiques $\{N_1(f \leq e), ..., N_n(f \leq e)\}$: et $\{\Pi_1(f \leq e), ..., \Pi_n(f \leq e)\}$, soit des histogrammes sur les degrés de satisfaction aléatoires du critère.

A.4.3. Approximation par une méthode de Monte-Carlo sur des ensembles aléatoires par l'approche homogène

[3] Le type de critère mettant en échec ce type de méthode est typiquement un critère mettant en jeu à la fois la borne gauche et la borne droite de l'intervalle flou, par exemple un critère de type $f \in [e1, e2]$.

Le principe de cette approximation est de considérer que les variables de f sont toutes de même type, à savoir des variables contraintes par des fonctions de croyances (puisque possibilités comme probabilités sont des fonctions de croyance) – f est donc un ensemble aléatoire. Puis on effectue une discrétisation sur ces fonctions de croyance, de manière à pourvoir appliquer un algorithme de Monte-Carlo.

Méthode de Monte-Carlo sur des fonctions de croyance

Lorsque les éléments focaux sont des intervalles ou des points et qu'il sont en nombre infini, les méthodes de Monte Carlo s'appliquent à partir de fonctions de croyance quasiment aussi simplement qu'à partir de distributions probabilité. En effet, il est facile d'effectuer des tirages aléatoires parmi les élément focaux d'une distribution de masse, en respectant la probabilité de tirage donnée par la fonction de masse (probabilité de tirer $A = m(A)$).

A chaque itération du Monte Carlo, on va donc tirer un intervalle I_{xi} par variable de x_i de f. On peut alors calculer l'image de ces intervalles par f. Dans le cas simple où f est continue et monotone dans chacun de ses argument, $f(I_{x1}, ..., I_{xk})$ est un intervalle, dont la borne inférieure est obtenue en appliquant f aux bornes inférieures des x_i croissants et aux bornes supérieures des x_i décroissants, et la borne supérieure de $f(I_{x1}, ..., I_{xk})$ est obtenue en appliquant f aux bornes supérieures des x_i croissants et aux bornes inférieures des x_i décroissants. A partir de N tirages, on obtient un ensemble de N intervalles : $\Phi = \{F_1, ..., F_N\}$ qui sont autant d'élément focaux d'une fonction de masse qui associe à chacun d'eux la masse 1 / N.

Cette approche permet de calculer approximativement une fonction de variables hétérogènes probabilistes et possibilistes continues. Pour les variables aléatoires on tire des valeurs au hasard selon la distribution; pour les variables posssibilistes, on produit des coupes de niveau au hasard en tirant un niveau de possibilité au hasard selon une loi uniforme, et en calculant la coupe de niveau correspondante. On obtient $\Phi = \{f(u_1, \ldots u_i, (M_{i+1})_{\alpha i+1} \ldots (M_k)_{\alpha k}), (u_1, \ldots u_i, \alpha_{i+1}, \ldots \alpha_n) \in T\}$.

Notons la différence avec le cas précédent : ici, on tire une coupe de niveau différente pour chaque variable possibiliste. En revanche, la méthode hétérogène considere que c'est le meme niveau de possibilite qui est choisi pour toutes les variables possibilistes, ce qui mène au calcul d'intervalles flous. Le calcul homogène proposé ici suppose une indépendance stochastique entre toutes les variables.

Par exemple, si l'on s'intéresse à la fonction $f(x, y) = a \cdot x + y$ et $E = \{f(x, y) \leq e\}$, avec des variables possibilistes traitées comme des densités d'ensembles aléatoires, on tire à chaque itération i deux intervalles (alpha coupes) : $[x_i, x_i']$ et $[y_i, y_i']$. Les F_i sont de la forme $[x_i + a \cdot y_i, x_i' + a \cdot y_i']$. On a donc deux courbes d'encadrement de la probabilité :

$Bel(f(x, y) \leq e) = Card(\{i, , x_i' + a \cdot y_i' \leq e\}) / N$
$Pl(f(x, y) \leq e) = Card(\{i, , x_i + a \cdot y_i \leq e\}) / N$

$PP((f(x, y) \leq e) = \sum_{i,\, xi + a\,.\, yi \leq e \leq,\, xi' + a\,.\, yi'} (x_i' + a\,.\, y_i' - e) / ((x_i' - x_i + a\,.\, y_i' - a\,.\, y_i)).\ N)$

Méthode directe dans le cas fini

Si les variables sont toutes représentées par des intervalles aléatoires $\{m_1, ..., m_k\}$avec un nombre fini déléments focaux U_i , alors il suffit de calculer $m_f(A) = \Sigma_{U1, ..., Uk\,,\, f(U1, ..., Uk) = A}\ m_1(U_1) * ... * m_k(U_k)$ sans passer par une methode de MonteCarlo, puisque le nombre de calcul est fini. On obtient encore un ensemble de N intervalles : $\Phi = \{F_1, ..., F_N\}$, cette fois pondérés par $m_f(F_i)$.

Pour tout événement E, on peut alors calculer un encadrement de sa probabilité :

$$Bel(E) = \Sigma_{\,Fi \subseteq E}\, m_f(F_i) \ \leq P(E) \leq Pl(E) = \Sigma_{\,Fi \cap E \neq \varnothing}\, m_f(F_i)$$

En particulier $Bel(f \leq e) = \Sigma_{\,supFi \leq e}\, m_f(F_i)$ et $Pl(f \leq e) = \Sigma_{\,infFi \leq e}\, m_f(F_i)$.

Une autre solution est d'extraire une distribution pignistique de cette fonction de croyance :

$$pp(x) = \sum_{i,\, x \in Fi} m_f(F_i)\ /\ ((\sup(F_i) - \inf(F_i)))$$

et si FF est sa fonction de répartition, $FF(e) = PP((f(x, y) \leq e)$ est la probabilité de demeurer en dessous du seuil.

Discrétisation d'une fonction de masse

Lorsque les éléments focaux de m ne sont pas en nombre fini (par exemple si m est une probabilité continue sur un intervalle de la droite réelle), ou s'ils sont en trop grand nombre, on peut approximer m par un m' moins spécifique au sens de l'inclusion des ensembles aléatoires.

On dit qu'une fonction de masse m est incluse dans m' ($m \subseteq m'$), ou que m est plus spécifique que m' ssi les trois conditions suivantes sont vérifiées :

(i) $\forall\ A \in Foc(m)\ , \exists\ B \in Foc(m')$ t.q. $A \subseteq B$

(ii) $\forall\ B \in Foc(m'), \exists\ A \in Foc(m)$ t.q. $A \subseteq B$

(iii) il existe une matrice d'affectation W, d'entrées $A \subseteq Foc(m)$, $B \subseteq Foc(m')$ telle que :

$$A \not\subset B \Rightarrow W(A, B) = 0,$$
$$\forall\ A \in Foc(m),\ m(A) = \sum_{B,\, A \subseteq B} W(A, B)$$
$$\forall\ B \in Foc(m'),\ m'(B) = \sum_{B,\, A \subseteq B} W(A, B)$$

En d'autres termes, la masse m(A) d'un élément focal de m sera partagée entre un ou plusieurs sur-ensembles B de A qui seront les éléments focaux de m', et la masse m' d'un B ne sera que la somme des masses cédées par les A qu'il contient. Un cas particulier intéressant est celui où l'on regroupe simplement les éléments focaux de m en surensembles (c'est typiquement ce que l'on fait en discrétisant m) :

$\forall A \in Foc(m)$, $\exists !\ B \in Foc(m')$ t.q. $A \subseteq B$ et $\forall B \in Foc(m')$, $m'(B) = \sum_{B,\, A \subseteq B} m(A)$

Les distribution de masses liées par une relation d'inclusion ont la propriété remarquable suivante (Dubois, Prade 1991) :

$$m \subseteq m' \Rightarrow [Bel, Pl] \subseteq [Bel\ ', Pl']$$

C'est à dire qu'en approximant m par une fonction de masse moins spécifique, on obtient une approximation extérieure de Bel et de Pl : Bel' ≤ Bel, Pl' ≥ Pl. De plus, si f est une fonction d'une variable x, et m_f l'image de m par f (un ensemble aléatoire tel que $m_f(A) = m(f^{-1}(A))$), alors

$$m \subseteq m' \Rightarrow m_f \subseteq m'_f$$

C'est à dire qu'en approximant m par une fonction de masse moins spécifique, l'image de cettte dernière est encore une approximation supérieure de l'image de la première. Si calculer m'_f est plus facile que de calculer m_f, on peut facilement calculer une approximation « encadrante » de m_f.

Une application directe de cette propriété est que la méthode de Monte Carlo appliquée ci dessus à des fonctions de masse continues, donc à des variables aléatoires ou à des variables possibilistes décrites par des distributions continues peut être remplacée par une méthode directe, exacte, avec un nombre de calculs finis, pour obtenir une approximation extérieure du résultat. Il suffit de discrétiser les distributions de possibilité et de probabilité en en construisant des approximations finies extérieures.

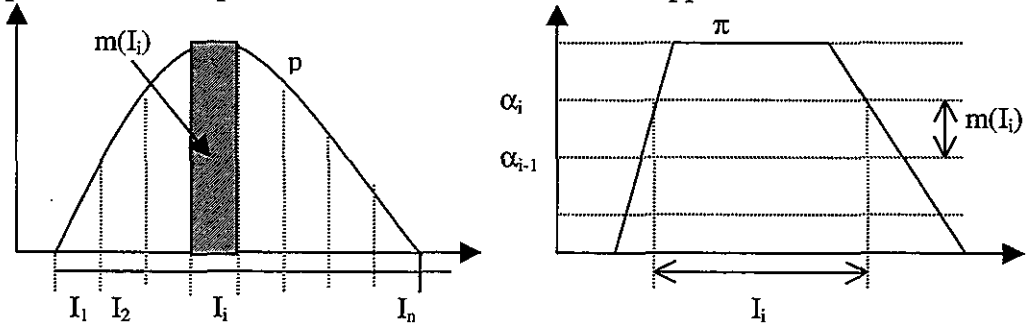


Fig. 6 **-** Discrétisation d'une distribution de probabilité, d'une distribution de possibilité

Discrétisation d'une distribution de probabilité: On va simplement partionner le support [a, b] de la distribution de probabilité en n intervalles I_1, ..., I_n , typiquement des intervalles de même largeur l = (b – a) / n: I_1 = [a, a + l], I_2 = [a + l, a + 2 . l], ..., I_n = [b – l, b] (on peut préférer partitionner [a, b] en intervalles de même probabilité). Ceci revient à transformer la fonction de masse m correspondant à p en une fonction de masse m' d'éléments focaux disjoints I_1, ..., I_n : $m(I_i) = P(I_i)$. Clairement, $p \subseteq m'$.
Discrétisation d'une distribution de possibilité: On va simplement extraire du support [a, b] de π n intervalles $I_1 \subseteq ... \subseteq I_n \subseteq [a, b]$, qui sont autant de coupes de niveau α_i de π. Typiquement, on peut prendre un δ = 1/n et fixer les α_i tous des δ : α_n = 1/n, α_{n-1} = 2/n,

..., $\alpha_i = \alpha_{i-1} + 1/n$, $\alpha_1 = 1$. On obtient donc la fonction de masse d'élément focaux $I_i = \{\omega, \pi(\omega) \geq \alpha_i\}$ et de masse $m'(I_i) = \alpha_i - \alpha_{i-1}$ (dans notre exemple de $m'(I_i) = \delta = 1/n$). Clairement, $\pi \subseteq m'$.

Principe d'un calcul homogène sur variables mixtes probabilistes et possibilistes

Résumons les principes exprimés ci dessus et appliquons les à notre fonction, $f(x, y) = ax + y$, où x est contraint par une distribution de probabilité p et y par une distribution de possibilités π :

- Discrétiser la distribution p en une fonction de masse m_x,
- Discrétiser la distribution en une fonction de masse m_y,
- Pour i = 1, N
  - Tirer selon m_x un élément focal $[x_i, x_i']$ de $Foc(m_x)$,
  - Tirer selon m_y un élément focal $[y_i, y_i']$ de $Foc(m_y)$,
  - Calculer $F_i = f([x_i, x_i'], [y_i, y_i']) = [x_i + a \cdot y_i, x_i' + a \cdot y_i']$ et le mémoriser dans $\boldsymbol{F}$,
- Calculer : $Bel(E) = Card(\{F_i \in \boldsymbol{F}, F_i \subseteq E\}) \leq P(E) \leq Pl(E) = Card(\{F_i \in \boldsymbol{F}, F_i \cap E \neq \varnothing\})$

  C'est à dire : $Bel(f(x, y) \leq e) = Card(\{i, , x_i' + a \cdot y_i' \leq e\}) / N$

  $Pl(f(x, y) \leq e) = Card(\{i, , x_i + a \cdot y_i \leq e\}) / N$

Bien entendu, il n'est pas nécessaire de mémoriser tous les F_i et on peut imaginer un algorithme un petit peu sophistiqué qui calcule incrémentalement Bel(E) et Pl(E). On peut aussi simplifier la représentation finale en replacant deux éléments focaux A et B tres semblables par leur union affectée de la somme des masses.

Méthode homogène par transformation en possibilité ou probabilité

Une dernière façon d'aborder le problème serait de transformer les variables aléatoires en variables possibilistes, et d'effectuer le calcul d'une distribution de possibilité sur f, ou inversement, de transformer les distributions de possibilités en distributions de probabilité et d'effectuer un calcul de variables aléatoires.

Dans le premier cas, on irait vers une perte d'information, puisque l'on noie une probabilité dans une famille, mais l'on effectue un calcul sain puisque calculant des bornes peut être trop larges de la famille de probabilités sous-jacente. Mais on risque d'avoir des résultats plus imprécis que les autres méthodes, même si les calculs seront plus simples.

Le second calcul semble plus hasardeux, puis qu'il biaise la connaissance possibiliste pour la mettre sous forme de probabilité, et plus difficile (le calcul d'une fonction sur des variables aléatoires est en général plus coûteux qu'un calcul d'intervalles , fussent des intervalles flous) – mais il rend une information plus précise, une probabilité au lieu d'un intervalle de probabilité. C'est en fait l'option choisie par les statisticiens

classiques qui supposent toutes les variables incertaines représentables par des distributions de probabilité. Cette option est methodologiquement contestable (on néglige l'incomplétude des données) et pas très nouvelle.

Annexe 2

Article soumis au Journal of Environmental Engineering

A HYBRID APPROACH FOR ADDRESSING UNCERTAINTY IN RISK ASSESSMENTS

Submitted to the Journal of Environmental Engineering (September 2001)

By Dominique Guyonnet[1], Bernard Bourgine[1], Hélène Fargier[2], Bernard Côme[3] and Jean-Paul Chilès[1]

[1] Envir. Spec., BRGM, BP 6009, 45060 Orléans, Cédex 2, France
[2] Math. Spec. Université Paul Sabatier, 31063 Toulouse, France
[3] Envir. Spec., ANTEA, BP 6119, 45061 Orléans, Cédex 2, France

Key words : Risk assessment, Uncertainty, Monte Carlo, Fuzzy calculus

ABSTRACT : Uncertainty is a major aspect of the estimation, using models, of the risk of human exposure to pollutants. The Monte Carlo method, which applies probability theory to address model parameter uncertainty, relies on a statistical representation of available information. In recent years, the theory of possibilities has been proposed as an alternative approach to address model parameter uncertainty in situations where available information does not allow a representative statistical analysis, due in particular to data scarcity. In practice, it may occur that certain model parameters can be reasonably represented by probability distributions, because there is sufficient data available to substantiate such distributions by statistical analysis, while others are better represented by fuzzy numbers (due to data scarcity). The question then arises as to how these two modes of representation of model parameter uncertainty can be combined for the purpose of estimating the risk of exposure. In this paper an approach (termed a hybrid approach) for achieving such a combination is proposed, and applied to the estimation of human exposure, via vegetable consumption, to cadmium present in the surficial soils of an industrial site located in the north of France. The application illustrates the potential of the proposed approach, which allows the uncertainty affecting model parameters to be represented in a fashion which is consistent with the information at hand.

INTRODUCTION

As risk assessments become increasingly used as aids in the decision-making process related to the management of contaminated land, the issue of uncertainty is of primary concern. Guyonnet et al. (1999) compared two alternative methods for addressing model parameter uncertainty in risk assessments : the popular Monte Carlo method (see for example Vose, 1996) which is based on probability theory, and fuzzy calculus which applies the theory of possibilities (Zadeh, 1965, 1978). While no judgement of value was made regarding the intrinsic validity of one approach compared to another, it was suggested that selection of the mode of representation of the uncertainty affecting a given model parameter should be consistent with the information which is available regarding that parameter. If a variable is represented by a probability distribution function (PDF), then it is implicit that its population has been sufficiently sampled as to obtain a meaningful statistical representation. In field situations pertaining to the risk of human exposure to soil pollutants, however, this is often

not the case. Information may be scarce and imprecise in which case uncertainty representation using fuzzy numbers may sometimes be more appropriate. One question addressed in Guyonnet et al. (1999) was : « well what does it change in practice ?». The main consequence in an environmental or health context is that the *a priori* assumption of PDF's, without justification by available information, may lead to an unconservative estimation (minimisation) of risk. In order to understand this, one must recall that the probability that two independent events A and B should occur simultaneously, is the product of the probabilities of both events. Therefore, during Monte Carlo random sampling, scenarios that combine low probability parameter values have all the less chance of being selected. If a very large number of iterations is used, these scenarios will be realised, but with very low relative frequencies. When the results of the Monte Carlo analysis are compared to an acceptance criterion (for example a reference daily pollutant dose), for a certain probability level (for example 95%), these scenarios will be eliminated because they fall within the 5% high outliers. Had model parameter uncertainty been considered in terms of possibilities rather than probabilities, these low-likelihood scenarios might not have been discarded, because fuzzy calculus does not transmit through multiplication the uncertainty of the parameter values onto that of the calculation result.

In practice, it often occurs that certain model parameters can be justifiably represented by PDF's, because the data exists to substantiate these probability distributions, while others are more adequately represented by fuzzy numbers. The question then arises as to how these two modes of representation of uncertainty can be combined in the assessment of risk. This paper proposes an approach for performing such a combination, and applies it to the estimation of the risk of human exposure to cadmium present in the surficial soils of an industrial site in the north of France. The primary purpose of this paper is to promote a methodology which is consistent with the information at hand. Too often in the past, Monte Carlo analyses have been reported where the model parameter PDF's have been literally « pulled out of a hat ».

THEORY

It is reminded that a fuzzy number describes the relationship between an uncertain quantity X, and a so-called membership function μ (between 0 and 1) which represents the degree of likelihood that X should take on a certain value x. It is somewhat analogous to a probability density function, which represents the probability that a random variable X takes on a certain value x. The probability distribution function (PDF), or cumulative probability function, represents the probability that X should take on a value lower than or equal to x. As shown in Dubois & Prade (1992), a fuzzy number can be thought of as a family of probability density functions. The reader will refer to Zadeh (1965, 1978) and Dubois & Prade (1988) for a detailed description of the theory of possibilities and of fuzzy numbers, while for example Dou et al. (1995), Bardossy et al. (1995), Freissinet et al. (1998), Guyonnet et al. (1999), Cazemier (1999) present applications to environmental problems.

Compared with a probability density function or a PDF, a fuzzy number is very poor in terms of intrinsic information. The PDF defines a variable entirely ; it applies to a system which is precisely known. A fuzzy number, on the other hand, provides information such as : the value of parameter A is likely to lie between values a_1 and a_2; values outside the range a_3 - a_4 are considered impossible. In field situations, this type of information is often a more realistic transcription of available data than are an average and a standard deviation for example.

An important consequence of combining probability distributions with fuzzy numbers is that the net result will necessarily be a fuzzy number. As the information conveyed by a fuzzy

number is poor, no mathematical procedure can compensate for this paucity and achieve the degree of system definition required by a probabilistic representation. Different approaches for combining probabilistic and possibilistic uncertainties can be defined. The approach described below, called the « hybrid approach », is relatively pragmatic and intuitively amenable to the practising environmental engineer.

In order to illustrate the hybrid approach, we will consider the estimation of a dose resulting from the exposure of a human target to soil pollutants. This dose is calculated using a « model », M, which is a function of a certain number of parameters :

$$Dose = M(P_1, \ldots, P_n, F_1, \ldots, F_m), \quad (1)$$

where M = model; $P_1, \ldots, P_n = n$ model parameters each represented by a PDF; $F_1, \ldots, F_m = m$ model parameters each represented by a fuzzy number. Note that the model can also involve precise (i.e. « crisp ») parameter values. Calculation of the dose is performed by combining the Monte Carlo random sampling technique, with the method of α-cuts (Dubois and Prade, 1988) for fuzzy calculus. The combination procedure is summarised below :

1. Generate n random numbers $(\chi_1, \ldots, \chi_n)$ from a uniform distribution and sample the n PDF's to obtain a realisation of the n random variables : $p_1, \ldots p_n$ (Fig. 1A)
2. Select a value α of the membership function (a level of possibility).
3. Calculate the *Inf* (smallest) and *Sup* (largest) values of $M(p_1, \ldots, p_n, F_1, \ldots, F_m)$, considering all values located within the α-cuts for each fuzzy number (see Fig. 1.B).
4. Affect these *Inf* and *Sup* values to the lower and upper limits of the α-cut of $M(p_1, \ldots, p_n, F_1, \ldots, F_m)$.
5. Return to step 2 and repeat steps 3 and 4 for another α-cut (note : α can be increased stepwise from 0 to 1 every 0.1 increments). The fuzzy result of $M(p_1, \ldots, p_n, F_1, \ldots, F_m)$ (the fuzzy dose) is obtained from the *Inf* and *Sup* values of $M(p_1, \ldots, p_n, F_1, \ldots, F_m)$ for each α-cut.
6. Return to step 1 to generate a new realisation of the random variables.

If steps 2 through 5 are repeated ω times, ω fuzzy doses are calculated (Fig. 1.C). For each value of the membership function (each value of α), the spread between the *Inf* and *Sup* values of the fuzzy results (see Fig. 1.C) is entirely a consequence of the Monte Carlo random sampling. It is therefore proposed to select the final *Inf* and *Sup* values of $M(P_1, \ldots, P_n, F_1, \ldots, F_m)$, for each value of α, by building a histogram of cumulative relative frequencies of the *Inf* and *Sup* values, and extracting the final *Inf* and *Sup* values for a certain level of probability. This is illustrated in Fig. 2. For each level of α, the histograms reproduce the spread of the *Inf* and *Sup* values. The final *Inf* and *Sup* values are taken such that there is a 5% probability of having values lower of higher respectively. This final fuzzy dose can be compared to an acceptance criterion, using the tools provided by possibility theory, as illustrated in the next section.

Note that in the general case, step 4 of the procedure above can be performed using a minimisation and maximisation algorithm. However, if the model is an equation involving simple operations such as multiplication and substraction, the *Inf* and *Sup* values (and their final values) can be identified directly.

APPLICATION TO AN INDUSTRIAL SITE

The proposed method is applied to a metallurgical industrial site located in the north of France. The surficial soils of this site are contaminated by a number of metallic pollutants among which cadmium, which have been deposited by the smoke emanating from a chimney located on the site. The quality of the surficial soils (upper ten centimeters) has been monitored in detail, and there is a relatively large amount of data available. These data are primarily total soil metal contents measured by atomic absorption spectrophotometry after extraction by fluorhydric acid. No selective extraction data (see for example Tessier et al., 1979) were available for this study. This detail is of importance since we are interested here in the exposure of a human target to cadmium in the soil via the consumption of vegetables. A large body of scientific evidence shows that there is poor correlation between the amount of cadmium absorbed by plants, and the total amount present in the soil (see for example Jopony and Young, 1993, Lorenz et al., 1997). Metal uptake by plants depends on a variety of factors among which metal speciation, plant specie, pH (Dijkshoorn et al., 1983, Singh et al., 1995), redox conditions, humidity, temperature (Chang et al., 1987), competition with other metals (Smilde et al., 1992, Chaney et al., 1999), etc. The variety of these factors explains why it is difficult in practice to develop a model of metal uptake by plants based on total soil concentrations, and therefore preferable to have direct site-specific measurement of metal uptake. This is fortunately the case for this industrial site : measurements of cadmium content were performed by Luttringer and de Cormis (1979) for a limited number of vegetables (in particular leeks) grown in the immediate vicinity of the soil sampling points. Measurements were performed on the edible vegetable parts, after they had been washed to eliminate the metal fraction present in the dust at the surface of the leaves. Cadmium was analysed by atomic absorption after plant calcination and attack by chlorhydric and fluorhydric acid.

Due to the relatively large number of surficial soil analyses (124), the spatial distribution of cadmium in the surficial soils can be investigated with geostatistical methods (see Chilès and Delfiner, 1999). Geostatistics is a special branch of statistics which applies to data which display a spatial structure. Its use is not compulsory for the hybrid approach proposed in this paper, but serves here to provide statistically representative estimators of soil cadmium concentration. As could be expected, the data show a decrease in soil cadmium concentrations with increasing distance to the chimney (see Fig. 3 where Cd concentrations are in logarithm). In order to apply the geostatistical tools (the variogram and kriging), the data were first transformed into logarithm and then decomposed into a trend and a residual around this trend, according to :

$$\ln(\mathrm{Cd}_s) = \ln(\mathrm{Cd}_s)_T + R \qquad (2)$$

where $\ln(\mathrm{Cd}_s)$ = logarithm of measured soil cadmium concentrations; $\ln(\mathrm{Cd}_s)_T$ = predicted values of $\ln(\mathrm{Cd}_s)$ according to the trend; R = residual around this trend. A correlation equation which describes the trend is (Fig. 3) :

$$\ln(\mathrm{Cd}_s)_T = -0.3 + 4.2\exp(-d/2) \qquad (3)$$

where d = distance to the chimney. The spatial distribution of the residual is then examined using the variogram. In a first step, directional variograms were calculated because it was anticipated that wind direction could have an influence on the spatial distribution of cadmium. But these directional variograms showed no clear anisotropy and therefore an omnidirectional variogram was used. It was fitted with a linear variogram model and a nugget effect

accounting for microstructures and/or measurement errors (Fig. 4). The next step consisted in using this variogram model to interpolate the residual R, by kriging. Note that an interpolated value is not a "true" value, but kriging ensures that the interpolation is not biased (on average, the kriging error is zero) and that it is optimal (the kriging variance is minimum). Noting the kriged residual as R_K, we obtain the soil cadmium concentration from :

$$Cd^*_{sM} = \exp\left(R_K + \ln\left(Cd_s\right)_T\right) \tag{4}$$

where Cd^*_{sM} is a median estimator of soil cadmium concentration, because kriging of R yields a median estimator of R (i.e., the true value has a 50% probability of being lower than the kriged value). Results depicted in Fig. 5 show relatively high values of soil cadmium concentrations close to the chimney (up to 27 ppm), and a gradual decrease to values below 1 ppm at a distance of approximately 4 km.

Since kriging achieves the minimisation of the estimation variance, a by-product is the kriging variance, or its square root ; the kriging standard deviation σ_K. If the kriging error (i.e., the difference between the kriged and true values) is assumed Gaussian, confidence intervals can be deduced from σ_K. We write :

$$R < R_K + t\sigma_K \tag{5}$$

where t is a factor which depends on the level of confidence assigned to R. For a level of confidence of 95%, for example, $t = 1.65$: R has 95% chances of being lower than $R_K + 1.65\ \sigma_K$. An estimation of soil concentration is obtained by combining Equations (2) and (4) :

$$Cd_s < \exp\left(R_K + \ln\left(Cd_s\right)_T\right) \cdot \left(\exp(\sigma_K)\right)^t \tag{6}$$

Since the first term in Equation (6) is the median estimator for Cd_s (Equation 4), $\exp(\sigma_K)$ can be considered as a multiplicative standard deviation. It is called here an "error factor" for short, and is presented in Fig. 6. This map of the error factor shows values up to 1.3 with hollows centred around measurement points. At these points, the remaining uncertainty is due to the measurement errors. The statistical soil cadmium concentrations which result from this analysis are combined below with fuzzy uncertainties related to the uptake of cadmium by vegetables and to the absorption of a dose by a human target.

Measured values of cadmium in leeks are plotted in Fig. 7 as a function of measured soil concentrations at the locations where the leeks were grown. As seen in this figure, the small number of measurements (five) hardly warrants a statistical analysis. They nevertheless provide very valuable information as they represent site-specific values of metal uptake by vegetables. This information can be analysed using a fuzzy correlation. Based on the shape of the measured data, the following correlation equation is defined :

$$Cd^*_{pl} = Or + (As - Or)\left(1 - \exp\left(-k\ Cd_s\right)\right) \tag{7}$$

where Cd_{pl}^* = estimated cadmium concentration in the plant (mg Cd / kg dry plant); Cd_s = measured cadmium concentration in the soil (mg Cd / kg dry soil); Or = Cd_{pl}^* at the origin ($Cd_s = 0$); As = Asymptote (Cd_{pl}^* at large values of Cd_s); k = parameter which controls the rate of increase. Parameters Or, As and k in Equation (7) are represented by fuzzy numbers, while the Cd_s values are represented by probability density functions defined by the geostatistical analysis. The choice of the fuzzy numbers was dictated by the measured data. In

Fig. 7, the dashed lines (from Equation 7) represent « likely » boundaries for cadmium uptake by leeks. These boundaries imply that if someone were to go and measure the concentration in leeks cultivated on this site, it is likely that he/she would obtain a value which falls within these boundaries. The full line represents what is considered as a reasonably conservative upper boundary for cadmium concentration in leeks grown on this site. While one could argue regarding the precise positions of these curves, it should be noted that this representation is at least consistent with the measured data, and that it can be easily adjusted to accommodate input from agronomic experts. The fuzzy numbers for parameters *Or*, *As* and *k* are presented in Fig. 8 (along with another parameter discussed below). The limits of these fuzzy numbers are deduced directly from Fig. 7. For example the cadmium content of the plant at the origin ($Cd_s = 0$) is considered likely to be between 0 and 0.3 (these values have a likelihood of 1 in Fig. 7.A) while a value of 0.6 is considered as a possible upper boundary (in Fig. 7.A, the likelihood of values above 0.6 is considered nil).

The objective of the calculation is to estimate a dose of cadmium absorbed by a human target. In fact, doses are calculated along a grid which covers the site in order to examine the spatial distribution of absorbed dose. The dose is calculated from :

$$Dose = \frac{Cd_{pl}^{*} \cdot 1000 \cdot Con \cdot DMC}{BW} \tag{8}$$

where Dose = Absorbed dose (μg Cd per day per kg body weight); *Con* = Leek daily consumption (kg leek per day); *DMC* = Leek dry matter content (weight percent); *BW* = Human target body weight (kg). The leek dry matter content (*DMC*) accounts for the fact that daily consumption is provided with respect to wet weight while the cadmium concentration in the leeks (Cd_{pl}^{*}) is relative to dry weight. The cadmium concentration in the leeks is obtained from the fuzzy correlation equation (Equation 7). As a simplifying hypothesis, it is considered that leeks are representative of vegetables with respect to human exposure to cadmium through vegetable consumption. The daily vegetable consumption and the vegetable dry matter content are selected based on data presented in INERIS (1999). Likely daily vegetable consumption is taken between 100 and 120 g/day, while an upper possible limit is taken as twice the higher value (240 g/day). The vegetable dry matter content (*DMC*) and the body weight are considered as constant : 15% and 70 kg respectively.

An example calculation performed with the hybrid approach is presented for a median value of Cd_{sM}^{*} = 7.97 ppm (from Fig. 5), and its corresponding error factor (Fig. 6) = 1.15. The probability distribution for Cd_s can be obtained by calculating a normal Gaussian distribution for mean = ln(7.97) and standard deviation = ln(1.15), and then taking the exponential of the results to return to the distribution on Cd_s. The hybrid approach can then be performed according to Figs. 1 and 2 (which represent the general case). However, as there is only one probabilistic variable involved, and because Equation (8) is very simple (minima and maxima can be identified directly), the calculation can also be performed without Monte Carlo random sampling. Using the values for Cd_{sM}^{*} and the error factor above, the maximum value of Cd_s for a 95 % confidence level is obtained from Equation (6) for $t = 1.65$: $Cd_{s\,max}$ = 10.04 ppm. Likewise, the minimum value of Cd_s is obtained for $t = -1.65$: $Cd_{s\,min}$ = 6.33 ppm. These values then serve in Equations (7) and (8) to obtain the *Inf* and *Sup* values of the Dose, for each value of α (level of likelihood), using the relevant *Inf* and *Sup* values of the fuzzy numbers involved in Equation (8). The result is depicted in Fig. 9. The calculation was performed for a large number of points using the same grid as the one used to generate Fig. 5.

The next step consists in examining the acceptability of the calculated doses. This acceptability is considered here with respect to a maximum reference dose. According to WHO (1994), the kidney is the main target of cadmium toxicity. In order to maintain

cadmium concentrations in the kidney cortex below 50 mg/kg, WHO (1994) recommends that cadmium absorption via food consumption should not exceed 1 μg per day and per kg body weight. The calculated fuzzy doses were compared to this daily reference dose (noted D_o) using the measure of possibility (see Dubois and Prade 1988 or Guyonnet et al., 1999) for the proposition : calculated fuzzy dose F exceeds the reference dose D_o. For a "crisp" reference dose, the measure of possibility is written :

$$\Pi(F > D_o) = \underset{u>D_o}{\text{Sup}}\, \mu_F(u) \qquad (9)$$

where $\mu_F(u)$ = membership function of F for any value u; Sup = the largest value. Fig. 10 provides a graphical illustration of Π for such a proposition. As long as the fuzzy dose is entirely below the reference dose, the possibility of D_o being exceeded is considered nil (Π = 0 ; Fig. 10.a). As the fuzzy dose intersects the reference dose, excess of D_o is considered possible with a possibility measure $\Pi = \alpha$ (Fig. 10.b). Once the reference dose intersects the plateau (Fig. 10.c), Π becomes equal to 1. Note that the latter case does not imply that the reference dose will be exceeded with « certainty ». Reasonable certainty occurs when another indicator of the validity of the proposition (i.e. the measure of necessity ; Dubois and Prade 1988) becomes equal to 1. Applying the possibility indicator to the calculated fuzzy doses, we obtain the spatial distribution of the possibility of exceeding the reference dose. This spatial distribution is depicted in Fig. 11. Possibilities of 0.45 of exceeding the reference dose are found in the close vicinity of the chimney, but decrease below 0.1 at a couple of kilometers.

DISCUSSION AND CONCLUSIONS

The map of Fig. 11 can be used as an aid in the decision-making process related to the management of this industrial site. It would be incumbent, however, upon the competent sanitary authority, to define which level of possibility of reference dose excess should be considered tolerable. Depending upon the context, values around 0.1-0.2 may seem consistent with a reasonable application of the precautionary principle. To require a possibility of zero will in many cases be too strict, and may result for example in excessive areas of land being ruled out for certain uses, or in excessive cleanup costs. It is reminded that the fuzzy calculus component of the proposed hybrid model considers all possible combinations of fuzzy parameter values, and does not transmit through multiplication the uncertainty of these parameter values onto that of the calculation result. It is more conservative than a purely Monte Carlo approach, and therefore constraints on acceptance criteria in terms of possibilities need not be as strict as in terms of probabilities.

In the preceding section, Fig. 7 illustrated how a fuzzy correlation could be established based on scarce data. Initially, a linear correlation equation was used, which was later considered to be unsatisfactory as it assumed that as soil concentration increased, plant uptake also increased in direct proportion. This is analogous to using a « bioconcentration factor » such as appears in several risk analysis tools (for example HESP; see Poels et al., 1990). But agronomists argued that plant uptake could not increase indefinitely and that in fact it will not exceed a certain level. Therefore a log-shaped correlation seemed more appropriate, and consistent with the data depicted in Fig. 7. This illustrates what, in our view, is an important asset of the fuzzy approach : it provides considerable flexibility for incorporating the know-how of experts in the relevant scientific fields. When dealing with multidisciplinary problems such as those which appear in an environmental context, this flexibility is of great value. The statistical approach is comparatively more rigid ; while issues such as multi-modal distributions, skewness etc. can be taken into account, the use of means and standard

deviations leaves little scope for "soft" information input from experts in the relevant fields. Yet experience shows that with respect to environmental issues, adequate decisions are best achieved when such input is sought.

In this paper a hybrid approach was proposed for combining probabilistic and possibilistic representations of model parameter uncertainty. As stated in Guyonnet et al. (1999), if data is available which substantiates a statistical representation of parameter value variability, then such a representation should certainly be preferred. Possibility theory is proposed as an alternative tool, to assist in situations where such data is not available. It is felt that to force-fit probability distribution functions on data without statistical justification, is not only scientifically unrigorous, but may also introduce serious biases in the decision-making process. The proposed hybrid approach presents the double advantage of preserving the strengths of statistical analysis, while providing the flexibility of the fuzzy approach when such an analysis is not substantiated by the data. It is presented as a means to improve consistency between calculation hypotheses and available information. It is felt that in the field of risk assessment, where there can hardly be high claims on the precision of model predictions, such consistency should be promoted.

In the application example presented above, the hybrid approach was applied in a forward mode to evaluate the possibility of exceeding a tolerable reference dose. An interesting issue is the inverse problem : « Given a tolerable reference dose, and considering the uncertainty (probabilistic or possibilistic) affecting model parameter values, which residual soil concentration guarantees that the possibility of the reference dose being exceeded, is below the value fixed by the competent sanitary authority ? ». This issue, which is of particular interest to risk-based corrective action, will be the topic of a forthcoming paper.

Acknowledgements

We thank Métalleurop France for kindly providing some of the soil data analysed in this paper, and Didier Dubois of the University of Toulouse for helpful advice. This study was funded in part by the French Ministry of the Environment, within the GESSOL Programme led by the French National Institute for Agronomic Research (INRA). Funding was also provided by the BRGM's Directorate of Research Activities.

APPENDIX I. REFERENCES

Bardossy, A., Bronstert, A., and Merz, B., (1995). "1-, 2- and 3-dimensional modeling of groundwater movement in the unsaturated soil matrix using a fuzzy approach." *Advances in Water Resources*, 18(4), 237-251.

Cazemier, D. (1999). " Utilisation de l'information incertaine dérivée d'une base de données sols. (Use of uncertain information derived from a soil data-base). " *Ph. D. Thesis of the National School of Agronomy*, Montpellier (France). 170 pp.

Chaney, R., Brow, S., Stuczynski, T., Daniels, W., Henry, C., Li, Y.-M., Siebielec, G, Malik, M., Angle, J., Ryan, J., and Compton, H. (1999) – " In-situ remediation and phytoextraction of metals from hazardous contaminated soils. " *In* : Innovative clean-up approaches : Investments in technology developments, results & outlooks for the future. Nov.2-4, Bloomingdale, USA. 29 pp.

Chang, A., Page, A., and Warneke, J. (1987) – " Long-term sludge application on cadmium and zinc accumulation in Swiss chard and radish. " *Journal of Environmental Quality*, 16, 217-221.

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

Dijkshoorn, W., Lampe, J., and VanBroekhovern L. (1983) – " The effect of soil pH and chemical form of nitrogen fertilizer on heavy metal contents in ryegrass. " *Fertilizer Research*, 4 :63-74.

Dou, C., Woldt, W., Bogardi, I., and Dahab, M. (1995). "Steady-state groundwater flow simulation with imprecise parameters." *Water Resources Research,* 31(11), 2709-2719.

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

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

Freissinet, C., Erlich, M., and Vauclin, M. (1998) - " A fuzzy logic-based approach to assess imprecision of soil water contamination modelling. " *Soil & Tillage Research*, 47, 1-17.

Guyonnet, D., Côme, B., Perrochet, P., and Parriaux, A. (1999) – " Comparing two methods for addressing uncertainty in risk assessments. " *Journal of Environmental Engineering*, 125(7), 660-666.

INERIS (1999) – " Méthode de calcul des valeurs de constat d'impact dans les sols. (Method for calculating risk-based concentration limits in soils). " *INERIS Unpublished Report*, April 1999, Verneuil-en Halatte, France.

Jopony, M., and Young, S. (1993) – " Assessment of lead availability in soils contaminated by mine spoil. " *Plant and Soil*, 151, pp. 273-278.

Lorenz, S., Hamon, R., Holm, P., Domingues, H., Sequeira, E., Christensen, T., and McGrath, S. (1997) – " Cadmium and zinc in plants and soil solutions from contaminated soils. " *Plant and Soil*, 189:21-31.

Luttringer, M., and de Cormis, L. (1979) – " La pollution par les métaux lourds à Noyelles-Godault et ses environs (Pas de Calais). (Pollution by heavy metals at Noyelles-Godault and surrounding area, Pas de Calais). " *Unpublished report of the National Institute for Agronomic Research (INRA)*, Montfavet (France), 12 pp.

Poels, C., Gruntz, U., Isnard, P., Riley, D., Spiteller, M., ten Berge, W., Veerkamp, W, Bonyinck, W., (1990) – " Hazard assessment of chemical contaminants in soil ". ECETOC Report No. 40, ISSN-0773-8072-40. European Chemical Industry Ecology & Toxicology Center, Brussels, Belgium.

Singh, B., Narwal, R., Jeng, A., and Almas, A. (1995) – " Crop uptake and extractability of cadmium in soils naturally high in metals at different pH levels. " *Commun. Soil Sci. Plant Anal.*, 26(13&14), 2123-2142.

Smilde, K., Van Luit, B., and Van Driel, W. (1992) – " The extraction by soil and absorption by plants of applied zinc and cadmium. " *Plant and Soil*, 143, 233-238.

Tessier, A.P., Campbell, G.C., and Bisson, M. (1979) – " Sequential extraction procedure for speciation of particulate trace metals. " *Analytical Chemistry* 51, 844-850.

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

WHO (1994) – " Quality directives for drinking water. Volume 1 : recommendations. 2nd Edition. " *World Health Organisation.* Geneva, Switzerland. ISBN 92 4 254460 4, 202 pp.

Zadeh L. (1965). "Fuzzy Sets." *Information and Control,* 8, 338-353.

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

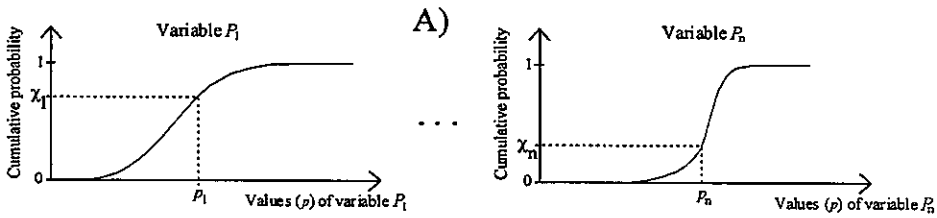
APPENDIX II. NOTATION

The following symbols are used in this paper :

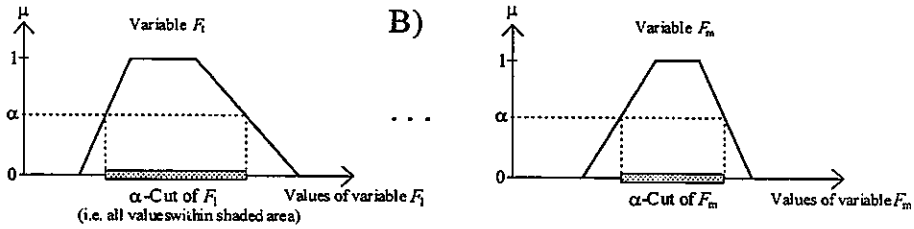
As = Asymptote for Equation (7);
BW = body weight;
Cd_{pl} = measured cadmium concentration in the plant;
Cd_{pl}^{*} = estimated cadmium concentration in the plant;
Cd_s = measured cadmium concentration in the soil;
Cd_{sM}^{*} = median estimator of cadmium concentration in the soil;
Con = leek daily consumption;
d = distance to the chimney;
D_o = daily reference dose;
DMC = Leek dry matter content (weight percent);
Inf = smallest value;
$F_1, \ldots, F_m = m$ model parameters each represented by a fuzzy number;
k = Constant which controls the rate of increase in Equation (7);
$\ln(Cd_s)$ = logarithm of measured cadmium concentration in the soil;
$\ln(Cd_s)_T$ = value of $\ln(Cd_s)$ predicted by a regression equation (trend);
M = model;
$Or = Cd_{pl}$ at the origin in Equation (7);
$P_1, \ldots, P_n = n$ model parameters each represented by a PDF;
PDF = probability density function;
R = residual;
Sup = largest value;
α = value of the membership function μ;
α-cut = all values of parameter F within shaded area in Fig. 1B;
$\mu_F(u)$ = membership function of F for any value u;
χ = random number;
σ_K = kriging standard deviation for R;
ω = number of Monte Carlo iterations;
Π = possibility measure.

FIGURES

Generate n random numbers (χ) to sample the n PDF's : obtain n values $p_1, \ldots, p_n$.



Apply method of α-cuts : look for *Inf* and *Sup* values of $M(p_1, \ldots, p_n, F_1, \ldots, F_m)$ on the α-cuts. Build fuzzy number of $M(p_1, \ldots, p_n, F_1, \ldots, F_m)$.



Repeat the procedure ω times : obtain ω fuzzy results of $M(P_1, \ldots, P_n, F_1, \ldots, F_m)$.

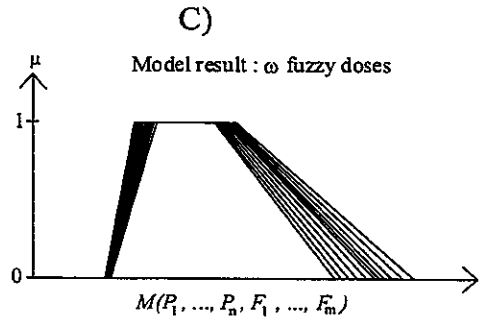


FIG. 1. Schematic illustration of the hybrid approach

Select a value of α and build the cumulative relative frequency diagrams of the *Inf* and *Sup* values of $M(p_1, \ldots, p_n, F_1, \ldots, F_m)$.

Select final *Inf* and *Sup* values of $M(p_1, \ldots, p_n, F_1, \ldots, F_m)$ for 95% confidence level.

Obtain final fuzzy result of $M(p_1, \ldots, p_n, F_1, \ldots, F_m)$ by repeating for each value of α.

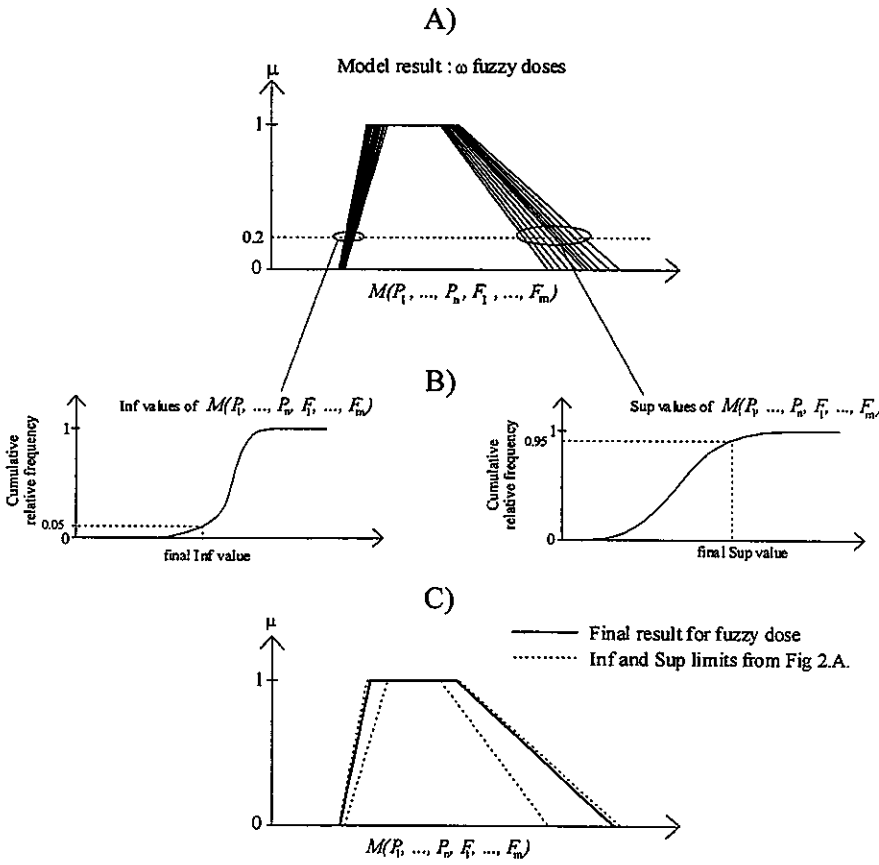


FIG. 2. Selection of the final fuzzy result

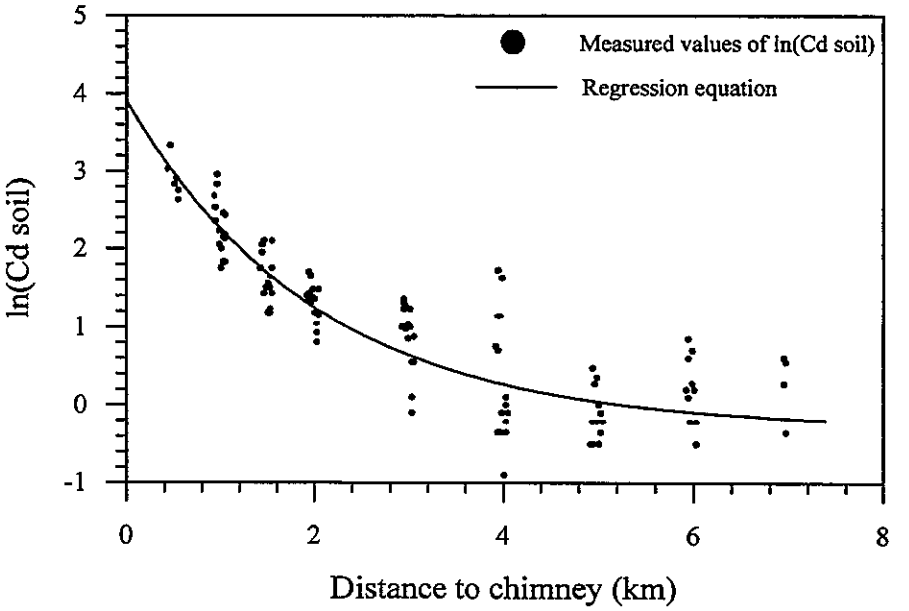


FIG. 3. Correlation between the logarithm of Cd soil concentrations, and distance between the chimney and the sampling point

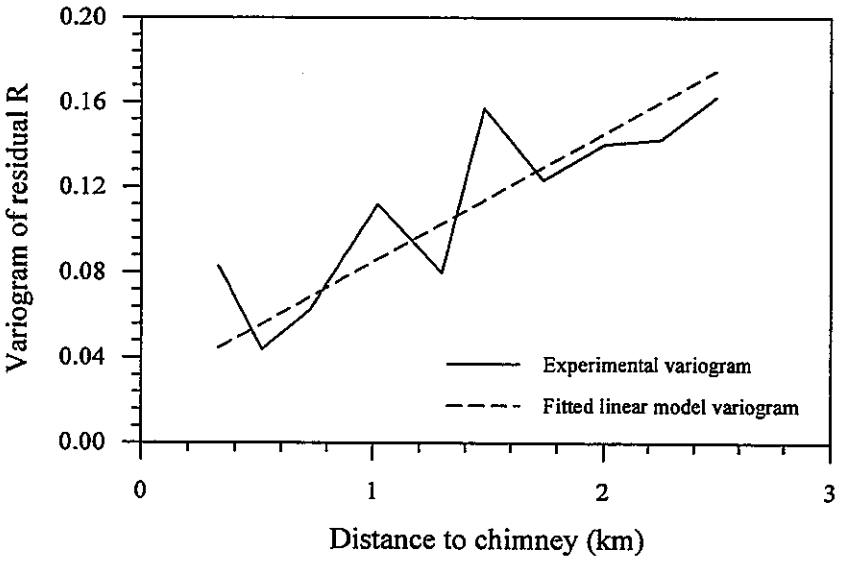


FIG. 4. Variogram of residual *R*. Fit with a linear variogram model

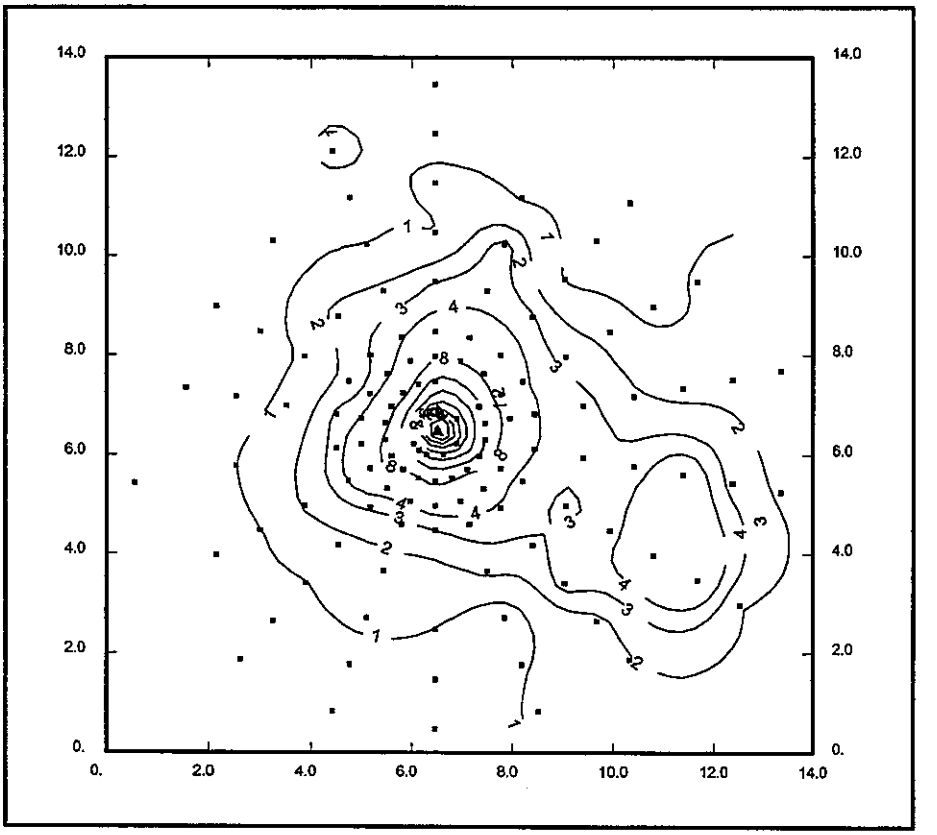


FIG. 5. Interpolated map of soil cadmium concentrations (Cd_s^* ; ppm). Graduation in km; Triangle = chimney location; Points = measurement points.

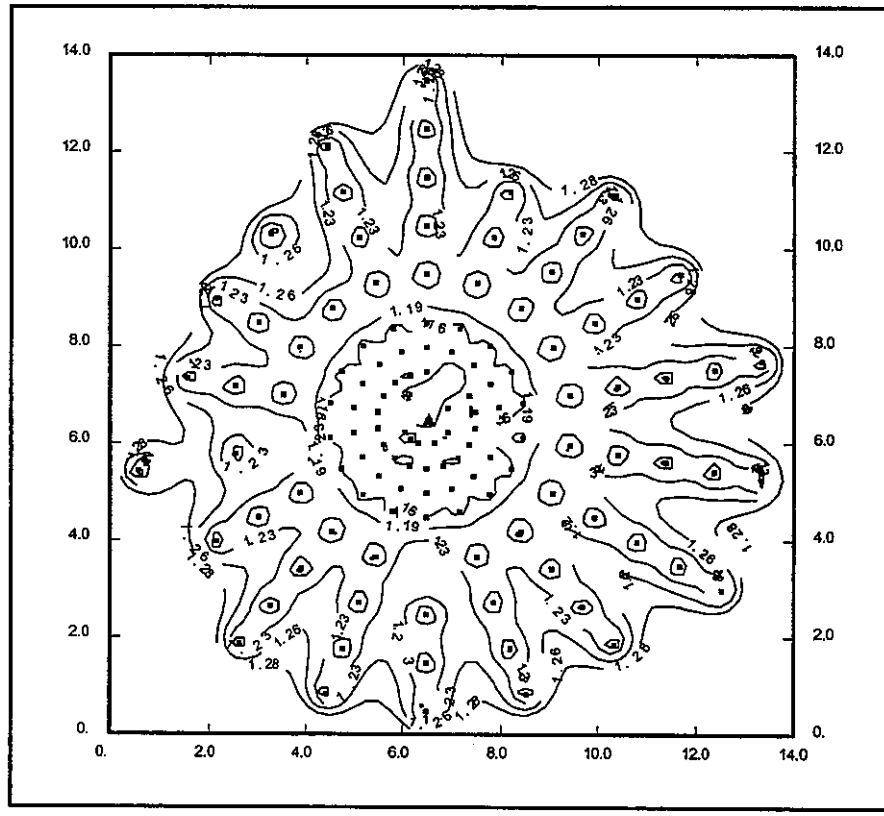


FIG. 6. Spatial distribution of the error factor, $\exp(\sigma_K)$, on the residual *R*. Graduation in km; Triangle = chimney location; Points = measurement points.

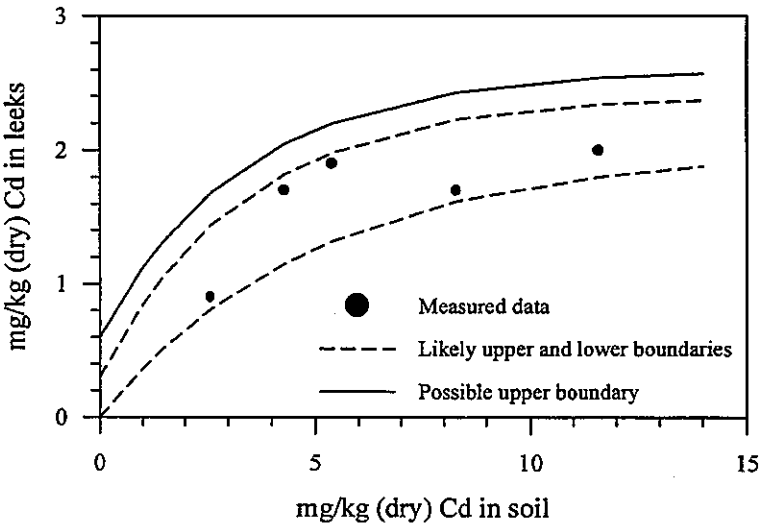


FIG. 7. Cadmium concentrations measured in leeks versus measured soil concentrations, and fuzzy correlation

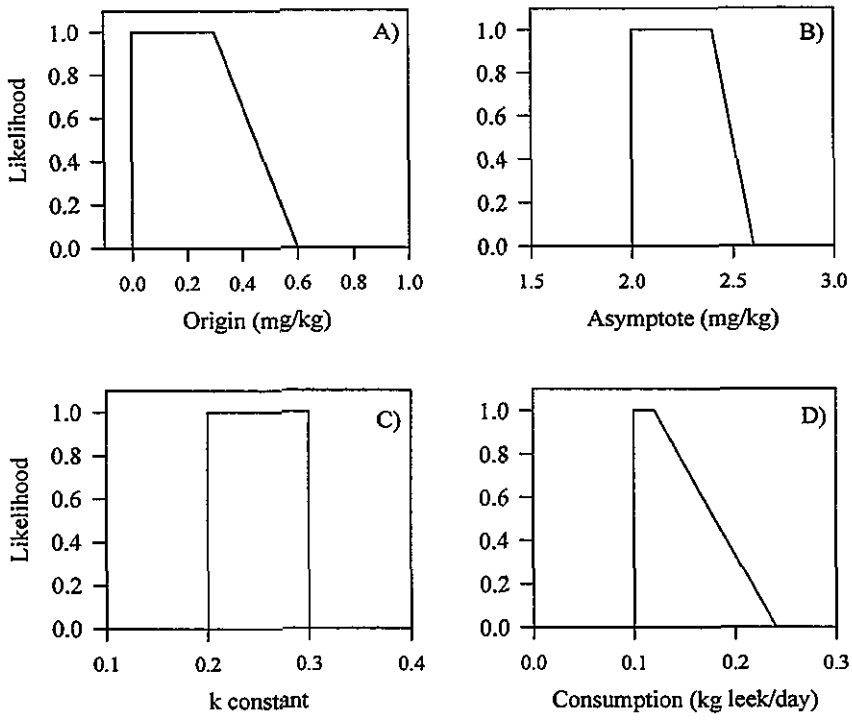


FIG. 8. Fuzzy numbers for several model parameters

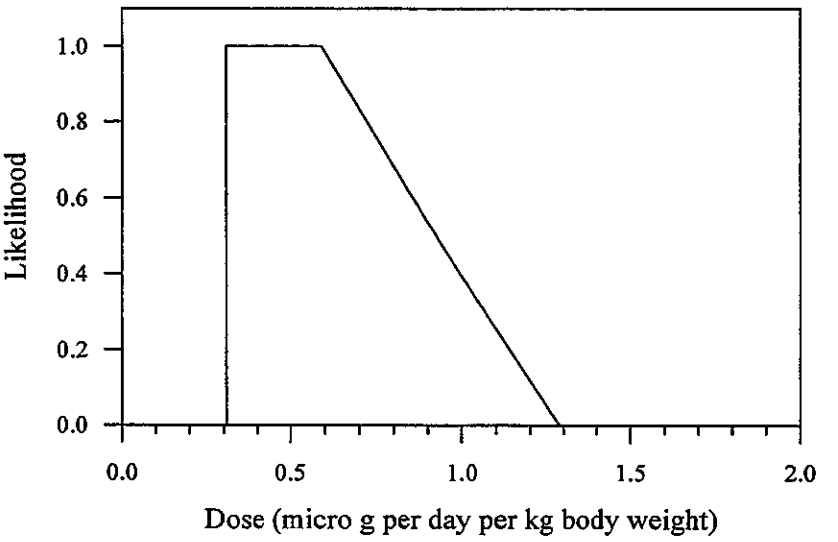


FIG. 9. Calculated dose for Cd_{sM}^{*} = 7.97 mg/kg, error factor = 1.15 and 95% confidence level on estimated soil cadmium concentration.

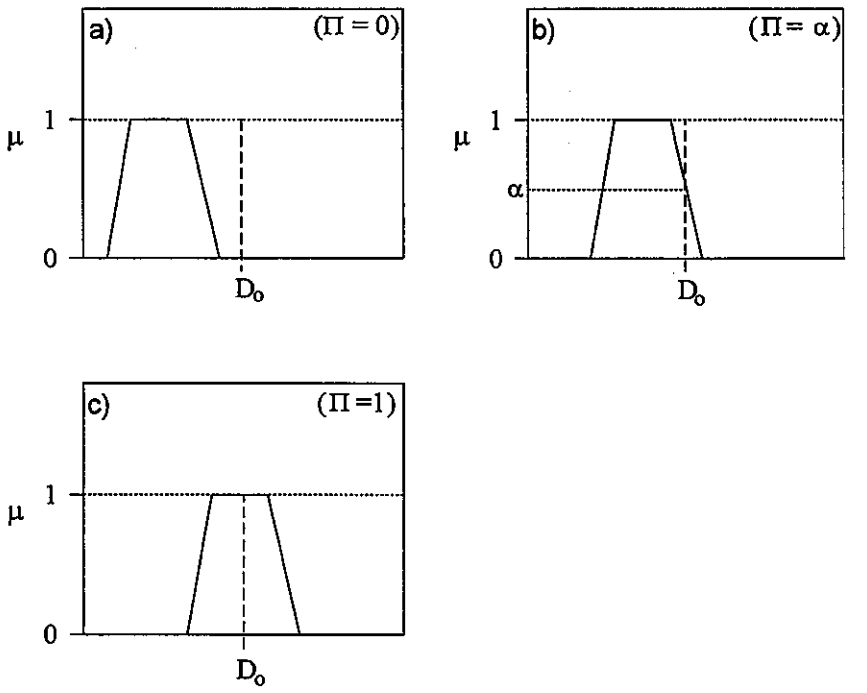


FIG. 10. Comparison between a fuzzy dose and a reference dose (D_o)

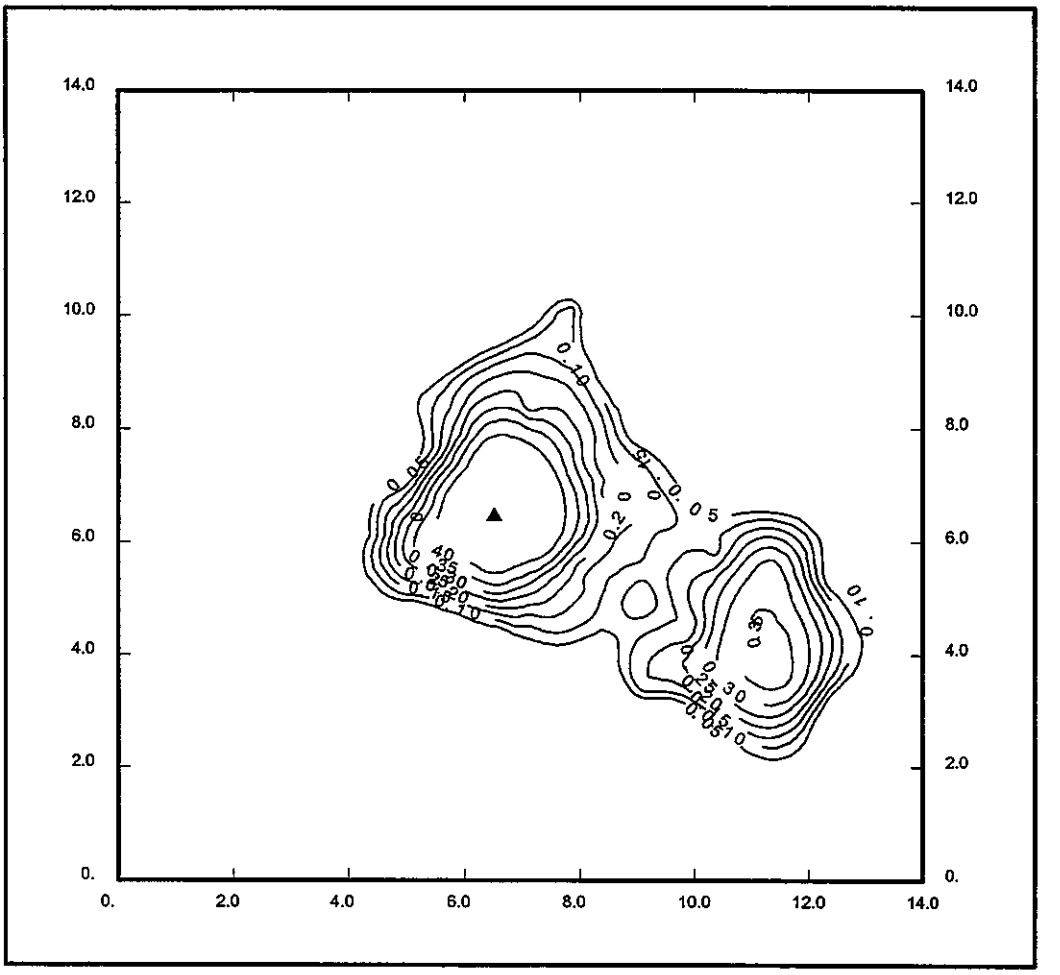


FIG. 11. Map of the possibility that the absorbed dose should exceed the reference dose (1 μg/d kg^{-1}). Graduation in km; Triangle = chimney location.

BRGM
SERVICE EPI
Unité SIS
BP 6009 – 45090 Orléans cedex 2 – France –Tél. : 33 (0)2 38 64 34 34