HAL
open science

Traitement d'imperfections des indicateurs agri-environnentaux

Zainab Assaghir, Philippe Girardin, Amedeo Napoli

To cite this version:

Zainab Assaghir, Philippe Girardin, Amedeo Napoli. Traitement d'imperfections des indicateurs agri-environnentaux. Manifestation des Jeunes Chercheurs en Sciences et Technologies de l'Information et de la Communication - MajecSTIC 2008, Oct 2008, Marseille, France. hal-00608814

HAL Id: hal-00608814

https://hal.science/hal-00608814v1

Submitted on 15 Jul 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Traitement d'imperfections des indicateurs agri-environnentaux

Application de la théorie de possibilités pour le calcul d'un indicateur agri-environnemental

Zainab Assaghir[*,**] **– Philippe Girardin**[*] **– Amedeo Napoli**[**]

** INRA Nancy-Colmar*
28 rue Herrlisheim – B.P. 20507– F-68021 Colmar
*** LORIA Nancy*
Campus Scientifique – B.P. 239– F-54506 Vandoeuvre-lès-Nancy
{assaghiz,napoli}@loria.fr–{zassaghir,girardin}@colmar.inra.fr

RÉSUMÉ. *Les indicateurs développés et utilisés en agronomie sont généralement calculés à partir de données et d'informations entachées d'imperfections. Dans cet article, on présente le calcul d'un indicateur agri-environnemental–fondé sur un système de règles de décision–en utilisant une méthode de fusion de la théorie des possibilités au sein de la notion de sous-ensemble maximaux cohérents et une arithmétique sur des intervalles pour calculer les bornes de l'indicateur. La modélisation de données est réalisée dans le cadre de la théorie des possibilités qui fournit des outils de représentation de l'imperfection. Cet article présente les résultats de l'indicateur obtenues par calcul d'intervalles et nous présentons la note de l'indicateur par une méthode de défuzzyfication.*

ABSTRACT. *The indicators developed and used in agronomy are generally calculated from starting data and information corrupted with imperfections. In this paper, We present the calculation of an environmental indicator–based on a system based on fuzzy rules–using notion of maximal coherent subsets on quantitative possibility distributions. Data are represented in the context of the theory of possibilities that provides tools representation of imperfection. This paper presents results by intervals and We introduced the note of indicator using defuzzyfication method.*

MOTS-CLÉS : *Théorie de possibilités, fusion, sous-ensembles maximaux cohérents, indicateur.*

KEYWORDS: *Possibility theory, fusion, maximal coherent subsets, indicator.*

1. Introduction

Dans le domaine de l'analyse de données, nous manipulons des données, le plus souvent numériques, qui sont censées donner une image aussi fidèle que possible de la réalité. Lorsque nous devons prendre une décision, nous devons étudier notre environnement de façon à obtenir le maximum de données pour en tirer une conclusion. En agronomie, pour aider l'agriculteur à améliorer la durabilité environnementale de ses pratiques agricoles, les chercheurs de l'INRA de Colmar développent depuis une dizaine d'années des indicateurs sous forme de notes ou d'indices (Girardin *et al.,* 2000). Ces indices sont entachées d'imperfections parce qu'ils sont calculés à partir de données provenant de plusieurs sources ou à partir des estimations ou d'évaluations approximatives. Les données sont accessibles sur le terrain et d'autres proviennent de sources d'informations. Une source d'informations peut être une base de données ou un expert. Le but de cet article est de traiter l'imperfection liée à un indicateur. Dans ce but, on s'intéresse à la théorie des possibilités (Dubois *et al.,* 1985) qui fournit des outils de représentation pour toutes les sortes d'informations. En particulier, lorsqu'il s'agit de trouver les valeurs plausibles de variables provenant de plusieurs bases de données, la théorie des possibilités offre un choix étendu entre le mode de fusion conjonctif et le mode disjonctif. Les informations collectées pour le calcul des indicateurs sont entachées d'imperfections. Ces informations sont issues de plusieurs bases de données ou elles sont estimées par des experts. D'une part, il faut calculer les intervalles flous des variables d'entrées. Ensuite, il faut utiliser ces intervalles pour calculer un intervalle de variation de l'indicateur.
Cet article est organisé comme suit : dans un premier temps, nous rappelons des notions de base de la théorie des possibilités, son intérêt dans le cadre de calcul de l'indicateur. Ensuite, nous présenterons le contexte de l'indicateur, ses objectifs, ses utilisateurs, son mode de calcul et ses variables. Dans la section 4, nous présenterons les méthode utilisées et comment elles s'appliquent dans notre cas. Enfin, nous présenterons les résultats d'une application de deux matières actives sur une parcelle et on présente également le résultat par défuzzyfication.

2. Théorie de possibilités

La théorie de possibilités a été définie par Zadeh et développée par Dubois et Prade à partir de la théorie des sous-ensembles flous (Dubois *et al.,* 1985). Les informations collectées pour le calcul des indicateurs sont entachées d'imperfections. Ces informations sont issues de plusieurs bases de données ou elles sont estimées par des experts. Or, il s'agit de trouver les valeurs plausibles des variables provenant de plusieurs bases de données de nature hétérogènes. La théorie de possibilités fournit des outils de représentation des informations sous forme de possibilités. Une *distribution de possibilités* $\pi(x)$ est une fonction d'un espace X dans l'intervalle $[0, 1]$. Cette distribution peut s'interpréter comme un ensemble d'intervalles emboîtés. Ces possibilités peuvent être combinées conjonctivement ou disjonctivement. Dans cet article, on utilise les fusions possibilistes au sein de la notion de sous-ensembles maximaux cohérents.

3. Indicateur : risque phytosanitaires pour les eaux souterraines (Reso)

Dans cette section, nous présentons l'objectif, les variables et le mode de calcul de Reso. Un indicateur est une note qui correspond à une valeur statistique, une mesure ou une série d'éléments quantitatifs ou qualitatifs. L'indicateur Reso estime le risque de l'impact des pratiques phytosanitaires vis-à-vis des eaux souterraines. Reso se calcule à partir de trois types de variables (voir tableau 1) pour chaque application de matières actives. Cet indicateur est un outil d'aide à la décision qui permet à l'agriculteur de limiter l'impact de ses pratiques phytosanitaires sur la qualité des eaux souterraines. Les techniciens agricoles ou les agriculteurs eux-mêmes peuvent l'utiliser pour établir un diagnostic a posteriori ou pour évaluer a priori de nouvelles pratiques.

Types	Variables	Description	Origine
Matière active	GUS	Indice d'ubiquité de Gustafson	Fonction(DT50,Koc)
	DT50	Temps de demi-vie	Bases de données
	Koc	Le coefficient de partage organique-eau	Bases de données
	DJA	La dose journalière admissible	Bases de données
Milieu	Potentiel de lessivage	Quantité lessivée estimée	Expert
Pratiques	Position d'application	Taux d'interception par la surface	Mesure

Tableau 1. *Variables de Reso*

3.1. *Mode de calcul de Reso*

la méthode de calcul de Reso s'appuie sur un système de règles floues (Van der Werf *et al.,* 1998). Les étapes de calcul de Reso sont :

– Classer les variables GUS, DJA, position et potentiel de lessivage en deux classes favorable et défavorable. Des limites pour toutes les variables sont choisies (Bockstaller *et al.,* 2003) (ex :la limite favorable de GUS est 1.8 et celle défavorable est 2.8).

– Construire la base de règles ($2^4 = 16$) dont les conclusions nommées scores sont issues d'un consensus entre experts. Comme par exemple :si toutes les variables sont favorables alors le score est 10.

– Calculer les fonctions d'appartenance aux classes. Des fonctions sinusoïdales sont associées aux variables de Reso. Elles sont construites à partir des variations de la variable. Comme par exemple : la fonction favorable de GUS s'écrit sous la forme : $0.5 + 0.5\cos(\pi(\text{GUS} - 1.8))$. Pour GUS = 2.5, le degré d'appartenance favorable est $f = 0.16$ et celui défavorable est $d = 1 - f = 0.84$.

– Calculer la valeur de vérité ω_i de chaque règle = minimum des fonctions d'appartenance des variables de chaque règle. Comme par exemple pour la matière active chloridazone, si toutes les variables sont favorables, ω_1 = $\min(f_1, f_2, f_3, f_4) =$

$\min(0.16, 0.9, 0, 0.65) = 0$ avec $f_k, k = 1, \ldots, 4$ sont les degrés d'appartenance des variables aux classes favorables.

– Calcul de risque Reso : la valeur de Reso vaut la moyenne des scores (s_i) pondérées par leur valeur de vérité (ω_i) des règles (Van der Werf *et al.,* 1998).

$$Reso = \frac{\sum_{i=1}^{16} s_i \times \omega_i}{\sum_{i=1}^{16} \omega_i} \qquad [1]$$

L'interprétation du résultat de Reso se fait de la manière suivante : le risque est considéré maximal pour l'environnement lorsque Reso = 0 et minimal pour Reso = 10. Si Reso > 7, les pratiques phytosaniataires sont considérées durables, leur impact sur la qualité des eaux souterraines comme faible. Dans ce cas on ne conseillera pas de changement de pratiques. Le résultat 7 est donc considéré comme le seuil d'acceptabilité pour l'environnement.

4. Exposition de problème et présentation de la méthode

Dans cet article, nous considérons n sources, fournissant chacune une information pour évaluer la variable mal connue x. Nous utilisons la méthode de sous-ensembles maximaux cohérents pour résumer toutes ces informations. Nous travaillons avec le dernier résultat un intervalle de variation de Reso.

4.1. *Sous-ensembles maximaux cohérents sur l'ensemble des intervalles*

L'utilisation d'un seul mode de fusion n'est pas toujours souhaitable surtout quand on connaît rien sur la fiabilité de sources. La méthode de sous-ensembles maximaux cohérents permet d'avoir un maximum d'informations sur l'ensemble d'informations fournies en faisant au mieux avec la fiabilité des sources (Rescher *et al.,* 1970). Cette méthode consiste à fusionner par un opérateur conjonctif des sous-ensembles de sources cohérentes entre elles, et d'utiliser un opérateur disjonctif sur ces derniers sous-ensembles. Dans ce but, un algorithme a été développé par Dubois et al. (Dubois *et al.,* 2000) pour chercher les sous-ensembles maximaux cohérents. Soient $I_i = [a_i, b_i], (i = 1, \ldots, n)$ des intervalles de $\mathbb{R}$. La méthode revient à prendre l'intersection maximale cohérente au sein de K_j sous-ensembles ($K_j \subseteq \{1, \ldots, n\}, j \leq n$) t.q. $\bigcap_{i \in K_j} I_i \neq \emptyset$ pour ensuite faire l'union sur les résultats.

Algorithme 1 : Sous-ensembles maximaux cohérents sur des intervalles

Input : n intervalles
Output : Liste de m sous-ensembles maximaux cohérents K_j
Liste=$\emptyset$;$j = 1$;$K = \emptyset$;
Mettre en ordre croissant $\{a_i, i = 1, \ldots, n\} \cup \{b_i, i = 1, \ldots, n\}$;
Les renommer $\{c_i, i = 1, \ldots, 2n\}$
avec $type(i) = a$ si $c_i = a_k$ et $type(i) = b$ si $c_i = b_k$;
for $i = 1$ to $2n - 1$ **do**
  if $type(i) = a$
    Ajouter source k à K t.q. $c_i = a_k$;
    if $type(i + 1) = b$
      Ajouter K à la liste ($K_j = K$) ;
      $j = j + 1$;
  else
    Enlever la source k de K t.q. $c_i = b_k$;

L'algorithme se base sur l'ordre des points extrêmes des intervalles en une séquence $(c_i)_i = 1, \ldots, 2n$. L'intersection maximale cohérente d'intervalles est atteint seulement qu'un c_i de type b (i.e. la borne supérieure d'un intervalle) est suivi par un c_{i+1} de type a (i.e. la borne inférieure).
L'exemple suivant montre comment on applique la méthode des sous-ensembles maximaux dans le cadre de fusion d'informations issues de plusieurs sources. Considérons la variable DT50 fournie par trois bases de données : ARS, Agritox et RIVM. Ces bases fournissent ces informations sous forme d'un intervalle de variation. L'information est résumée dans le tableau 2.

Sources	ARS	Agritox	RIVM
Intervalle	[10,42]	[46,53]	[21,56]

Tableau 2. *Exemple : informations des sources*

L'algorithme trouve deux sous-ensembles maximaux cohérents : K_1 = {ARS,RIVM} et K_2 = {RIVM,Agritox}. Après application de la méthode des sous-ensembles maximaux cohérents, le résultat est (ARS $\cap$ RIVM) $\cup$ (RIVM $\cap$ Agritox) = [21,42] $\cup$ [46,53]

4.2. *Calcul d'intervalles de Reso*

Reso est une moyenne des scores pondérée par les valeurs de vérité (voir 3.1). Reso s'écrit sous la forme Reso$(\omega_i) = \frac{\sum_i \omega_i \times s_i}{\sum_i \omega_i}$ avec ω_i et s_i sont respectivement la valeur de vérité et le score de la règle i, pour $i = 1, \ldots, 16$. Il s'agit de trouver les bornes de Reso si $\omega_i \in [\omega_i^-, \omega_i^+]$. Les scores sont des réels entre $[0, 10]$. Partant de l'hypothèse qu'on peut ordonner les scores (i.e. $s_i \leq s_j$ pour tout $i < j$), Fortin et al. (Fortin *et al.,* 2008) ont demontré qu'il existe un rang $k \in [1, 15]$, tels que pour tout

$i \leq k$, Reso(.) est décroissante suivant w_i, et pour tout $i > k$, Reso(.) est croissant respectant w_i. k correspond au rang à partir duquel la fonction Reso(ω_i) change de signe suivant ω_i. Or, on peut calculer les bornes inférieures et supérieures de Reso, pour tout $k = 1, \ldots, 15$, telles que :

$$Reso_k^{inf} = \frac{\sum_{j=1}^{k}(w_j)^+.s_j + \sum_{j=k+1}^{n}(w_j)^-.s_j}{\sum_{j=1}^{k}(w_j)^+ + \sum_{j=k+1}^{n}(w_j)^-} \quad [2]$$

et

$$Reso_k^{sup} = \frac{\sum_{j=1}^{k}(w_j)^-.s_j + \sum_{j=k+1}^{n}(w_j)^+.s_j}{\sum_{j=1}^{k}(w_j)^- + \sum_{j=i+1}^{n}(w_k)^+} \quad [3]$$

Ce qui nous permet de déterminer un domaine de variation de Reso, défini par : [min(Reso$_k^{inf}$),max(Reso$_k^{sup}$)] pour $k = 1, \ldots, 15$.

5. Applications numériques

Cette section présente un calcul de Reso pour deux matières actives appliquées aux cultures de betterave et de maïs en utilisant les méthodes mentionnées déjà citées. Pour les caractéristiques de la matière active (Koc, DJA et DT50), chaque variable est donnée par différentes bases de données comme par exemple ARS (américaine), RIVM (néerlandaise), le comité de liaison ou Agritox (France). Les sources d'informations fournissent leur point de vue sous forme des intervalles, soit des valeurs. Dans certains cas les bases ne donnent pas d'avis sur la variable. On applique la méthode des sous-ensembles maximaux cohérents afin d'obtenir les distributions finales ces variables. La variable GUS est une fonction de DT50 et Koc. Elle est définie suivant l'équation GUS = $\log$(DT50)$\times(4 - \log$(Koc)). Les variables Koc et DT50 varient respectivement entre [Koc$^-$,Koc$^+$] et [DT50$^-$,DT50$^+$]. En utilisant la notion d'arithmétique sur les intervalles (Moore, 1979), on calcule les bornes de la variable GUS telles que : GUS$^-$ = $\log$(DT50$^-$)$\times(4 - \log$(Koc)$^+$) et GUS$^+$ = $\log$(DT50$^+$)$\times(4 - \log$(Koc)$^-$). Le potentiel de lessivage est une information pour laquelle on n'a pas beaucoup d'informations. Ce potentiel est donné par un expert sous forme d'un indice entre 0 et 1 et estimant la quantité lessivée de pesticide en fonction de caractéristiques du sol. Pour ce type d'informations, on propose un découpage en α-coupes (pour cet exemple, $\alpha = 0.9$). Notons qu'on pourrait prendre d'autres valeurs de α pour le calcul de Reso et deux valeurs proches n'affectent pas la décision de l'agriculteur. La position d'application dépend de la date d'application (15 juillet) et de la couverture de sol. Pour cet exemple, nous calculons Reso pour une parcelle de betterave (chloridazone) et de maïs (glyphosate). Le tableau 3 est un récapitulatif de résultats obtenus. Dans certains cas, on a convexifié la distribution pour ne pas perdre d'informations après la fusion.

Les valeurs de vérité, $\omega_1, \ldots, \omega_{16}$, sont calculées pour toutes les règles. Ces valeurs sont calculés à l'aide d'une extension du minimum sur les intervalles (Dubois *et al.,* 1985). ce qui veut dire $\min([a, b], [c, d]) = [\min(a, c), \min(b, d)]$ avec $[a, b]$

et $[c, d]$ deux intervalles de $\mathbb{R}$. Par exemple, pour la matière active chloridazone : si toutes les variables de Reso sont favorables, les intervalles des degrés d'appartenance de GUS, DJA, position et potentiel de lessivage sont respectivement : [0,0.15], [0.65,0.65], [0.97,0.97] et [0.02,0.02]. D'où ω_1 varie dans l'intervalle [0,0.97].

Variables / Matière active	DT50	KOC	DJA	GUS	Lessivage	Position
Chloridazone	[21,46]	120	0.025	[2.53,3.19]	[0.81,0.91]	0.9
Glyphosate	[18,47]	[167,2640]	0.3	[0.73,2.97]	[0.81,0.91]	0.75

Tableau 3. *Résultats des variables sans défuzzyfication*

Nous calculons les bornes inférieures et supérieures de Reso, pour $k = 1, \ldots, 15$, en appliquant les équations (2) et (3). Les résultats de Reso sont [4.3,9.82] pour le chloridazone et [2,10] pour le glyphosate. Bien que cet intervalle est limité, les utilisateurs potentiels des indicateurs préfèrent une valeur unique comme résultat de l'indicateur pour faciliter leur décision. Dans ce cas, on proposera une défuzzyfication des distributions résultantes de la méthode de sous-ensemble maximaux cohérents. La défuzzyfication est la conversion de la distribution en une valeur. Dans notre exemple, nous utilisons la méthode de centre de gravité. Le tableau 4 montre les résultats des variables de Reso après défuzzyfication. Ce qui donne comme résultat de Reso 8.59 pour le chloridazone et 10 pour le glyphosate.

Variables / Matière active	DT50	KOC	DJA	GUS	lessivage	Position
Chloridazone	36.3	120	0.025	2.99	0.88	0.9
Glyphosate	31.6	928	0.3	1.55	0.88	0.75

Tableau 4. *Résultats des variables après défuzzyfication*

6. Discussion et conclusion

Dans ce travail, les conclusions des règles de décision sont fixées. La méthode utilisée pour fusionner les données provenant des sources hétérogènes nous a permis de prendre au mieux toute l'information fournie par toutes les bases de données. Comme on a vu plus haut, la méthode fondée sur la notion de sous-ensembles maximaux cohérents ne fait aucune supposition concernant les sources d'informations. Cette méthode est simple, elle ne requiert que l'information fournie par les sources et intuitivement séduisante. La méthode s'intéresse plus à la redondance des différentes sources en accordant plus de poids aux informations qui se répètent.
Quant à la méthode de calcul d'intervalles, le principal avantage de la méthode est sa grande simplicité. Les opérations arithmétiques utilisées dans le modèle de calcul de l'indicateur Reso sont faciles à réaliser. De côté défauts, cette méthode ne fait pas de

différence entre les intervalles calculés, ils sont tous considérés comme équivalents.
Il est nécessaire de comparer les résultats obtenus par la fusion utilisant les sous-ensembles maximaux cohérents avec d'autres modes de fusion (voir par exemple (Oussalah *et al.,* 2003)). Ainsi que pour le résultat de l'indicateur, il serait intéressant de voir d'autres algorithmes de calcul de la moyenne de pondérations floues comme ceux développés dans (Lee *et al.,* 1997).
Dans cet article, nous avons calculé un indicateur agri-environnemental évaluant le risque lié aux pesticides vis-à-vis des eaux souterraines en tenant compte de l'imperfection due à l'expertise lors de choix de données. La notion de sous-ensembles maximaux cohérents est utilisée au sein des fusions de la théorie des possibilités pour calculer une distribution finale de la quantité mal connue. Nous avons calculé un intervalle de variation de l'indicateur. Une note de l'indicateur est calculée avec la défuzzyfication des distributions par la méthode de centre de gravité.

7. Bibliographie

Bockstaller C., Girardin P., *Mode de calcul des indicateurs d'INDIGO*, INRA, 2003.

Dubois D., Hajek P., Prade H., « Knowledge-Driven versus Data-Driven Logics », *J. of Logic, Language and Information*, vol. 9, n° 1, p. 65-89, 2000.

Dubois D., Prade H., *Théorie des possibilités*, Masson, Paris, 1985.

Fortin J., Dubois D., Fargier H., « Gradual numbers and their application to fuzzy interval analysis », *IEEE TRANS. ON FUZZY SYSTEMS*, vol. 16, n° 2, p. 388-402, 2008.

Girardin P., Bockstaller C., Van Der Werf H., « Assessment of potential impact of agricultural practices on the environment the AGRO*ECO method », *Environmental Impact Assessment*, vol. 20, p. 227-239, 2000.

Lee D., Park D., « An efficient algorithm for fuzzy weighted average », *Fuzzy sets and systems*, vol. 87, n° 1, p. 39 - 45, 1997.

Moore R. E., *Methods and Applications of Interval Analysis*, Siam, Philadelphia, 1979.

Oussalah M., Maaref H., Barret C., Prade H., « From adaptive to progressive combination of possibility distributions », *Fuzzy Sets and Systems*, vol. 139, n° 3, p. 559-582, 2003.

Rescher N., Manor R., « On inference from inconsistent premises », *Theory and Decision*, vol. 1, p. 179-217, 1970.

Van der Werf H., Zimmer C., « An indicator of pesticide environmental impact based on a fuzzy expert system », *Chemosphere*, vol. 36, n° 10, p. 2225-2249, 1998.