Vers une méthode de caractérisation et de quantification des incertitudes dans le cadre d'une fusion de données hétérogènes multicapteurs dans le domaine de la pollution atmosphérique

Aymeric Ambert*, Mickael Germain*, Yacine Bouroubi*

* Département de géomatique appliquée, Faculté des lettres et sciences humaines
2500, boulevard de l'Université, Sherbrooke (Québec) J1K 2R1
aymeric.ambert@usherbrooke.ca

Résumé. La lutte contre la pollution atmosphérique est un enjeu majeur du 21e siècle. La gestion des données massives liées à la diversification des supports de mesure est un défi et engendre des problématiques inédites en termes de volume de qualification et de traitement de l'information. Les méthodes de fusion de données sont autant de solutions au problème posé par l'utilisation de données massives. La prise en compte de chaque capteur en tant qu'élément d'information dans le cadre de la fusion entraîne néanmoins un risque quant à l'incertitude globale des données à prendre en considération. Le présent article vise à établir une approche pour réduire l'incertitude relative pour chaque source de données en utilisant la fusion évidentielle. Se basant sur un modèle attributaire de données existantes, l'article propose de définir des indicateurs de performances permettant de valider ou non un tel modèle.

1 Introduction et état de l'art

Le 22 septembre 2021, l'OMS (Organisation Mondiale de la Santé) a rendu public son nouveau guide concernant la pollution atmosphérique considérant en préambule que "la pollution atmosphérique est l'une des principales menaces environnementales pour la santé" (OMS, 2021). Lutter contre la pollution atmosphérique est donc devenu un enjeu majeur de santé publique. L'utilisation d'un capteur de pollution correspond à un besoin spécifique et doit prendre en compte la précision, l'incertitude et la capacité d'adaptation à de potentiels nouveaux besoins de celui-ci (ex : Piedrahita et al. (2014)). Pour y parer, la recherche et l'industrie, via l'essor de nouvelles technologies comme l'intelligence artificielle, s'oriente de plus en plus vers l'utilisation big data, c'est-à-dire la massification des données, notamment pour pallier l'incomplétude et l'incertitude de données mesurées (Hariri et al., 2019). Dans le contexte de la pollution atmosphérique et du présent article, il s'agit justement de fusionner des données issues de capteurs de différents types. Il existe déjà de nombreux travaux de recherche liés à la fusion de données (Khaleghi et al., 2013). Néanmoins la notion d'incertitude revêt un caractère tout à fait singulier, car sa prise en compte pour chaque capteur impacte l'information globale

délivrée par la fusion elle-même et notamment les fonctions de croyance qui y sont associées Si cette incertitude est souvent prise en compte dans des travaux (Yang et Han (2016), Deng et al. (2017)), la propagation de celle-ci au sein d'une fusion de données reste un enjeu. Cet article vise donc à développer une approche méthodologique de classification du niveau de pollution aux particules se basant sur l'utilisation de capteurs de pollution en utilisant d'une part la fusion de données via la théorie de l'évidence (ex : Tong et al. (2021)) et d'autre part sur une méthode de propagation de l'incertitude visant à délivrer la meilleure information possible.

2 Concepts théoriques

2.1 Fusion de données

La fusion de données représente l'action de combiner différentes données issues de sources diverses. Elle peut être de plusieurs types (Bloch, 2003) : fusion bayésienne (qui se base sur la théorie des probabilités), fusion évidentielle (qui se base sur les travaux de Dempster Shaffer - Dempster (1967)) et la fusion floue qui se base sur la logique floue. Le présent article se concentre sur la fusion évidentielle. L'utilisation de la théorie de l'évidence dans la fusion de données est assez commune et touche de nombreux domaines (imagerie médicale, reconnaissance aérienne, robotique, etc.). Certains articles récents l'appliquent même dans le domaine de la pollution (Rahmati et Melesse, 2016). Le principe de la fusion de données évidentielle consiste en l'utilisation de fonction de croyance (belief functions) et utilise la notion de masse (ex : Tong et al. (2021)). La fusion évidentielle comporte trois étapes principales , la première consiste dans la modélisation des fonctions de masses. Quelque soit un univers des possibles Ω, la fonction de masse m se définit comme la croyance dans la survenance d'un évènement ω_0 dans un sous ensemble A $\subseteq \Omega$ ou Ω représente l'univers des possibles et vérifie la condition suivante :

$$\sum_{A \subseteq \Omega} m(A) = 1 \quad avec \quad m(A) \in [0,1] \tag{1}$$

La deuxième étape consiste dans la combinaison des différentes sources via la règle de Dempster. Il existe de nombreux types de combinaisons qui dépendent du niveau de confiance (reliability) accordable aux différentes sources (Osswald et Martin, 2006). Ici, en cas de données trés fiables, on utilisera la combinaison conjonctive (2), Sinon il convient d'utiliser la combinaison disjonctive (3).

$$\forall A \subseteq \Omega, (m_1 \oplus m_2)(A) = \sum_{B \cap C = A} m_1(B).m_2(C) \tag{2}$$

$$\forall A \subseteq \Omega, (m_1 \oplus m_2)(A) = \sum_{B \cup C = A} m_1(B).m_2(C) \tag{3}$$

Enfin, la dernière étape consiste en la prise de décision, celle-ci peut se faire de différentes manière en fonction de la valeur de fonctions : la credibilité (BEL) et la plausibilité (PLS).

$$\forall A \subseteq \Omega, \quad \textit{BEL}(A) = \sum_{\emptyset \neq B \subseteq A} m(B) \qquad et \quad \textit{PLS}(A) = \sum_{B \cap A \neq \emptyset} m(B) \tag{4}$$

2.2 Incertitude

Toute information issue d'une source et quel que soit son format possède une imprécision et une incertitude : *L'imprécision* peut concerner l'ensemble des attributs d'une donnée, par exemple, dans le cas de capteurs, la position, l'horodatage ou encore la mesure en elle-même. Elle peut se définir comme l'écart entre la valeur réelle et la valeur mesurée d'un attribut. *L'incertitude* quant à elle peut se définir comme la véracité d'une mesure ou comme son degré de confiance. Dans le cas des capteurs, l'incertitude est liée à la position du capteur, à son environnement, à l'utilisation qui en est faite.

Du point de vue de la fusion évidentielle, l'incertitude de la mesure revêt un caractère primordial quant au choix de la combinaison (conjonctive ou disjonctive), mais également dans le calcul des masses. En effet, l'ignorance, i.e. la masse appliquée à l'univers des possibles Ω dépend directement de l'incertitude (ici on pourrait presque parler de confiance) liée à la source. Le problème du calcul d'incertitude réside d'une part dans la localisation de ses sources mais également dans la conjonction de plusieurs de ces sources comme génératrices d'incertitude.

3 Méthodologie

La présente section vise à construire une méthode permettant de fusionner efficacement trois sources de données de format, et de résolutions différentes, tout en minimisant l'incertitude.

3.1 Données

Le but de ce travail réside dans la quantification des particules fines de diamètre inférieur à 2.5 μm ($PM_{2.5}$) sur la ville de Sherbrooke au Quebec (Canada) à l'aide de capteurs de pollution de différents types : capteurs fixes (que l'on notera S_1 pour la suite) ; capteurs mobiles (S_2) placés sur les bus de la ville (ex : **figure 1**) ; capteurs "Citoyens" (S_3). Ces capteurs fonctionnent de manière optique, ils analysent un flux d'air puis par algorithme en déduisent la concentration en $PM_{2.5}$ au sein de ce volume. Les données possèdent toutes, quelque soit la source, les mêmes attributs : la position (X,Y) ; l'horodatage (t) ; la vitesse instantanée (nulle dans le cas de S_1) ; l'identifiant unique du capteur (qui correspond également a l'identifiant de l'utilisateur pour S_3) ; la mesure de $PM_{2.5}$ (notée Val_{PM}).

3.2 Imprécision de mesure et impact sur le choix de classe

Dans un premier temps, il convient de déterminer les classes permettant d'adjoindre un niveau de pollution à partir des recommandations annuelles de l'OMS (Tab1). Chaque capteur

	I	II	III	IV	V	VI
Classe de $PM_{2.5}$ ($\mu g/m^3$)	[$\leq$5]	[5,10]	[10,15]	[15,25]	[25,35]	[$\geq$35]

TAB. 1 – *Classes de pollutions dérivées des recommandations de l'OMS.*

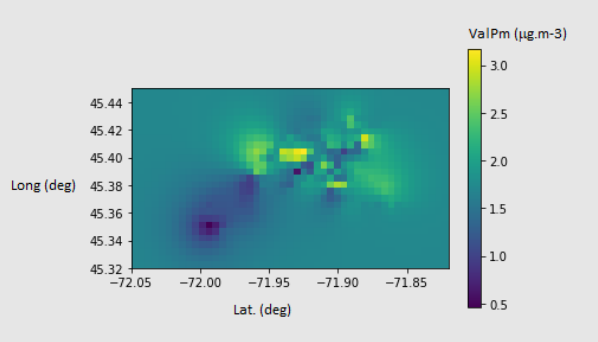

FIG. 1 – *Représentation de l'information issue du krigeage des données mobiles S2 sur une journée.*

possède sa propre imprécision de mesure donnée par le constructeur. Cette imprécision notée ΔMes impacte donc potentiellement chaque mesure, i.e.son appartenance a l'une des classes. Pour résoudre le problème, on peut donc jouer sur le concept de masse. Ainsi pour un capteur parfait on peut écrire :

$$\forall C \subset \Omega, \quad m(C) = 1 - m(\Omega) \tag{5}$$

Avec $\Omega = \{[0,5],[5,10]...[35,\infty]\}$ l'ensemble des classes et $m(\Omega)$ l'ignorance liée à la source (voir **3.3**).

Dans le cas d'un capteur possédant une imprécision même minime, il est possible que $Mes + \Delta Mes$ et $Mes - \Delta Mes$ ne soient pas strictement inclus dans les mêmes classes, auquel cas la probabilité que la mesure affichée soit dans une des 2 classes qui la jouxte est considérée comme identique et l'on implémente alors de la manière suivante :

$$\forall (C,D) \subset \Omega, \qquad Mes + \Delta Mes \in C \quad ET \quad Mes - \Delta Mes \in D,$$
$$\begin{cases} m(C) = \frac{1-m(\Omega)}{2} \\ m(D) = \frac{1-m(\Omega)}{2} \end{cases} \tag{6}$$

3.3 Incertitude, et combinaison

La fusion évidentielle se base prioritairement sur le concept de fonction de masse. La masse d'une classe C correspond à la confiance et donc dépend de l'imprécision de la mesure. Néanmoins, une source peut être précise dans ses attributs et pour autant ne pas être fiable, on parle alors d'incertitude. Ici l'incertitude sera déterminée par la masse de l'ensemble des possibles Ω. En effet la masse accordée à Ω correspond à la confiance qu'on donne au fait que la valeur de la mesure soit en fait non contenue dans une classe, mais contenue potentiellement dans toutes les classes. Dans notre cas, l'ensemble des classes correspondant à un espace infini, $m(\emptyset) = 0$, l'incertitude correspond à $m(\Omega)$, en effet les fonctions de masse m sont définies sur 2^{Ω} et $\emptyset$ correspond à l'ouverture au monde.

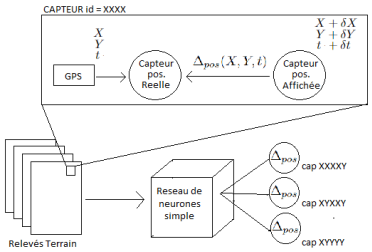

FIG. 2 – *Calcul de $\Delta_{pos}(X, Y, t)$ par apprentissage profond.*

Sources d'incertitude et prise en compte Parmi les attributs de la donnée, la position et l'horodatage sont nativement incertains pour S_2 et S_3, de plus S_3 est utilisé par des citoyens ce qui engendre également une incertitude difficilement quantifiable. Il est impossible de connaître à priori, le niveau d'incertitude issu de ces éléments, sa quantification doit donc se faire soit à priori empiriquement (i.e. lui appliquer une valeur arbitraire) soit, et c'est l'option choisie ici, par apprentissage.

Ainsi pour chaque type de source, l'ignorance s'implementera de la manière suivante :

$$\begin{aligned} m_{S_1}(\Omega) &= 0 \\ m_{S_2}(\Omega) &= Inc_{Pos}(v, \delta X, \delta Y, \delta t) \\ m_{S_3}(\Omega) &= Inc_{Pos}(v, \delta X, \delta Y, \delta t) + Inc_{User}(Id) \end{aligned} \tag{7}$$

Avec $\{\delta X, \delta Y, \delta t\}$, l'écart entre la position réelle et la position mesurée (voir paragraphe suivant), v la vitesse instantanée du capteur, Id l'identifiant du capteur, Inc la fonction d'incertitude calculée. Pour les capteur fixes, l'ignorance est nulle pour les sources d'incertitude mentionnées ci-dessus, néanmoins, il peut exister d'autres sources d'incertitudes non étudiées ici, relatives au capteurs fixes telles que l'humidité ou le vent, qui peuvent affecter ses mesures.

Incertitude positionnelle L'incertitude relative à la position du capteur est complexe à implémenter car elle dépend avant tout de la "dérive" du capteur, c'est à dire de l'écart entre la valeur mesurée et la position spatio-temporelle réelle du capteur. Il faut donc connaître pour chaque capteur la valeur de $\Delta_{pos}(X, Y, t)$ $=\{\delta X, \delta Y, \delta t\}$, où X, Y représentent la position spatiale et t l'horodatage. Pour ce faire, la méthode choisie va utiliser l'apprentissage profond par réseau de neurones convolutifs (*CNN*). Le jeu de données d'entrée comporte simplement de 2 tableaux à 3 colonnes (voir **figure 2**) ayant chacun les coordonnées (X,Y,t) de contrôle : GPS des stations fixes (pour les capteurs citoyens) et GPS du bus (pour les stations mobiles). Une fois les valeurs $\{\delta X, \delta Y, \delta t\}$ déterminées il convient de mesurer les impacts sur l'incertitude de la mesure de $PM_{2.5}$ engendrée par les écarts de position. Il s'agit ici de déterminer pour un écart positionnel donné l'écart induit sur la valeur de $PM_{2.5}$ mesurée. La **figure 3** montre le principe général. Il s'agit ici donc de créer un nouveau réseau de neurones dont les données d'entrées prendraient en compte les valeurs de $PM_{2.5}$ des stations fixes et des stations mobiles, leur positionnement absolu (comprenant les valeurs de $\Delta_{pos}(X, Y, t)$), et, la vitesse instantanée du capteur. En sortie il y aurait donc une fonction qui donnerait l'écart maximal

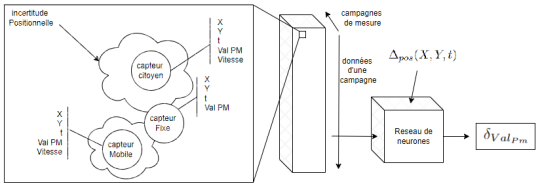

FIG. 3 – *Calcul de l'incertitude positionnelle.*

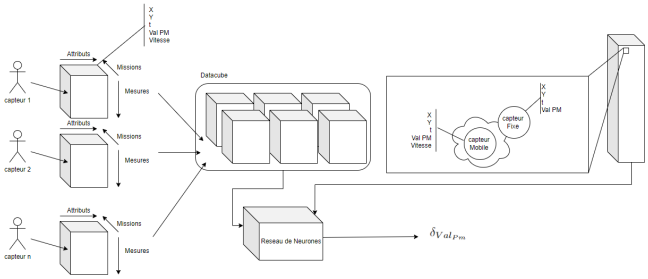

FIG. 4 – *Calcul de l'incertitude d'utilisation.*

de valeur de $PM_{2.5}$ (noté $\delta^{Pos}_{Val_{Pm}}$) en fonction de la vitesse du capteur et de son incertitude positionnelle. En normalisant cette valeur on obtient alors l'incertitude relative à la position et la vitesse par capteur (en pourcentage).

$$Inc_{Pos}(v, \delta X, \delta Y, \delta t) = \delta^{Pos}_{Val_{Pm}} \tag{8}$$

Incertitude d'utilisation Pour ce qui est de l'incertitude liée à l'utilisation, de la même manière, il est possible de combiner les information de valeurs de $PM_{2.5}$ par utilisateur (par id de capteur) en les comparant avec d'autres utilisateurs , et les valeurs voisines de capteurs fixes et de capteurs mobiles. Ici aussi l'apprentissage profond servira de base au calcul (**figure 4**) car il permettra d'apprendre des campagnes de mesures de chaque utilisateur une incertitude moyenne sur la décision. On obtient alors $\delta^{Use}_{Val_{Pm}}$ qui n'est plus fonction de la vitesse ou de la position mais seulement reliée à l'utilisateur.

$$Inc_{User}(Id) = \delta^{Use}_{Val_{Pm}} \tag{9}$$

Combinaison, décision, indicateurs Pour réaliser la combinaison, si l'ignorance est nulle ou très faible, alors on utilisera la fonction conjonctive , à l'inverse si l'ignorance est forte on utilisera la fonction disjonctive. Soit $m_i(A)$ la masse de la classe A issue du capteur S_i on a :

$$\forall A \subseteq \Omega, (m_1 \oplus m_2 \oplus m_3)(A) = \sum_{B \cup C \cup D = A} m_1(B).m_2(C).m_3(D) = m_c(A) \tag{10}$$

Pour choisir entre deux classes issues des masses, compte tenu des hauts niveaux d'incertitudes prévus (notamment via les sources de capteurs citoyens), on devra choisir un mode un mode de décision par faible dominance (Denœux, 2019) : Ainsi, soit 2 classes C_1 et C_2 , C_1 dominera faiblement C_2 si $BEL(C_1) \geq BEL(C_2)$ et si dans le même temps $PLS(C_1) \geq PLS(C_2)$, auquel cas la classe choisie sera C_1. Il convient dès lors également de déterminer des indicateurs de performances d'une telle méthode afin de pouvoir la valider. Il existe de nombreux capteurs fixes sur le territoire (AERONET, LIDAR, etc.) qui permettraient de comparer les résultats de la fusion sur des points spécifiques. Néanmoins, la difficulté réside dans l'extension de ces contrôles sur l'ensemble du territoire. Dès lors l'étude va utiliser deux campagnes de mesures : la première menée en 2022 sera progressive (période capteur fixe, période capteurs fixe + mobiles, période capteurs fixes + mobiles + citoyens) permettra de calculer les incertitudes liées à chaque source et l'indicateur de performance liée au temps de calcul, la deuxième en 2023 permettra de vérifier les indicateurs de performance finaux que sont l'écart entre les valeurs de $PM_{2.5}$ mesurées par les autres capteurs et le modèle aux points fixes.

4 Discussion et perspectives

Au sein de la méthode présentée ici, que ce soit pour l'incertitude positionnelle ou l'incertitude liée à l'utilisation des capteurs, la détermination des incertitudes suit un modèle semblable (données, réseau de neurone, calcul d'incertitude). Néanmoins, il n'est traité ici que deux grands types d'incertitude liées aux capteurs. Or dans le cadre de la pollution atmosphérique se limiter à ces simples domaines n'est pas suffisant car il existe de nombreux facteurs pouvant impacter directement ou indirectement la mesure et l'incertitude de la source. Dans ce contexte la méthode utilisée dans cet article pourrait être dérivée (après test sur des données) de la manière suivante : Soit une source de donnée quelconque S, p les paramètres de la source S $(pos, hor., valPM, \Delta_{Mes})$, Ω_p : l'ensemble des facteurs impactant le paramètre p.

- **Etape 1** : Déterminer par réseau de neurones Ω_p du parametre p de la source S et joindre un facteur d'importance a chaque élément d'Ω_p ;
- **Etape 2** : Déterminer l'incertitude globale (l'ignorance) de la source en fusionnant les incertitudes d'Ω_p ;
- **Etape 3** : Utiliser cette ignorance dans la fusion multi-sources.

Malgré tout, l'incertitude ainsi calculée si elle cherche à se rapprocher le plus possible d'une réalité terrain se heurte à la gestion dite des artefacts, i.e.des éléments extérieurs temporaires ayant un impact direct mais très peu prévisible sur les données. L'une des perspectives à explorer après la mise en place et la validation de la méthode ci-dessus consisterait donc en l'ajout de données supplémentaires (hétérogènes).

Références

Bloch, I. (2003). *Fusion d'informations en traitement du signal et des images*. Hermes Science Publications.

Dempster, A. (1967). Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Stat. 38*, 325–339.

Deng, X., F. Xiao, et Y. Deng (2017). An improved distance-based total uncertainty measure in belief function theory. *Applied Intelligence 46*(4), 898 – 915.

Denœux, T. (2019). Decision-making with belief functions : A review. *International Journal of Approximate Reasoning 109*, 87 – 110.

Hariri, R. H., E. M. Fredericks, et K. M. Bowers (2019). Uncertainty in big data analytics : survey, opportunities, and challenges. *Journal of Big Data 6*(1). All Open Access, Gold Open Access.

Khaleghi, B., A. Khamis, F. O. Karray, et S. N. Razavi (2013). Multisensor data fusion : A review of the state-of-the-art. *Information Fusion 14*(1), 28–44.

OMS (2021). *WHO global air quality guidelines : particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide : executive summary*. World Health Organization.

Osswald, C. et A. Martin (2006). Understanding the large family of dempster-shafer theory's fusion operators - a decision-based measure. In *2006 9th International Conference on Information Fusion*, pp. 1–7.

Piedrahita, R., Y. Xiang, N. Masson, J. Ortega, A. Collier, Y. Jiang, K. Li, R. Dick, Q. Lv, M. Hannigan, et L. Shang (2014). The next generation of low-cost personal air quality sensors for quantitative exposure monitoring. *Atmospheric Measurement Techniques 7*(10), 3325 – 3336.

Rahmati, O. et A. M. Melesse (2016). Application of dempster–shafer theory, spatial analysis and remote sensing for groundwater potentiality and nitrate pollution analysis in the semi-arid region of khuzestan, iran. *Science of the Total Environment 568*, 1110 – 1123.

Tong, Z., P. Xu, et T. D. ux (2021). An evidential classifier based on dempster-shafer theory and deep learning. *Neurocomputing 450*, 275–293.

Yang, Y. et D. Han (2016). A new distance-based total uncertainty measure in the theory of belief functions. *Knowledge-Based Systems 94*, 114 – 123.

Summary

Atmospheric pollution control is becoming a key public health issue for our 21st century. The growing diversification of sensors and the growing variety and number of data they produce, lean some important issues for qualifying and processing information. Belief functions theory and data fusion could be a part of the solution of these issues. By the way the fact that fusion take each sensor in consideration for global information is a threat to the global uncertainty of merged information. This paper aims at building a method using belief theory and deep learning to get a merged information from a known attribute data model, while reducing uncertainty. Beyond the method, this paper aims at creating performance indicators to post validate the model.