Fusion de données imparfaites dans un système d'information agro-environnemental

Une approche basée croyance

Karima Zayrit

CReSTIC (URCA)
IUT de Reims Châlons Charleville
Rue des Crayères
51687 Reims cedex 2
karima.zayrit@univ-reims.fr

RÉSUMÉ. *La mise en place d'OBSERVOX, un observatoire des pratiques agricoles et de leur pression sur la ressource en eau dans le bassin versant de la Vesle, requière l'exploitation d'informations incertaines et imprécises dans un environnement multi-sources. Dans ce travail nous nous intéressons à la gestion des imperfections de données spatiales agro-environnementales, et à la fusion de résultats de classifieurs crédibilistes dans le cadre de la théorie des fonctions de croyance afin de fournir aux utilisateurs du système d'information un résultat qualifié par un degré de confiance.*

ABSTRACT. *OBSERVOX, a community environmental information system for the monitoring of agricultural practices and their pressure on water resources in the Vesle basin, requires the use of information which is imprecise and uncertain in a multi-sources context. In this work, we deal with the management of agronomical data imperfection and with the use of the belief function theory in uncertain classifier decision fusion. Our aim is to always provide information qualified by a belief degree.*

MOTS-CLÉS : *Fusion de données, système d'information environnemental, agronomie, fonctions de croyance, données imparfaites.*

KEYWORDS: *Data merging, environmental information system, agronomy, belief functions, imperfect data.*

1. Introduction

Ce travail s'inscrit dans le cadre d'Observox : un observatoire des pratiques agricoles et de leur pression sur la ressource en eau au niveau du bassin versant de la Vesle. La mise en place de ce dispositif est parti du constat que la ressource en eau sur le territoire étudié était contaminée par des pesticides d'origine agri-viticole, d'où le besoin de mieux comprendre les pratiques qui peuvent en être responsable.

Observox a pour objectif d'informer les différents acteurs du territoire concerné sur l'évolution des pratiques agricoles en proposant une historisation (pérenne) de l'information et de les aider dans leurs efforts pour améliorer leurs pratiques (choix des molécules, enherbement, etc.). L'observatoire se charge aussi d'assurer la fluidité de la circulation des informations entre acteurs afin d'aider à la mise en place de solutions permettant de diminuer les risques encourus par la ressource en eau ; il vise également à proposer une reconstruction spatialisée et qualifiée des informations relatives à l'utilisation des pesticides.

Pour sa construction, Observox s'appuie sur une démarche méthodologique, s'inspirant de (Passouant *et al.*2007), visant à faire dialoguer, participer et mobiliser les différents acteurs du territoire (agriculteurs, viticulteurs, techniciens, gestionnaires de l'eau, citoyens, responsables politiques) autour de la protection et de la qualité de la ressource en eau, en fonction des attentes et des besoins de chacun.

Il vise donc à l'organisation et à la diffusion de l'information sur les pratiques phytosanitaires, la nature des produits, et leurs liens avec l'évolution des systèmes de culture. Un tel système nécessite de traiter de la qualité de l'information et notamment de porter attention à la prise en compte de l'imperfection des connaissances dès leur acquisition, puis lors de leur utilisation afin de construire des décisions plus justes en sachant bien de quoi il s'agit. Parmi les travaux qui ont traité de la question, on peut citer (Mardkheh *et al.*2012) qui dans le cadre de développement d'un outil géo-décisionnel pour améliorer l'évaluation du risque d'érosion propose une représentation floue des zones côtières à risques.

Dans ce but, les théories mathématiques de l'incertain offrent un cadre théorique permettant de traiter et de raisonner sur des informations imparfaites. Parmi les principales, nous pouvons citer la théorie des probabilités, la théorie des sous-ensembles flous, et la théorie des fonctions de croyance (Dempster1968, Shafer1976).

Afin d'atteindre ces objectifs, nous avions effectué un premier travail en nous appuyant sur la logique floue, en particulier dans le cadre de la propagation de l'imprécision quantitative et spatiale (Zayrit *et al.*2011). Parallèlement à cette démarche nous nous intéressons à la représentation de l'imperfection par la théorie des fonctions de croyance qui permet l'exploitation d'informations multi-sources, imprécises et incertaines et qui a été exploitée avec succès dans diverses applications notamment dans un contexte environnemental (Corgne *et al.*2003). Dans ce travail nous exploitons la théorie des fonctions de croyance, afin que le système d'information(SI) fournisse pour chaque endroit du territoire étudié une information sur le type de culture présent,

qualifiée avec un indice de confiance. Un des enjeux majeurs dans le cadre de l'évolution des systèmes d'information est d'intégrer les différents outils de représentation et d'inférence sur les données qualifiées.

Dans cet article, nous nous intéresserons dans la section 2 à la mise en place de l'observatoire et au besoin de gestion de l'imperfection. La section 3 abordera les principales notions et principes de la théorie des fonctions de croyances. La section 4 illustrera ceux-ci dans le cadre de données agronomiques simulées. La conclusion et les perspectives de ce travail seront présentées dans la section 5.

2. Mise en œuvre du système d'information de l'observatoire

Observox peut être vu comme un système d'information pérenne, partagé et co-construit par l'ensemble des acteurs concernés par l'enjeu (Passouant *et al.*2007). Dans notre contexte, il permet de capitaliser, gérer, traiter et restituer une information pertinente à l'échelle du territoire à l'ensemble des acteurs concernés par l'enjeu global « Pratiques agri-viticoles et qualité de l'eau vis-à-vis des produits phytosanitaires ».

Néanmoins en l'abscence d'une connaissance exhaustive des pratiques agri-viticoles réalisées, la mise en place d'un tel système nécessite de lever plusieurs verrous. Les besoins liés à la construction d'Observox peuvent être exprimés à travers les points suivants :

– Comment affecter à un élément de l'espace un type de culture à partir du résultat de différentes classifications ?

– Comment intégrer la fiabilité des sources pour la prise de décision finale ?

Ainsi, la démarche choisie implique que le système puisse gérer des informations :

– Multi-sources : les données sont issues de sources d'information variées tant pour les pratiques phytosanitaires que pour la localisation des surfaces cultivées ; elles peuvent provenir d'organismes institutionnels (par ex. : registre parcellaire graphique, Corine Land Cover) mais aussi d'enquêtes terrain auprès des agriculteurs du territoire étudié ou de dires d'experts et peuvent être complétées à l'aide de photo-interprétation d'images satellitaires ou aériennes.

– Multi-variées et multi-composantes : elles peuvent être de natures différentes (quantitatives, qualitatives, spatiales et temporelles) et mettent en action de nombreuses variables.

– Multi-échelles : elles sont collectées à différents niveaux d'échelles tant spatialement (échelle nationale, échelle du bassin, échelle de la parcelle) que temporellement (ex : rotation de culture par année).

– Partielles et évolutives : des données peuvent être manquantes en fonction des échelles, et les indices observés peuvent varier au cours de l'exploitation d'Observox.

Afin de répondre aux objectifs à l'aide d'indices les plus fiables possibles, il est nécessaire d'intégrer dans le système d'information des outils gérant la qualité de l'information (Devillers et Jeansoulin2005). En effet, dans notre cadre, la formalisation des imperfections dans les connaissances sur les pratiques agricoles dans le temps et l'espace doit être accessible aux différents acteurs et fournir une visualisation enrichie dans un système d'information géographique. Ainsi, lors de la construction de l'observatoire, une première étape a été de renseigner l'ensemble des métadonnées sur la qualité de la donnée pour l'ensemble des sources d'information à notre disposition via l'outil MDweb à l'instar de la démarche proposée dans (Desconnets *et al.*2007).

Dans cet article, nous nous intéressons à l'exploitation d'informations incertaines et imprécises dans un environnement multi-sources. Dans ce but, nous avons choisi de travailler dans le cadre de la théorie des fonctions de croyance. En termes de fusion d'information, ce formalisme possède des outils qui nous permettent de combiner des informations provenant de sources différentes, de gérer le conflit entre les différentes connaissances, de pondérer les sources d'information en fonction de leur fiabilité, et ainsi d'aider à la prise de décision. Notre objectif est d'exploiter ces différents outils pour la gestion des données imparfaites et l'établissement de la fiabilité du système.

3. Théorie de l'évidence et modèle des croyances transférables

Connue sous le nom de « théorie de Dempster-Shafer » ou « théorie de l'évidence », la théorie des fonctions de croyance a été introduite dans (Dempster1968, Shafer1976). Elle permet de modéliser et de gérer l'information imprécise et l'incertaine dans un même formalisme, via les fonctions de masse m, de plausibilité Pl et de croyance Bel. Le modèle de croyance transférable (MCT), ou *transferable belief model* (TBM), proposé par Smets (Smets1994) est un cadre formel basé sur la définition des fonctions de croyance pour la représentation de la connaissance et la prise de décision. Le MCT s'appuie sur deux niveaux de perception de l'information, le niveau crédal qui concerne l'étape de la modélisation et de l'agrégation des données, et le niveau pignistique qui est dédié à la prise de décision en transformant les fonctions de masse en des fonctions de probabilités pignistiques.

3.1. *Niveau crédal*

3.1.1. *Représentation des connaissances*

Afin de modéliser la connaissance incertaine, on commence par identifier l'ensemble fini des n hypothèses ou solutions à un problème donné, appelé cadre de discernement et noté : $\Omega = \{h_1, h_2, .., h_n\}$; les fonctions de masse sont définies sur l'ensemble 2^Ω des parties de Ω, et pas uniquement sur les singletons comme c'est le cas pour les probabilités (un exemple d'application fourni dans la partie 4).

Une fonction de masse de croyance est définie mathématiquement par une fonction m sur 2^Ω à valeur dans $[0, 1]$ qui vérifie la condition suivante :

$$\sum_{A \in 2^\Omega} m(A) = 1. \quad [1]$$

La masse $m(A)$ représente le degré de confiance affecté à la proposition A. Chaque sous ensemble $A \subseteq \Omega$ tel que $m(A) > 0$ constitue un élément focal de m. $m(\Omega)$ représente l'ignorance.

A l'aide de ces masses, on peut déterminer la fonction de croyance Bel qui associe à chaque élément de 2^Ω la confiance minimale que l'on peut avoir dans celui-ci. Etant donné une fonction de masse m, la fonction Bel est définie pour tout $A \in 2^\Omega$ comme la somme des masses des hypothèses incluses dans A comme suit :

$$Bel(A) = \sum_{B \subseteq A, B \neq \emptyset} m(B). \quad [2]$$

La fonction de plausibilité Pl mesure le degré de croyance maximum que l'on a sur un événement. Pl est définie pour tout $A \in 2^\Omega$ comme la somme des masses des hypothèses dont l'intersection avec A n'est pas nulle :

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B). \quad [3]$$

L'intervalle $[Bel(A), Pl(A)]$, appelé intervalle de confiance ou de croyance, représente l'incertitude que l'on a sur un événement ; sa largeur est représentative de l'imprécision que l'on a sur cet événement.

3.1.2. *Fusion*

Afin de fusionner des informations incertaines venant de plusieurs sources, plusieurs règles ont été proposées dans la litterature. La plus généralement utilisée est celle de Dempster qui permet de combiner des fonctions de masse à partir de sources d'information distinctes. Cette règle est associative et commutative.

Soient deux sources S_1, S_2 fournissant respectivement les fonctions de masse m_1 et m_2, leur combinaison selon la règle de Dempster est définie sous sa forme normalisée par :

$$m_{1 \oplus 2}(A) = m_1 \oplus m_2(A) = \frac{\sum_{A_1 \cap A_2 = A} m_1(A_1) * m_2(A_2)}{1 - K} \quad [4]$$

$$K = \sum_{A_1 \cap A_2 = \emptyset} m_1(A_1) * m_2(A_2). \quad [5]$$

K représente le degré de conflit entre les sources m_1 et m_2. Le conflit peut se produire si les sources ne sont pas fiables ou si elles donnent des informations sur des phénomènes différents. Plus K est grand, plus le conflit entre les sources est important, et moins les résultats de la combinaison sont sûrs.

3.2. *Niveau pignistique : prise de décision*

La dernière étape dans le processus de la fusion par le formalisme des fonctions de croyance est la prise de décision. Cette étape a pour objectif la prise de décision (sélectionner un élément de 2^{Ω}) à partir des masses de croyance définies au niveau crédal. Plusieurs règles de décision sont envisageables telles que :

Le maximum de plausibilité consiste à prendre, parmi les singletons, celui ayant le maximum de plausibilité.

$$Dec(\Omega) = argmax_{A \in 2^{\Omega}, |A|=1} Pl(A). \quad [6]$$

Le maximum de croyance est une règle plutôt pessimiste, car au lieu des plausibilités, les croyances Bel, plus strictes, sont utilisées.

$$Dec(\Omega) = argmax_{A \in 2^{\Omega}, |A|=1} Bel(A). \quad [7]$$

Le maximum des probabilités pignistiques, proposé par Smets (Smets1994), est un compromis entre le maximum de plausibilité et le maximum de croyance. Considérée comme une règle prudente, elle se base sur une répartition équitable des masses placées sur les éléments non singleton vers les singletons qui les composent. La probabilité pignistique est définie pour un singleton de 2^{Ω} par :

$$BetP(A) = \sum_{B \in 2^{\Omega}, A \in B} \frac{m(B)}{|B| * (1 - m(\emptyset))}. \quad [8]$$

Le singleton sélectionné à l'issue de la décision est donc :

$$Dec(\Omega) = argmax_{A \in 2^{\Omega}, |A|=1} BetP(A). \quad [9]$$

Nous présentons dans la section suivante une application dans le contexte d'Observox.

4. Application à la fusion de données agronomiques

L'objectif de l'exploitation de la théorie des fonctions de croyance dans la structure de notre système d'information est la gestion de données multi-sources en prenant en considération leur qualité. En effet, nous souhaitons que le système fournisse pour chaque endroit du territoire l'information sur le type de culture pratiquée et la confiance associée, ainsi que la fiabilité du système en cet endroit. La masse de croyance associée au type de culture renvoyé (à l'aide de la probabilité pignistique par exemple) nous fournit une information sur la qualité de la réponse (décision) du système.

Nous nous plaçons dans le cadre d'un exemple simulé mais réaliste où nous disposons de résultats de classifications crédibilistes à l'echelle du pixel, issus d'un traitement d'images satellitaires S_1 et d'une classification fournie par un expert S_2 en une

zone du territoire étudié selon la liste de culture $\Omega = \{$Blé, Orge, Maïs$\}$. Les informations sur les types de culture sont issues du Registre Parcellaire Graphique. Nous cherchons à identifier le type de culture à cet endroit donné. Les masses de croyances (resp. m_1 et m_2) fournies par ces classifieurs sont présentées dans le tableau 1.

Hypothèse	m_1	m_2	Hypothèse	m_1	m_2
$\emptyset$	0	0	{Blé,Orge}	0.4	0.3
{Blé}	0.1	0	{Blé,Maïs}	0	0.6
{Orge}	0	0	{Orge,Maïs}	0	0
{Maïs}	0.5	0	Ω	0	0.1

Tableau 1. *Résultats des classifieurs S_1 et S_2 sur $\Omega = \{$Blé, Orge, Maïs$\}$*

Ainsi, nous obtenons le tableau 2 de combinaison pour lesquels au moins une des masses issues des classifieurs est non nulle.

	$m_2(\{$Blé,Orge$\})$=.3	$m_2(\{$Blé,Maïs$\})$=.6	$m_2(\Omega)$=.1
$m_1(\{$Blé$\})$=.1	{Blé}=.03	{Blé}=.06	{Blé}=.01
$m_1(\{$Maïs$\})$=.5	K=.15	{Maïs}=.30	{Maïs}=.05
$m_1(\{$Blé,Orge$\})$=.4	{Blé,Orge}=.12	{Blé}=.24	{Blé,Orge}=.04

Tableau 2. *Tableau de combinaison des masses m_1 et m_2 sur $\Omega = \{$Blé, Orge, Maïs$\}$*

Nous pouvons observer que le conflit (K) est de 0.15, ce qui n'est pas signicatif. En appliquant la combinaison de Dempster, nous obtenons les masses présentées dans le tableau 3.

Hypothèse	$m_{1\oplus 2}$	Pl	Bel	$BetP$
{Blé}	0.4	0.59	0.4	0.49
{Maïs}	0.41	0.41	0.41	0.41
{Blé,Orge}	0.19	-	-	-

Tableau 3. *Résultats de la fusion des classifieurs S_1 et S_2 sur $\Omega = \{$Blé, Orge, Maïs$\}$*

L'étude des résultats donne : une lecture très pessimiste voudrait que l'on choisisse le maïs (maximum de croyance) tandis qu'une lecture moins pessimiste (maximum de probabilité pignistique) ou optimiste (maximum de plausibilité) nous donnerait le blé. Il faut cependant nuancer les choix possibles en essayant de réduire le conflit en pondérant les sources d'information et en privilégiant une approche négociée avec les dires d'experts. Cette étape de fusion de connaissance permet ensuite de fournir aux utilisateurs un résultat qualifié par un degré de confiance qui doit être stocké dans le système d'information avec les données fusionnées et leur lignage.

5. Conclusion et perspectives

Les premiers travaux entrepris dans le cadre de l'Observatoire des pratiques phytosanitaires sur le bassin de la Vesle (OBSERVOX) ont permis de valider la pertinence d'une approche permettant d'appréhender l'imperfection de l'information de son acquisition à son utilisation. Le cadre théorique des fonctions de croyance permet d'enrichir d'un degré de confiance la connaissance que l'on intégre dans le système d'information environnemental. Afin de valider l'approche, nous sommes en train de mener une étude mélant analyse terrain et fusion crédibiliste sur une zone restreinte. Une future étape visera à pouvoir formaliser dans un outil de catalogage intégré au SI la diversité des modèles mathématiques de représentation de l'imperfection dans les données. La construction de l'observatoire s'accompagnera de la construction d'un modèle conceptuel proche de ceux proposés dans (Pinet2012) augmenté de la gestion de l'imperfection.

6. Bibliographie

Corgne S., Hubert-Moy L., Dezert J., Mercier G., « Land cover change prediction with a new theory of plausible and paradoxical reasoning », *Proc. of Fusion*, 2003, p. 8-11.

Dempster A. P., « A generalization of Bayesian inference », *Journal of the Royal Statistical Society*, vol. 30, n° 2, 1968, p. 205–247.

Desconnets J.-C., Libourel T., Clerc S., « Cataloguer pour diffuser les ressources environnementales », *INFORSID*, 2007, p. 344-361.

Devillers R., Jeansoulin R., Eds., *Qualité de l'information géographique*, Traité IGAT, Hermes, 2005.

Mardkheh A., Mostafavi M., Bédard Y., « Dealing with Uncertainty in Coastal Risk Assessment : Fuzzy Representation of Coastal Risk Zones », 2012.

Passouant M., Caron P., Loyat J., Tonneau J.-P., Barzman M., « Observatoire des agricultures et des territoires : mise à l'épreuve d'une méthode de conception », , 2007, page online.

Pinet F., « Entity-relationship and object-oriented formalisms for modeling spatial environmental data », *Environmental Modelling and Software*, vol. 33, n° 0, 2012, p. 80-91.

Shafer G., *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, N.J., USA, 1976.

Smets P., « The Transferable Belief Model », *Artificial Intellignence*, vol. 66, 1994, p. 191–234.

Zayrit K., Desjardin É., de Runz C., Akdag H., « Propagation of spatial imprecision in imprecise quantitative data in agronomy », *International Symposium on Spatial Data Quality*, , 2011, p. 145–150, INESC Coimbra.