APPARIEMENT DE DONNÉES GÉOGRAPHIQUES UTILISANT LA THÉORIE DES CROYANCES

par Ana-Maria Raimond

Laboratoire COGIT
Institut géographique national,
2-4 avenue Pasteur 94165 Saint-Mandé Cedex
ana-maria.olteanu@ign.fr

Dans un contexte général d'intégration de bases de données géographiques, nous présentons dans cet article une approche d'appariement de données, basée sur la théorie des fonctions de croyance. Les don - nées géographiques présentent des imperfections et celles-ci doivent être prises en compte dans le proces - sus d'appariement de données. Afin d'apparier les données géographiques, nous avons défini et combiné trois critères d'appariement de données, basés sur la géométrie, l'information sémantique et l'information toponymique. Nous avons testé notre approche sur deux jeux de données représentant les points remar - quables du relief. Les résultats ont été évalués en termes de précision et de rappel ; des valeurs de précision et de rappel proches des 100% ont été obtenues.

Mots-clés : *appariement de données, théorie des fonctions de croyance, ontologie.*

1Introduction

L'intégration de bases de données géographiques est un sujet qui suscite un intérêt dans le monde de l'information géographique depuis plusieurs années. Actuellement, il existe de nombreuses bases de données géographiques (BDG) qui couvrent le même territoire du monde réel à des échelles géométriques et sémantiques différentes. La multiplicité des BDG est due à l'existence d'un nombre croissant de données géographiques. En effet, leur saisie se fait plus facilement grâce à l'arrivée de nouveaux outils, les besoins en données géographiques précises sont en pleine croissance et les rythmes de mise à jour sont différents selon les thèmes et besoins. Les données géographiques sont modélisées par des géométries différentes (par exemple, une rivière peut être modélisée par une géométrie linéaire ou bien par une géométrie surfacique), elles sont destinées à répondre à plusieurs applications (visualisation, analyse) et elles proviennent de différents modes d'acquisition (sources, processus). Il y a une indépendance entre les bases de données géographiques existantes (Ruas 2002), ce qui pose certains problèmes à la fois aux producteurs et aux utilisateurs.

Dans ce contexte, le processus d'intégration de données semble être une solution, en répondant d'une part aux besoins des producteurs (la mise à jour, l'évaluation de la qualité des données ainsi que la détection des incohérences), et d'autre part aux besoins des utilisateurs (étude de différentes zones adjacentes ou pour faciliter les analyses mêlant différents points de vue).

Tous ces besoins conduisent à la fois les producteurs et les utilisateurs à vouloir établir des liens entre les bases de données géographiques, processus nommé appariement de données. Cet article se concentre sur l'appariement de données géographiques en s'appuyant sur des connaissances imparfaites qui viennent des spécifications ou des données elles-mêmes. Nous avons utilisé la théorie de fonctions de croyance (Shafer 1976) pour fusionner plusieurs connaissances provenant de différents critères tels que la géométrie, la toponymie et la sémantique, afin de trouver les correspondances entre les objets géographiques homologues appartenant à deux BDG à différentes échelles. Notons que dans la théorie des croyances, le terme de source d'information est utilisé pour définir les connaissances d'une source de données, tandis que dans cet article nous utilisons le terme de critère d'information.

L'article est organisé de la manière suivante. Dans un premier temps, nous situons notre travail par rapport aux travaux déjà existants. La problématique

étant définie, nous décrivons dans la section 3 notre approche basée sur la théorie des fonctions de croyance. L'initialisation des masses de croyance est abordée dans la section 4. Enfin, des résultats sont présentés dans la section 5.

2 État de l'art sur l'appariement de données géographiques

L'appariement de données géographiques est un outil qui permet d'associer les données de deux ou plusieurs BDG et de produire des liens explicites entre les objets homologues (Walter et Fritsch 1999). Il est utilisé dans plusieurs applications telles que l'intégration de données géographiques (Devogele 1997; Mustière 2006), la mise à jour automatique (Gombosi et al. 2003), l'analyse de la qualité des données (Bel Hadj Ali 2001) ou bien la détection des incohérences entre les BDG (Sheeren et al. 2004).

Dans la littérature il existe de nombreux algorithmes qui s'avèrent efficaces pour certains types de données ou dans des zones particulières. Les algorithmes d'appariement s'appuient sur plusieurs facteurs dont : la géométrie et les attributs des objets, les relations topologiques existant entre les objets et l'échelle de la base de données. Dans cette section, nous présentons différents algorithmes d'appariement de données géographiques existant dans la littérature. Nous distinguons deux types d'approches : d'une part les approches pour les points isolés, c'est-à-dire des données qui sont indépendantes les unes des autres, et d'autre par les approches pour les réseaux.

L'appariement de données isolées est basé principalement sur des mesures de distances entre les géométries. (Bel Hadj Ali 2001) propose un algorithme d'appariement adapté aux données surfaciques s'appuyant sur la géométrie et sur des mesures prenant en compte les surfaces et les contours, telles que l'intersection, la distance surfacique, etc. Dans (Beeri et al 2004), une approche probabiliste est adoptée, basée sur des critères purement géométriques. L'appariement peut aussi comparer les noms des objets lorsque des toponymes sont présents (Levensthein 1965; Cohen et al. 2003). Pour les réseaux linéaires, plusieurs algorithmes d'appariement ont été proposés dans la littérature. (Walter et Fritsch 1999) propose une méthode statistique qui apparie des réseaux routiers de deux BDG différentes à la même échelle, et qui est basée sur des critères géométriques et topologiques. D'une manière générale, les algorithmes d'appariement sont spécifiques aux données et aux BDG à apparier et ils sont basés sur des critères différents. Ainsi, il existe des algorithmes qui s'appliquent aux BDG représentant une même réalité à des niveaux d'abstraction différents (Devogele 1997 ; Mustière 2006) ou au même niveau d'abstraction (Voltz 2006).

La comparaison de la sémantique au niveau des classes est nécessaire pour apparier les schémas et elle est également utile pour apparier les données, même si elle est très peu utilisée. La comparaison de la sémantique s'appuie sur les ontologies (Gesbert 2005). Nous pouvons noter que les méthodes d'appariement, qu'elles soient appliquées sur des données ponctuelles, linéaires ou surfaciques ou qu'elles soient utilisées pour apparier des jeux de données à la même échelle ou à des échelles différentes, sont basées sur un enchaînement de différents critères. Ces derniers, fondés sur la géométrie, sur les attributs ou sur les graphes, sont appliqués, en général, l'un après l'autre. De plus, la majorité des approches ne prennent pas en compte les imperfections dans les données.

En conséquence, notre objectif est de trouver une méthode d'appariement de données qui prend en compte toutes les imperfections (incertitude, incomplétude, imprécision) et tous les critères en même temps. Nous considérons que la théorie des croyances est pertinente pour atteindre ces objectifs, parce que, d'une part, elle modélise toutes les imperfections y compris l'incomplétude, et que, d'autre part, elle permet de combiner les critères et les hypothèses afin de prendre une décision.

3 Le contexte général de la théorie des croyances

La théorie des croyances, nommée aussi le modèle de Dempster-Shafer ou la théorie de l'évidence, a été introduite par Shafer (Shafer 1976) à la suite des travaux de Dempster sur les probabilités inférieure et supérieure (Dempster1967), en se basant sur des fonctions de croyance.

3.1 Le cadre de discernement

La théorie des croyances considère un univers de référence appelé le cadre de discernement θ, θ ={H1, H2,…, HN}, composé d'un ensemble de N hypothèses. À partir du cadre de discernement,

notons 2^θ l'ensemble de tous les sous-ensembles de θ défini de la manière suivante: (1)

$$2^\Theta = \{\phi, \{H_1\}, \{H_2\}, \{H_1, H_2\} \ldots \{H_1 \ldots H_N\}, \Theta\}$$

où, {Hi, Hj} représente l'hypothèse que la solution à un problème donné est une des deux c'est-à-dire soit Hi soit Hj. Nous appelons cette hypothèse une proposition.

La théorie des croyances est basée sur des fonctions de croyance. Une fonction de croyance associe à une proposition, $A \in 2^\theta$, une valeur nommée masse de croyance et notée m(A) qui représente le degré avec lequel on croit en cette proposition. Par exemple, si nous considérons que le processus d'appariement est basé sur la géométrie des objets géographiques, plus les objets à comparer sont proches, plus on croit qu'ils sont homologues et en conséquence la masse de croyance a une valeur élevée. Les fonctions de croyance sont définies de la manière suivante : (2)

$$m : 2^\Theta \rightarrow [0,1],$$
$$\sum_{A \subseteq \Theta} m(A) = 1$$

Toute proposition $A \in 2^\theta$, telle que m(A)>0, est nommée élément focal. Nous considérons seulement les éléments focaux afin de combiner l'information et de prendre une décision. Notons que, lorsque les éléments focaux se réduisent aux singletons H_i, la notion de masse de croyance est assimilable à celle de probabilité.

3.2 L'opérateur de combinaison de Dempster

La théorie des croyances permet la fusion de plusieurs critères (par exemple la géométrie, la nature) en employant l'opérateur de combinaison de Dempster. Supposons deux sources d'information S_1 et S_2. La source d'information S_1 (respectivement S_2) soutient une proposition avec une masse de croyance $m_1(A)$ (respectivement $m_2(A)$). Notons m_{12} la masse résultante de la combinaison de ces deux sources soutenant la même proposition A. Par exemple, afin de décider si deux objets géographiques appartenant à deux BDG différentes doivent être appariés ou pas, nous considérons deux critères: un critère géométrique et un critère toponymique qui compare les toponymes. Sous l'hypothèse que les deux objets sont homologues, le premier critère croit que c'est le cas parce que géométriquement les objets sont très proches et il attribue à cette hypothèse une masse de croyance importante, alors que le deuxième critère n'est pas sûr parce que les deux toponymes ne sont pas similaires et donc il attribue une masse de croyance moins importante à cette hypothèse. Afin de prendre une décision, les deux critères sont combinés en utilisant l'opérateur de Dempster de la manière suivante : (3 et 4)

$$m_{12}(A) = m_1(A) \otimes m_2(A) = \frac{1}{1 - m_{12}(\varnothing)} \sum_{\substack{B \cap C = A \\ B,C \in 2^\Theta}} m_1(B) m_2(C)$$

où

$$m_{12}(\varnothing) = \sum_{\substack{B \cap C = \varnothing \\ B,C \in 2^\Theta}} m_1(B) m_2(C)$$

Lorsque les critères sont combinés, il est possible que les deux critères soient en conflit. Dans ce cas, le conflit est attribué à l'ensemble vide, conformément à l'équation 4, et il est utilisé dans le cadre de l'opérateur de Dempster pour normaliser la masse de croyance combinée, m_{12}. Ainsi, la masse de croyance associée au conflit est redistribuée proportionnellement aux éléments focaux (Shafer 1976 ; Smets 1988).

4 La théorie des croyances dans un contexte d'appariement de données spatiales

D'une manière générale, le processus d'appariement de données consiste, pour chaque objet appartenant à une BDG dite de référence, à rechercher ses homologues dans l'autre BDG dite de comparaison. Les données géographiques présentent des imperfections (par exemple la localisation peut être imprécise, les toponymes peuvent présenter des dissimilitudes en raison de la variabilité linguistique, l'utilisation du nom officiel et du nom d'usage pour la même entité géographique, etc.).

En utilisant des critères d'appariement en série, des erreurs peuvent se propager et donc le résultat d'appariement peut être erroné. En conséquence, l'imperfection doit être prise en compte dans le processus d'appariement et les critères doivent être appliqués en même temps afin d'obtenir une information plus pertinente. La théorie des croyances offre les outils nécessaires pour modéliser l'imperfection à travers les fonctions de croyance et fusionner différentes connaissances à travers l'opérateur de Dempster.

Dans cette section, nous décrivons notre approche basée sur la théorie des croyances développée en trois étapes (Olteanu 2007) :

- La première étape consiste à initialiser les masses de croyance pour chaque candidat à l'appariement et pour chaque source. Dans notre cas, une source est un critère guidant l'appariement, comme la proximité spatiale ou la ressemblance des toponymes.

- La deuxième étape consiste à fusionner les masses de croyance par candidat.

- Enfin, la troisième étape est basée sur une combinaison des masses de croyance résultant de la deuxième étape, ce qui consiste à fusionner les candidats entre eux.

Les deux premières étapes sont considérées comme une approche locale, parce que les candidats sont traités indépendamment les uns des autres, alors que la troisième est considérée comme une approche globale parce que tous les candidats sont analysés ensemble.

4.1 L'approche locale: définition du cadre de discernement

Nous considérons deux BDG construites pour des besoins d'utilisation différents, provenant de sources différentes et qui présentent différents niveaux de détail. Dans ce papier, nous proposons une méthode d'appariement qui calcule des liens entre les objets dans un sens, de la base moins détaillée vers la base plus détaillée. Ainsi, pour chaque objet appartenant à la base moins détaillée, appelé objet de référence, tous les objets de l'autre base plus détaillée qui se trouvent à une certaine distance (choisie empiriquement) sont sélectionnés et sont appelés objets candidats à l'appariement.

Dans notre cas, nous définissons un cadre de discernement local pour chaque objet de référence et tous les candidats sont des hypothèses, donc des homologues potentiels. Nous avons constaté qu'il y a des objets qui n'ont pas d'homologues dans l'autre base et donc ils ne sont pas appariés. Cela nous amène à définir une nouvelle hypothèse NA signifiant la solution « l'objet n'est pas apparié ». Ainsi, le cadre de discernement est exhaustif, c'est-à-dire la solution se trouve parmi les hypothèses définies et l'ensemble vide n'est utilisé que pour modéliser le conflit dû à la non-fiabilité des sources. Une approche similaire a été adoptée par (Royère 2002) dans le cadre d'une application de localisation d'un véhicule sur une carte.

Le cadre de discernement pour un objet de référence est défini ci-après : (5)

$$\Theta_{REF} = \{C_1, C_2, ..., C_N, Nm\}$$

où N représente le nombre de candidats et C_i représente l'hypothèse que C_i est l'homologue de l'objet de référence en cours d'analyse.

Afin de calculer les fonctions de croyance, nous définissons une approche locale : chaque candidat est analysé indépendamment des autres. En conséquence, nous modélisons les connaissances en utilisant des sources spécialisées : chaque source se spécialise et se prononce sur une seule hypothèse (Appriou 1991).

Étant donné 2^θ, nous définissons S_i, un sous-ensemble de 2^θ de la manière suivante : (6)

$$S_i = \{C_i, \neg C_i, \Theta\}$$

- C_i représente l'hypothèse que l'objet de référence en cours d'analyse est apparié avec le candidat C_i.

- $\neg C_i = \{C_1, C_2, ...C_{i-1}, ...C_N, NA\}$ représente l'hypothèse que l'objet de référence en cours d'analyse est apparié avec un autre candidat que C_i ou pas apparié du tout.

- $\theta = \{C_1, C_2, ...C_i, ...C_N, NA\}$ représente l'hypothèse que le critère ne peut pas se prononcer sur ce candidat, signifiant l'ignorance.

4.2 L'initialisation des masses de croyance

Dans cet article, nous proposons trois critères d'appariement de données. Ils représentent les sources d'informations définies dans le cadre de la théorie des croyances et sont présentés ci-dessous.

4.2.1 Le critère géométrique

Le critère géométrique s'appuie sur la distance euclidienne, d_E, entre la localisation de l'objet de référence et celle du candidat à l'appariement. Nous considérons que plus le candidat est proche, plus il y a des chances qu'il soit l'homologue de l'objet de référence, comme le montre la figure 1a). Dans la figure 1a), T_2 représente le seuil de sélection des candidats, qui représente la distance maximale de recherche de candidats, et T_1 définit le seuil de confiance qui associe une croyance moins forte aux candidats éloignés géométriquement de l'objet de référence en cours d'analyse. L'imprécision de la localisation des objets géographiques fait qu'il peut y avoir des homologues qui sont relativement éloignés. Pour éviter d'éliminer complètement un candidat qui est loin de l'objet de référence, nous considérons que la masse de croyance attribuée à l'hypothèse « le candidat C_i n'est pas l'objet homologue » n'est jamais 0, mais qu'elle varie de 1 à 0,1.

4.2.2 Le critère toponymique

Le deuxième critère consiste à comparer les toponymes des objets de deux BDG en utilisant la distance de Levenshtein, d_L (Levenshtein 1965), calculée de la manière suivante : (7)

$$d_T = \frac{d_L(toponyme_1, toponyme_2)}{\max(L_1, L_2)}$$

où d_L représente la distance de Levenshtein, L_1 représente la longueur du toponyme$_1$ et L_2 représente la longueur du toponyme$_2$. Précisons que nous utilisons le terme de distance pour évaluer la ressemblance entre deux toponymes, et non pas dans le sens mathématique du mot distance. À titre d'exemple, étant donné deux toponymes « boulevard du général de Gaulle » et « bld du g^{al} de Gaulle » la distance d_T est égale à 0,7.

Comme nous pouvons le remarquer dans la figure 1b), les courbes sont différentes de celles du critère géométrique, afin d'exprimer le fait que nous sommes moins confiants en ce critère. En conséquence, nous gérons le cas d'ambiguïté, lorsque par exemple deux toponymes indiquant le même objet du monde réel sont comparés, l'un étant le nom officiel et l'autre le lieu-dit, par exemple « place Charles de Gaule » et « place de l'Étoile ». Pour cela, nous proposons de diminuer la masse de croyance attribuée à l'hypothèse $\neg C_i$ (ce n'est pas le candidat l'objet homologue) et d'augmenter la masse de croyance attribuée à l'ignorance, θ. Ainsi, si la distance d_T est supérieure au seuil T_1 (par exemple 30% des lettres ne se ressemblent pas), les masses de croyance attribuées à l'hypothèse -C_i est l'objet candidat homologue- et à l'hypothèse -le critère ne sait pas-, sont égales à 0,5.

4.2.3 Le critère sémantique

L'analyse détaillée des données géographiques montre qu'il y a des objets géographiques qui ont le même toponyme, qui sont proches les uns des autres, mais qui ne possèdent pas la même nature et en conséquence ne peuvent pas être mis en correspondance, comme par exemple un sommet avec un col. Ainsi, nous définissons un troisième critère qui utilise des propriétés sémantiques.

Dans la figure 1c), nous illustrons les modélisations des fonctions de croyance pour le critère sémantique. Ce critère n'est pas le critère le plus important, il est possible qu'il existe beaucoup de candidats qui ont la même nature que l'objet de référence. C'est pour cela que nous considérons que si la distance sémantique entre l'objet de référence et un candidat à l'appariement est 0 (les objets sont homologues sémantiquement parlant), la masse de croyance attribuée à l'hypothèse C_i (c'est le candidat C_i l'objet homologue) est égale à 0,5, donc le critère n'attribue pas une forte croyance à ce candidat. Au contraire, si la distance sémantique est supérieure au seuil T1, le critère croit que le candidat C_i n'est pas le bon candidat.

4.3 Combinaisons des critères et des candidats

Une fois que les masses de croyance ont été initialisées, nous pouvons combiner les critères par candidat en utilisant l'opérateur de Dempster. Cette approche est une approche locale parce que les candidats sont analysés séparément, sans prendre en compte les autres candidats. La troisième étape, nommée l'approche globale consiste à combiner les candidats entre eux, c'est-à-dire combiner entre eux les résultats obtenus pour chaque candidat dans l'étape précédente. Plus précisément, les résultats obtenus pour deux candidats sont combinés, ensuite le résultat avec le troisième candidat, et ainsi de suite.

Afin de prendre une décision, nous avons utilisé le maximum de probabilité pignistique. La décision est prise après l'étape de fusion globale et après avoir normalisé les masses résultantes en utilisant l'opérateur de Dempster. Dans ce papier nous abordons la problématique liée au calcul des liens de cardinalité 1-1. Conformément aux spécifications des deux bases de données utilisées pour tester notre approche, un objet géographique représente une réalité. Ainsi, le choix de la mesure basée sur la probabilité pignistique a été privilégié en raison de la cardinalité du lien souhaité : on privilégie les hypothèses simples.

5 Applications aux points remarquables de relief

Les tests ont été réalisés sur deux jeux de données de l'Institut géographique national, BDCARTO© (365 objets) et BDTOPO© (1965 objets) représentant les points remarquables du relief. Les jeux de données présentent des niveaux de détail différents, le premier étant moins détaillé que le deuxième.

Premièrement, les données, comme par exemple les montagnes, les sommets, les pics, les vallées, les

cols, etc., sont imprécises d'une part par définition : la limite entre une vallée et une montagne n'est pas parfaitement définie, et d'autre part parce que les différences entre les concepts utilisés dans les bases peuvent être flous, comme c'est le cas pour sommet et pic.

Deuxièmement, les objets homologues peuvent avoir différents toponymes, en particulier en raison de la variabilité linguistique issue des erreurs de frappe ou des différences de prononciation, par exemple « Munhoa » et « Monhoa », de l'imprécision, quand des entités possédant le même toponyme ont des localisations différentes, par exemple « place du général de Gaulle » à Paris et à Lyon, ou à cause de l'utilisation du nom officiel et du nom d'usage pour la même entité géographique.

Troisièmement, les concepts ne pressentent pas le même niveau de détail. Par exemple dans la BDCARTO il y a des concepts qui sont regroupés et représentés avec la même valeur de l'attribut nature : "sommet, crête, colline", alors que dans la BDTOPO ces concepts sont bien séparés.

En conséquence, utiliser seulement un critère basé sur la géométrie des objets ne donne pas de bons résultats parce que l'objet homologue n'est pas toujours l'objet le plus proche. De la même manière, utiliser seulement le critère toponymique ou sémantique peut engendrer des incohérences. Notre approche a été implémentée en Java dans le SIG open-source GeOxygene (Badard et Braun 2003).

6 Évaluation et résultats

L'évaluation des résultats est une étape très importante dans le processus d'appariement de données. Afin d'évaluer les résultats d'appariement, nous avons comparé ces derniers avec un appariement interactif. Les liens issus des deux modélisations du critère de comparaison des toponymes ont été évalués en terme de précision et de rappel, comme le montre le tableau 1. La précision représente le nombre de liens pertinents trouvés par rapport au nombre total des objets sélectionnés, alors que le rappel est le nombre de liens pertinents trouvés par rapport au nombre total d'objets pertinents (Beeri et al. 2004).

La première ligne du tableau 1 montre les valeurs de la précision et du rappel lorsque le processus d'appariement utilise seulement deux critères (géométrique et toponymique) et nous remarquons que 96.4% des liens d'appariement sont justes parmi ceux trouvés et que seulement 95.9% des objets ont été appariés. Ceci est dû aux deux cas de conflits apparus entre les critères, c'est-à-dire des objets possédant le même toponyme mais se trouvant très loin l'un de l'autre.

	Précision globale	Rappel global
Critères géométrique et toponymique	96.4 %	95.9%
Critères géométrique, toponymique et sémantique	99.1 %	99.1 %

Tableau 1 : Évaluation qualitative du processus d'appariement.

Lorsque nous utilisons les trois critères définis dans la partie 4.2 (deuxième ligne du tableau 1), nous remarquons que la précision et le rappel ont augmenté pour atteindre 99.1%. Notons que ces résultats sont satisfaisants et que le pourcentage de 0.9% de non-réussite est dû d'une part à un lien d'appariement 1 : m, ce problème de liens multiples n'étant pas traité dans cet article, et d'autre part à deux cas où les objets homologues devraient être appariés mais ne le sont pas, par erreur.

La figure 2 illustre un résultat d'appariement de données. Elle montre l'importance de l'information sémantique dans le processus d'appariement. Nous observons dans la figure 2 à gauche que le processus d'appariement n'apparie pas les objets homologues parce qu'ils sont éloignés et que leurs toponymes sont assez différents. Au contraire, les objets homologues sont appariés lorsque le critère sémantique a été rajouté.

7 Conclusion

Dans ce papier, nous avons présenté une approche d'appariement de données basée sur la théorie des croyances. Celle-ci conduit à combiner des critères d'appariement de données afin d'apparier des données qui présentent des imperfections. Nous avons testé notre approche sur des données géographiques réelles représentant des points remarquables du relief et nous avons comparé les résultats obtenus en utilisant d'une part deux critères (géométrique et toponymique) et d'autre part trois critères (géométrique, toponymique et sémantique). Les résultats ont été évalués en terme de précision et de rappel. Nous avons obtenu une précision et un rappel proche de 100% en utilisant les trois critères.

Quelques cas particuliers restent néanmoins à résoudre, tels que les appariements 1 à M ou les objets en conflit, pour lesquels nous n'avons pas pris de décision. Ainsi, une perspective de travail serait de développer un opérateur associatif de redistribution de conflit et d'introduire une nouvelle source d'information, telle que la nature des objets, pour augmenter le nombre d'objets appariés et améliorer les résultats.

Remerciements

L'auteur voudrait remercier Sébastien Mustière et Anne Ruas pour la relecture de ce papier et pour la qualité de leur remarques.

Bibliographie

Appriou A., 1991, « Probabilités et incertitudes en fusion de données multi-senseurs », *Revue scientifique de technique de la Défense*, 11, p.27-40.

Badard T., Braun A., 2003, « Oxygène une plate-forme inter-opérable pour le déploiement de services Web géographiques », *Bulletin d'information scientifique et technique de l'IGN*, n° 74, p. 113-120.

Bel Hadj Ali A., 2001, *Qualité géométrique des entités géographiques surfaciques –Application à l'appariement et définition d'une typologie des écarts géométriques*, thèse, Université de Marne-la-Vallée.

Beeri C., Kanza Y., Safra E., Sagiv Y., 2004, « Object Fusion in Geographic Information Systems », dans *Proceedings of the 30th VLDB Conference*, Toronto, Canada.

Bruns H.T., Egebhofer M, 1996, "Similarity of Spatial Scenes", dans *Seventh International Symposium on Spatial Data Handling, Delft*, Netherlands, August, London, Taylor & Francis, p.173-184.

Cohen W.W., Ravikumar P., Fienberg S.E., 2003, « A Comparison of String Distance Metrics for Name-Matching Tasks », dans *Proceedings of the IJCAI*, 9-10 August, Acapulco, Mexico, p. 73-78.

Dempster A., 1967, "Upper and lower probabilities induced by multivalued mapping", *Annals of Mathematical Statistics*, vol. AMS-38, p.325-339.

Devogele T., Parent C., Spaccapietra S., 1998, « On spatial database integration », *International Journal of Geographical Information Science*, 12(4), 1998, p. 335-352.

Gombo_i M., _alik B, Krivograd S., 2003, « Comparing two sets of polygons », *International Journal of Geographical Information Science*, 17 (5), p.431-443.

Levenshtein V., 1965, "Binary codes capable of correcting deletions, insertions and reversals", *Doklady Akademii Nauk SSSR*, 4 (163), p.845-848.

Mustière S., 2006, "Results of Experiments on Automated Matching of Networks at Different Scales", dans *ISPRS Workshop, Multiple representation and interoperability of spatial data*, Germany, 22-24 February, p. 92-100.

Olteanu A.-M., Mustière S., Ruas A., 2005, « Matching Imperfect Data », dans *Proceedings from 7th International Symposium on Spatial Accuracy Assessement in Natural Resources and Environmental Sciences*, 2006, p. 694-704.

Olteanu A.-M., 2007, "A Multi-Criteria Fusion Approach for Geographical Data Maching", dans *International Symposion in Spatial Data Quality (ISSDQ)*.

Royère C., 2002, *Contribution à la résolution du conflit dans le cadre de la théorie de l'évidence : application à la perception et à la localisation des véhicules intelligents*, thèse, Université de Compiègne.

Ruas A.,2002, *Généralisation et représentation multiple*, Lavoisier.

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

Sherren D., Mustière S., Zucker J-D., 2004, « How to Integrate Heterogeneous Spatial Databases in a Consistent Way? », dans *Conference on Advanced Databeses and Information Systems,* Budapest, September 2004, p. 364-378.

Smets Ph., 1988, *Belief Functions.Non Standard Logics for Automated Reasoning*, Smets Ph., Mamdani A., Dubois D. and Prade H. ed., London, Academic Press, p. 253-286.

Voltz S., 2006, "An Iterative Approach for Matching Multiple Representations of Street Data", dans *ISPRS Workshop, Multiple representation and interoperability of spatial data Hanover, Germany, 22-24 February, p. 101-110.*

Walter V., Fritsch D., 1999, « Matching Spatial Data Sets: Statistical Approach », *International Journal of Geographical Information Science*, 13(5), p. 445-473.

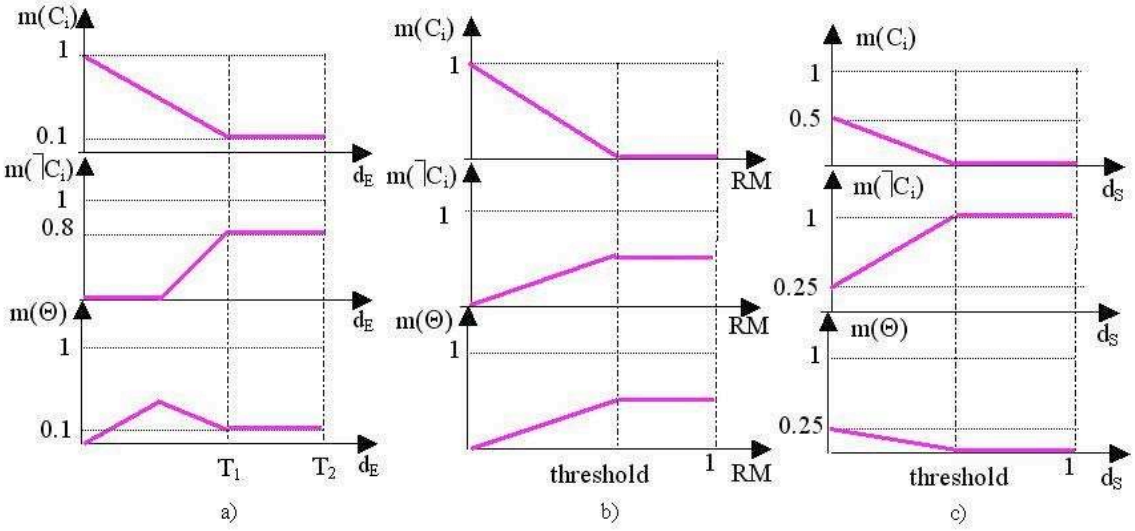

Figure 1 : Modélisation des critères, géométrique a), toponymique b) et sémantique c)

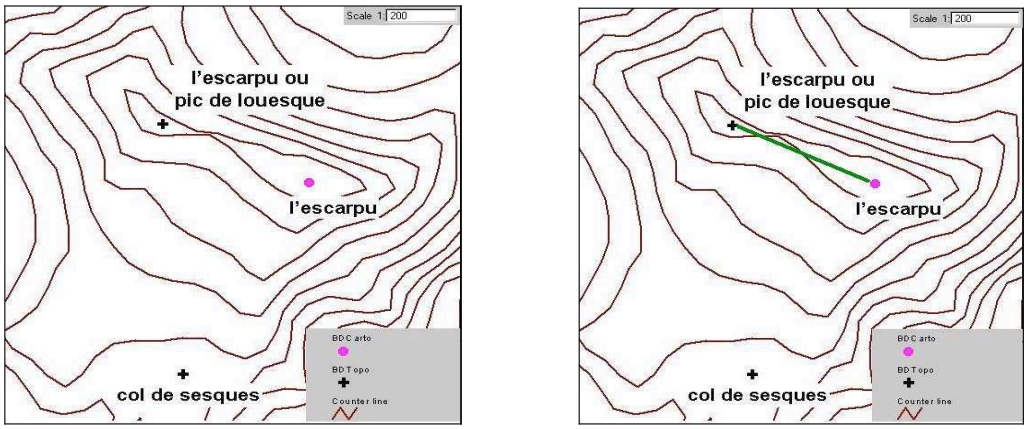

Figure 2 : Résultats d'appariement de données utilisant deux critères (à gauche) ou trois critères (à droite)