GéoPeuple

Rapport numéro	L2.3-1
Titre	Appariement Cassini - RGE (géométries ponctuelles)alop
Rédigé par	Benoit Costes (COGIT/IGN)
État (en cours / final)	Final
Relu par	Sébastien Mustière
Date	mars 2012

0.1 Préambule

Ce document présente la méthode utilisée dans le projet GéoPeuple pour mettre en relation les données issues de la vectorisation des feuilles de Cassini avec la BDTopo (base de données vectorielle de précision métrique, composante du Référentiel à Grande Échelle constitué par l'IGN).
Dans la suite de ce rapport et pour en simplifier l'écriture, nous appelons "objet Cassini" un objet de la base de données construite par l'EHESS par vectorisation des feuilles de Cassini, et "objet BDTopo" un objet de la BDTopo.
Les données sont en projection Lambert 93.

0.2 Contexte et spécificité de l'appariement des données anciennes

L'appariement de données géographiques est un problème complexe car il s'appuie sur la notion de ressemblance entre objets ; ressemblance de forme, de lieu, de nature, de relation spatiale ; très difficiles à établir [Olteanu, 2008]. Dans notre étude, une difficulté majeure supplémentaire vient se greffer à la problématique de l'appariement, provenant de la complexité des données elles-même, et se traduisant à différents niveaux.

0.2.1 Une différence de temporalité

Les données à apparier proviennent de cartes dessinées durant la seconde moitié du 18è siècle. Elles sont donc antérieures aux données de la BDTopo, utilisées comme données de référence, de près de 250 ans. Cette différence majeure de temporalité se traduit naturellement par des transformations des objets représentés sur les fonds de carte (évolution, disparition, absorption, changement de statut, etc.). En effet, le temps fait son oeuvre, et de nombreux événements façonnent l'espace français, qu'il soit naturel, urbain, rural ou encore administratif (guerre, régime politiques, remembrements, avancées technologiques, exodes ruraux et expansion des villes, etc.). Un exemple marquant concerne les remembrements communaux modifiant épisodiquement le profil des communes de France au cours du temps [Motte et al., 2003]. Il est également fréquent d'observer des divergences sémantiques entre objets

Cassini et BDTopo correspondants : certains villages du 18è siècle (hameau ou écart) sont aujourd'hui enregistrés en tant que "ruines", quelques anciens moulins dans la catégorie des "lieux-dits-habités" actuels, etc.

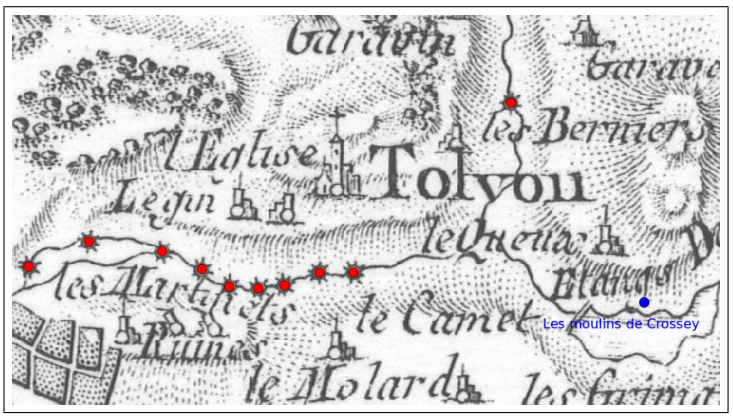


FIGURE 1 – Les anciens moulins (en rouge) ont presque tous disparus. Seul le lieu-dit actuel "les moulins de Crossey" (cercle vert) atteste potentiellement de leur existence et peut leur être apparié.

Enfin, on note parfois de fortes évolutions toponymiques des lieux-dits sur cet intervalle de temps : modifications orthographiques, ajouts,suppressions, substitution de mots ou groupes de mots, etc.

0.2.2 Une différence de niveau de détail et de représentation

Les cartes de Cassini sont à l'échelle du 1/86000ème, et présentent un niveau de granularité très faible comparé à la totalité des données de la BDTopo. Il est donc important de ne sélectionner dans la BDTOpo que les classes ayant un niveau de détail proche de celui des données Cassini. Par exemple, un objet Cassini de la classe "Religieux" et de nature "Église" représente en fait l'église et son lieu-dit, et possède une géométrie ponctuelle. Dans la BDTopo, un lieu dit est représenté par un ponctuel marquant l'emplacement de son toponyme et correspondant au centroid de la zone d'activité (classe PAI ZONE HABITATION), ainsi que par l'ensemble de ses bâtiments, routes, etc. Nous convenons d'établir la correspondance de l'objet Cassini uniquement avec le ponctuel de la classe PAI ZONE HABITATION.

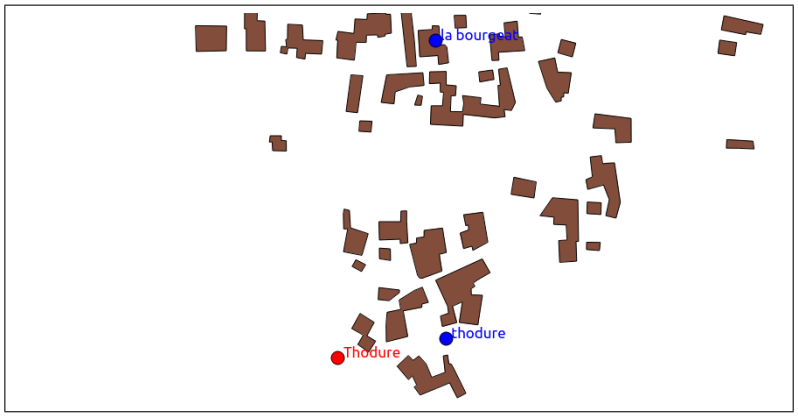


FIGURE 2 – Différents niveaux de détail : l'objet Cassini ponctuel "Église" (en rouge), et les entités correspondantes d'aujourd'hui (en bleu l'objet de la classe PAI ZONE HABITATION, en marron les bâtiments).

0.2.3 Des données imparfaites

L'évaluation du géoréférencement des cartes de Cassini a mis en évidence des décalages de position entre les données, variant de 5 à 750 mètres, conséquence de l'imprécision des fonds de cartes. Cette imprécision peut s'expliquer par des approximations de localisation des objets lors du levé des cartes à l'aide de méthodes de triangulation, mais aussi par l'usage de taille de symboles relativement gros par rapport à l'échelle de la carte.
L'étude des fonds de carte de Cassini souligne également le problème de l'incertitude des données exploitées : quelle confiance peut-on accorder aux informations retranscrites sur les cartes ? Ces incertitudes peuvent par exemple concerner des hameaux adjacents portant le même toponyme, une rivière passant au nord d'un village alors qu'elle le contourne par le sud aujourd'hui, des positionnements relatifs de lieux-dits les uns par rapport aux autres non cohérents avec leur emplacement actuel, comme illustré figure 3, etc.

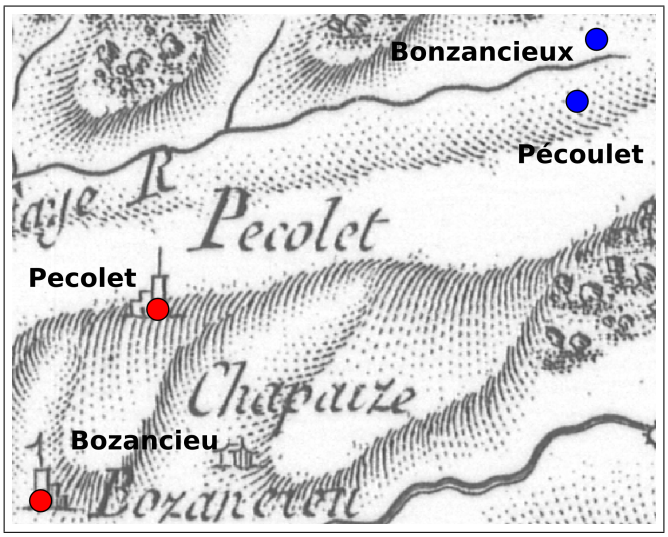


FIGURE 3 – Modification de la position relative entre les villages de Bozancieux et de Pécoulet : Bozancieux est au nord de Pécoulet dans la BDTOPO (en bleu) mais au sud sur la feuille Cassini de Grenoble (en rouge).

Enfin, en plus des problèmes d'imprécision et d'incertitude, se pose également celui de la complétude des données, notamment de la toponymie. Il existe effectivement de nombreux objets sur les fonds de cartes, qui ne possèdent aucun nom (hameaux, écarts, églises, moulins, etc.).

0.2.4 Conclusion

Les données à apparier sont donc très complexes par nature. De fait, il apparaît que les techniques d'appariement privilégiant les relations géométriques entre objets ne sont pas nécessairement adaptées.

0.3 Appariement des géométries ponctuelles

Nous nous intéressons dans cette section à l'appariement des objets Cassini des classes "religieux" (de type église, chapelle, prieuré, etc.), "non religieux" (de type hameau, écart, château, etc.) et "Moulin et activités industrielles" (de type moulin à eau, mines, etc.), qui ont tous une représentation géométrique de type ponctuel.

0.3.1 Travaux relatifs à l'appariement d'objets ponctuels

La plupart des processus d'appariement [Walter and Fritsch, 1999 ; Sui et al., 2004 ; Blasby et al., 2004 ; Zhang et al., 2005 ; Haunert, 2005 ; Voltz, 2006 ; Lüscher et al., 2007] ont une approche fondée uniquement sur un critère spatial. Après recalage et simplification des données, les objets à apparier sont sélectionnés de façon aléatoire, puis la définition d'une zone tampon (un *buffer*) autour de ces objets permet de filtrer les possibles candidats à l'appariement. Ce sont des mesures de distance, la distance euclidienne ou d'autres distances géométriques ([Devogele, 1997 ; Beeri et al., 2004 ; Voltz, 2006 ; Mustière and Devogele, 2008]), qui interviennent ensuite pour décider de l'appariement effectif des candidats aux objets. Dans ces approches, le processus d'appariement est évalué par des mesures de confiance [Clodoveu and Fonseca, 2007 ; Mustière and Devogele, 2008] ou en terme de précision et de rappel [Beeri et al., 2004 ; Safra et al., 2006]. La *précision* des liens d'appariement (respectivement des objets non appariés) est le rapport entre le nombre de vrais positifs (respectivement de vrais négatifs) et la somme des vrais positifs et des faux positifs (respectivement des vrais négatifs et des faux négatifs). Le *rappel* des liens d'appariement (respectivement des objets non appariés) est le rapport entre le nombre de vrais positifs (respectivement de vrais négatifs) et le nombre de liens attendus, c'est-à-dire établis manuellement. L'objectif d'évaluation recherché est donc de déterminer le nombre de vrais positifs (liens d'appariement correctement établis), de vrais négatifs (objets correctement non appariés), de faux positifs (liens d'appariement établis par erreur) et de faux négatifs (objets non appariés par erreur).
Enfin, il existe peu de méthodes permettant un appariement d'objets à géométrie ponctuelle en dehors de celles proposées par [Beeri et al., 2004] et [Olteanu, 2008]. L'approche basée sur la théorie de l'évidence que propose [Olteanu, 2008] a l'avantage de tenir compte de l'imperfection des données et de combiner trois critères : géométrique, sémantique et toponymique. De façon classique, *le critère géométrique* évalue la proximité de deux objets avec la distance euclidienne. *Le critère sémantique*, qui mesure la similarité de nature entre les objets, utilise la distance de Wu-Palmer [Wu and Palmer, 1994] et s'appuie sur la taxonomie générale des objets géographiques de la BDTOPO proposée par [Abadie and Mustière, 2008]. *Le critère toponymique* analyse la ressemblance lexicale entre deux toponymes, via l'usage de la distance de Levenshtein [Levenshtein, 1965].
L'approche proposée par [Olteanu, 2008] repose sur cinq étapes (cf. figure 4.a) : sélection des candidats, initialisation des masses de croyance, fusion des critères, fusion des candidats et prise de décision. Un des problèmes majeur posé par cette approche réside dans l'étape d'initialisation des masses de croyance car elle nécessite un paramétrage délicat. En effet, la théorie des fonctions de croyance utilise les masses de croyance comme un filtre qui s'applique de façon particulière à chaque critère, filtre qui peut atténuer plus ou moins l'influence du critère et qui exprime le degré de croyance en l'appariement entre l'objet étudié et son candidat pour le critère mesuré. Ce filtre est généralement représenté par une fonction floue, qui doit préalablement être définie. [Olteanu, 2008] utilise à cette fin des fonctions linéaires mais précise qu'il serait intéressant d'étudier d'autres modélisations de fonctions. Par ailleurs, l'algorithme est très coûteux en temps de calcul.

0.3.2 Propositions d'améliorations pour l'appariement multi-critères basé sur la théorie des fonctions de croyance

Dans notre contexte d'intégration de données anciennes imparfaites, la méthode d'appariement multi-critères basée sur la théorie des croyances proposée par [Olteanu, 2008] apparaît comme étant la plus adaptée. Cependant sa complexité algorithmique est un inconvénient majeur. Nous proposons une approche simplifiée qui en reprend les principales étapes tout en s'affranchissant de la théorie de l'évidence. Notre algorithme apparie ainsi 35 fois plus vite à configuration matérielle équivalente (estimation réalisée à partir de 350 objets). Nous utilisons par la suite le terme de "poids" plutôt que celui de "masse de croyance" et l'étape de pondération des critères remplace celle d'initialisation des masses de croyance. Nous proposons également de nouvelles améliorations à cette approche – représentées en vert figure 4 – destinées à faciliter son paramétrage et l'analyse des résultats. Afin

de vérifier les résultats, un appariement manuel est au préalable effectué. Cent points sont extraits aléatoirement sur trois zones pour chacune des classes étudiées, puis appariés manuellement, soit au total 900 objets Cassini, représentant un peu plus du dixième des données totales.

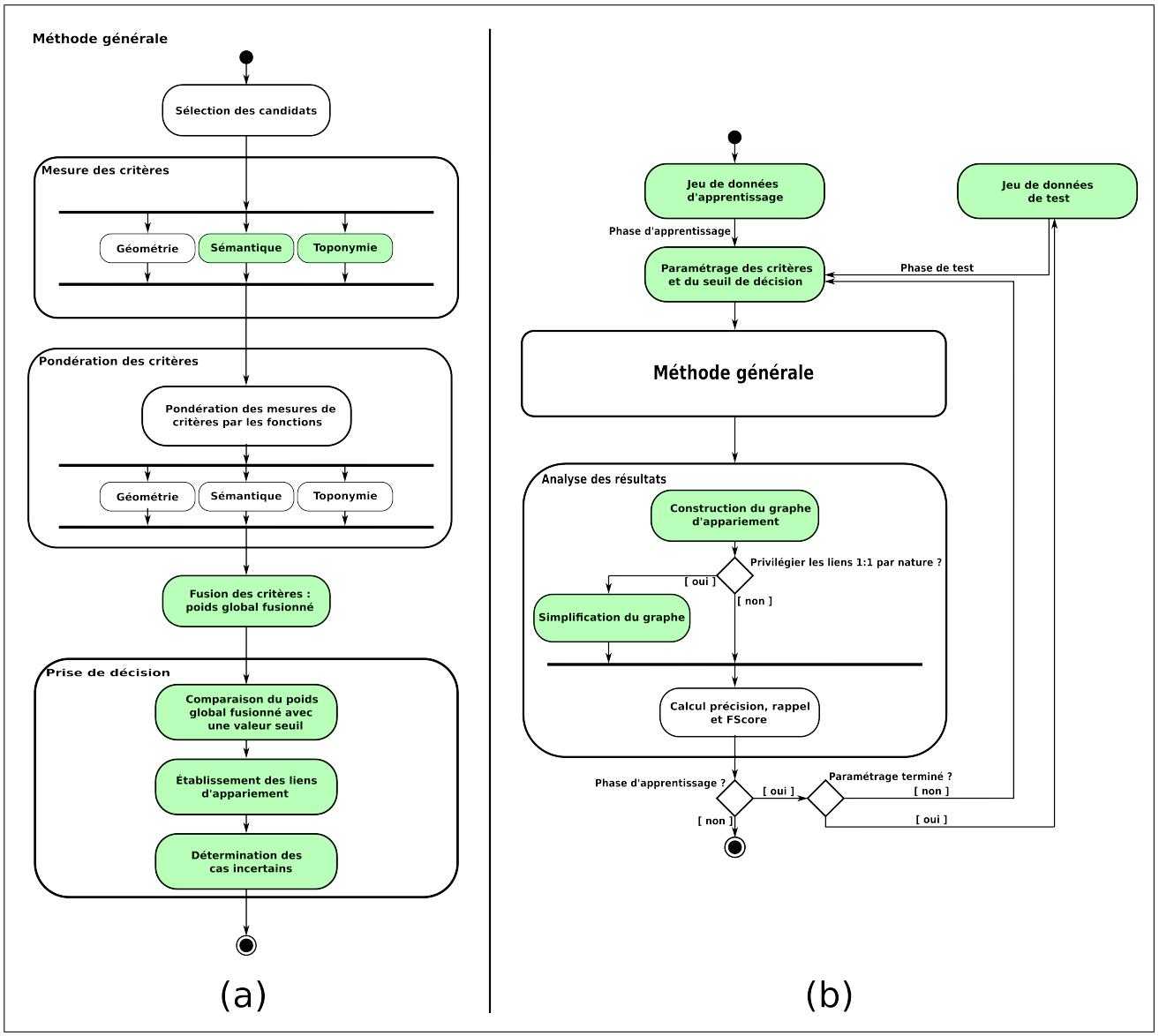


FIGURE 4 – Méthode s'appuyant sur l'approche proposée par [Olteanu, 2008] (a) et méthode de paramétrisation automatique (b).

Sélection des candidats et évaluation des critères

0.3.2.0.1 Sélection des candidats Autour de chaque objet Cassini, une zone tampon, ou *buffer*, est définie afin de déterminer quels sont les objets de la BDTOPO les plus probables de lui être appariés. La taille de ce *buffer* dépend de la zone étudiée, en raison des différences de précision de chaque feuille Cassini, soulignées par les résultats du géoréférencement. Ces seuils ont été fixés empiriquement pour chaque zone en fonction de l'erreur maximale déterminée lors de l'appariement manuel, avec une marge d'erreur de 10% : 1400m pour Reims, 1800m pour Saint-Malo et 2600m pour Grenoble. En cas d'abscence d'appariement manuel en amont, les résultats de l'étude du géoréférencement peuvent être utilisés afin de fixer comme valeur de seuil pour une zone donnée un multiple de l'erreur quadratique moyenne issue du géoréférencement.

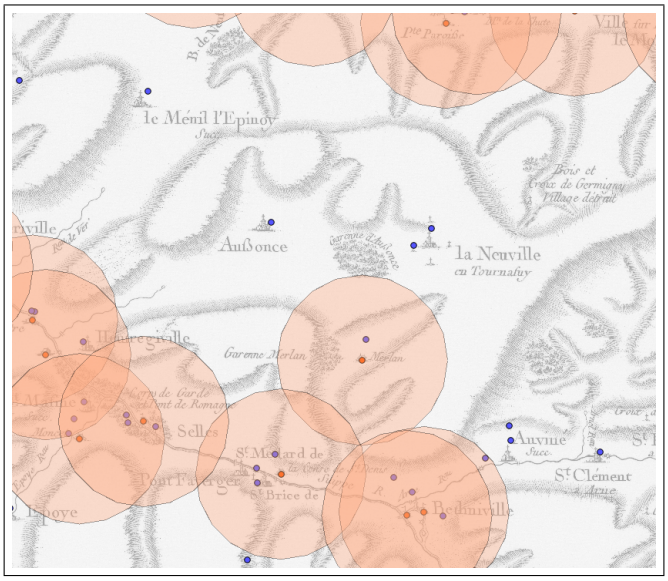


FIGURE 5 – Étape de sélection des candidats, sur un critère d'éloignement.

0.3.2.0.2 Modification de la distance sémantique De nombreux concepts saisis dans la base de données des objets Cassini n'existent soit pas du tout ("prieuré", "gentilhommière", etc.), soit de manière incomplète dans la taxonomie de [Abadie and Mustière, 2008] (pas de "moulin à vent en bois" mais seulement des "moulins à vent"). En l'absence d'une ontologie de domaine spatio-temporelle, couvrant les époques ici traitées, nous modifions le calcul de la distance sémantique utilisée par [Olteanu, 2008] de la manière suivante. D'une part, après lemmatisation des noms de nature d'objets, nous vérifions si chacun des lemmes ne sont pas présents dans la taxonomie, avant de calculer une distance de Wu-Palmer minorée (en fonction de la proportion de lemmes absents de la taxonomie) ; ceci permet ainsi d'associer tous les moulins (à vent, à eau, à vent en bois, à vent en pierre) aux objets de nature "moulin" de la BDTOPO. **1.** Chaque nature (de l'objet étudié et du candidat) est lemmatisée : mise en forme (suppression des accents, des majuscules, etc.), simplification (suppression de certains articles, pronoms etc.), séparation en une liste de mots ; **2.** pour chaque liste, nous concaténons tous ses éléments pour former la plus grande phrase. Si cette phrase existe dans la taxonomie nous la sélectionnons, sinon, nous la raccourcissons en supprimant le dernier concept et nous réitérons l'étape précédente ; **3.** si une phrase a été retenue pour les deux listes, nous calculons la distance de Wu et Palmer associée, en lui appliquant une minoration, fonction de la proportion de lemmes absents de la taxonomie

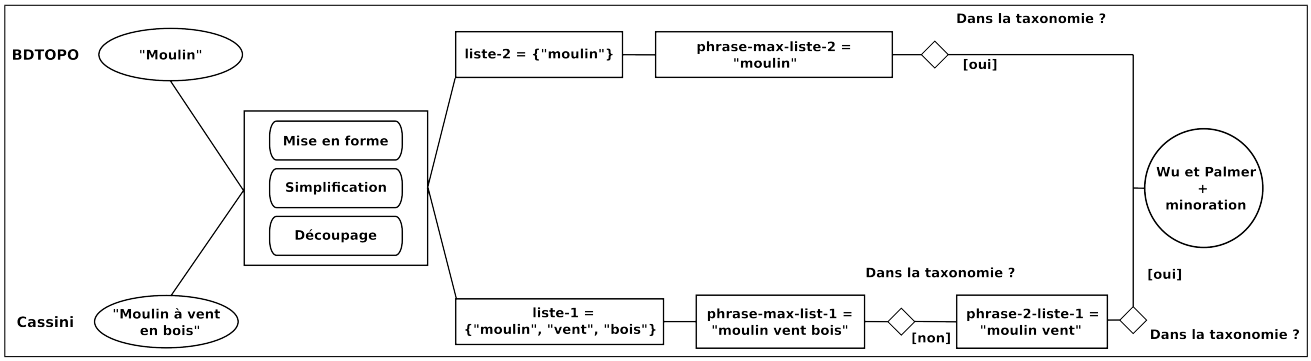


FIGURE 6 – Exemple de calcul de la distance sémantique par lemmatisation entre les concepts "Moulin" et "Moulin à vent en bois".

D'autre part, les toponymes des objets eux-mêmes sont analysés pour vérifier s'ils ne contiennent pas eux-mêmes une nature d'objet présente dans la taxonomie. Des liens d'appariement sont ainsi établis

entre un "moulin à eau" de Cassini et un lieu-dit de la BDTOPO, de nature "lieu-dit habité" et de toponyme "le moulin du clos", "moulin" étant un concept de la taxonomie.

0.3.2.0.3 Modification de la distance toponymique La toponymie joue un rôle primordial dans l'appariement des données anciennes mais de grandes différences orthographiques existent entre les toponymes actuels et ceux de la carte de Cassini. Comme leur comparaison exige de la souplesse dans la mesure du critère toponymique, la distance de Damerau-Levenshtein, extension plus tolérante de la distance de Levenshtein, est utilisée. De plus nous associons à la distance toponymique une valeur faible fixée à environ la moitié de la valeur maximale prise par le critère toponymique lorsque certains cas particuliers sont en jeu : (i) concaténation de termes, *e.g.* "Pont Aven" devenant "Pontaven" ; (ii) interversion de termes, *e.g.* "La Chapelle Felcourt" devenant "Felcourt la Chapelle" ; (iii) ajout ou suppression de termes, *e.g.* "Saint-Souplet" devenant "Saint-Souplet sur Py" ou "Hauteville" devenant "Hauteville-sur-Mer".

0.3.2.1 Modification de l'étape de prise de décision

Après que les critères aient été combinés, l'algorithme de [Olteanu, 2008] fusionne les candidats avant de choisir celui qui maximise une fonction de probabilité appelée probabilité pignistique. Nous préférons adopter un modèle de décision plus simple, en comparant la moyenne des poids des critères, appelée poids global fusionné, à un seuil défini préalablement, moins coûteux en terme de calculs. Nous qualifions les liens d'appariements d'incertains lorsque :

- la valeur du poids global fusionné est très proche (inférieure) à la valeur du seuil global
- la nature de l'objet de référence est de type "naturelle" (bois, forêt, eau, etc.), alors que la nature du candidat est de type "artificielle" (moulin, hameau, église, etc.), et réciproquement.
- le pourcentage de candidats appariés de même nature que le candidat étudié, pour tous les objets de référence appartenant à la même classe ,est faible.
- il existe déjà un lien d'appariement entre l'objet de référence et une entité de la même classe que le candidat, dont la valeur du poids global fusionné est plus faible que pour le lien étudié (c'est typiquement le cas lorsqu'une église Cassini est appariée avec deux objets de la classe "bâti remarquable : une église et une chapelle).

On constate en moyenne entre 10 et 30 cas incertains pour 400 ou 500 objets.

0.3.2.2 Vers une paramétrisation adaptée de l'appariement

Comme les mesures de précision et de rappel sont aussi importantes l'une que l'autre, le protocole de paramétrisation automatique développé se base sur l'utilisation d'une fonction qui les combine équitablement, appelée *F-score* et définie comme suit :

$$F\text{-}score = 2 * \frac{precision * rappel}{precision + rappel} \tag{1}$$

0.3.2.2.1 Modélisation des fonctions L'influence du choix des fonctions de pondération, notées f_i, sur les résultats de l'appariement est maintenant étudié. Pour chaque critère, quatre mêmes familles de fonctions sont testées. Soit min et max les valeurs minimale et maximale de la fonction étudiée, et soit S le seuil pour lequel la fonction prend sa valeur maximale, les fonctions de pondération, notées respectivement f_1, f_2, f_3 et f_4, vérifient $\forall i \in [\![1;4]\!], f_i(0) = min$ et $\forall i \in [\![1;4]\!], \forall x \geq S, f_i(x) = max$, et ont pour équations :

$$f_1(x) = min + (max - min)\frac{1 - e^{\frac{\lambda x}{S}}}{1 - e^{\lambda}} \text{ si x} < S, \forall \lambda \in \mathbb{R} \quad (2)$$

$$f_2(x) = min + (max - min)\frac{ln(\frac{\lambda x}{S} + 1)}{ln(\lambda + 1)} \text{ si x} < S, \forall \lambda \in \mathbb{R} \quad (3)$$

$$f_3(x) = \frac{max - min}{S}x + min \text{ si x} < S \quad (4)$$

$$f_4(x) = min + (max - min)\sqrt{\frac{x}{S}} \text{ si x} < S \quad (5)$$

La figure 7 illustre les courbes représentatives de ces fonctions de pondération :

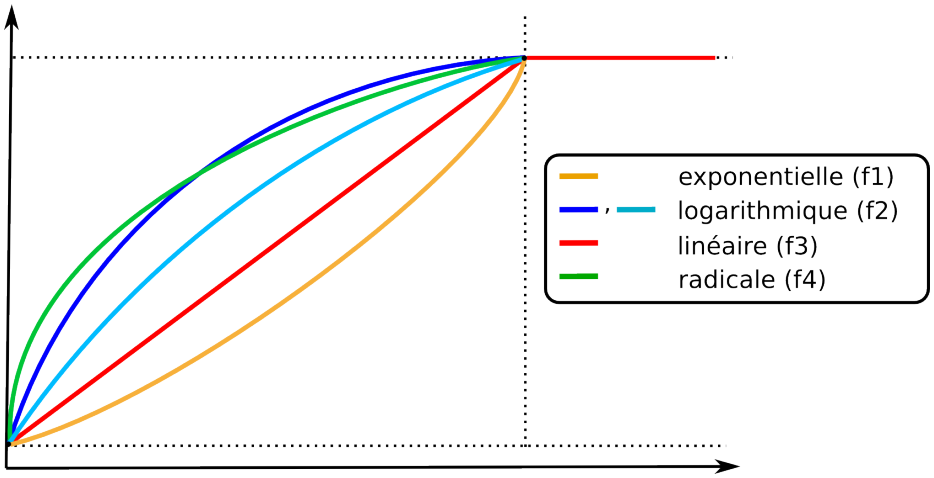


FIGURE 7 – Courbes représentatives de quelques fonctions de pondération.

Le tableau 1 présente pour différentes combinaisons de fonctions et pour une série de paramètres fixés arbitrairement, les résultats du processus d'appariement sur les données appariées manuellement de la classe "non religieux" de la feuille de Reims.

Courbes critères [géométrique, sémantique, toponymique]	**Précision LAP**	**Rappel LAP**	**Précision NAP**	**Rappel NAP**	**F-score moyen**
[EXP(λ=1), LOG(λ=8), LINEAIRE]	97%	97%	96%	100%	**97%**
[EXP(λ=1), LOG(λ=8), EXP(λ=1)]	95%	91%	91%	100%	**94%**
[EXP(λ=1), EXP(λ=8), EXP(λ=1)]	87%	93%	94%	96%	**92%**
[LINEAIRE, LINEAIRE, LINEAIRE]	70%	97%	97%	76%	**83%**
[RADICALE, RADICALE, RADICALE]	100%	72%	78%	100%	**86%**
[LOG(λ=8), LOG(λ=8), LOG(λ=8)]	100%	72%	78%	100%	**86%**

TABLE 1 – Résultats d'appariement (classe "non religieux", feuille de Reims) pour différentes combinaisons de fonctions : précisions et rappel des liens d'appariement (LAP), précision et rappel des objets non appariés (NAP), *F-score* moyen. Paramètres utilisés pour les obtenir : geoMin=0.2, geoMax=0.5, geoS=1000, semMin=0.0, semMax=0.6, semS=0.3, topMin=0.1, topMax=0.7, topS=0.9, seuil=0.41.

Nous constatons que la qualité de la procédure d'appariement dépend du choix des fonctions de pondération. De plus, dans l'exemple ci-contre, les meilleurs résultats sont obtenus par l'utilisation d'une courbe différente par critère : une exponentielle pour la géométrie, une logarithmique pour la sémantique et une linéaire pour la toponymie. Pour un critère donné, le type de courbe déterminé par la paramétrisation dépend peu de la zone considérée, mais surtout de la classe étudiée. Les paramètres du processus d'appariement à définir sont donc le seuil global de prise de décision et, pour chaque

critère, le type et les valeurs min, max et S de sa fonction de pondération. Pour trois critères étudiés, il y a donc 13 paramètres à déterminer $(1 + 4 * 3)$.

0.3.2.2.2 Paramétrisation automatique des fonctions et détermination du seuil Dans la logique d'évaluation d'un appariement de données géographiques, il convient de s'intéresser autant aux liens d'appariement qu'aux objets non appariés, notamment dans un contexte d'appariement de données anciennes. Le processus de paramétrisation automatique repose sur un apprentissage supervisé. La base d'exemples des appariements manuels est scindée en deux : une base d'apprentissage et une base de tests (respectivement 2/3 et 1/3 des données). Le processus est itératif et prend en entrée un jeu de paramètres à tester. La moyenne des *F-scores* pour les liens d'appariement et pour les objets non appariés, que nous appellerons par la suite *F-score* moyen, est recalculée pour chaque nouveau paramètre. Les paramètres d'appariement recherchés sont ceux qui maximisent le *F-score* moyen dans la phase d'apprentissage. La base de tests sert à valider la paramétrisation issue de l'apprentissage.

0.3.2.2.3 Optimisations Le processus de paramétrisation repose sur l'imbrication d'autant de boucles qu'il y a de paramètres à déterminer. Sachant que notre algorithme d'appariement s'exécute en 2 secondes en moyenne, le calcul pour 13 paramètres prenant chacun 3 valeurs durerait : $2 * 3^{13}$ secondes soit plus de 36 jours. Nous avons donc du y apporter certaines optimisations : optimisations algorithmiques, de code, placement en amont et stockage des opérations indépendantes des paramètres, et minimisation du volume de données à utiliser. Nous revenons ici sur ce dernier point qui constitue l'optimisation la plus sensible en temps de calcul et mémoire vive utilisée.
Nous appelons pour ce paragraphe, objet Cassini aléatoire un des objet Cassini apparié manuellement. Pour une classe données et sur une zone, l'utilisation de l'ensemble des objets Cassini n'est pas nécessaire. Cependant, n'utiliser uniquement que les données appariées manuellement ne suffit pas car il est nécessaire de disposer du contexte autour de ces données. Dans l'exemple de la figure 8, on se

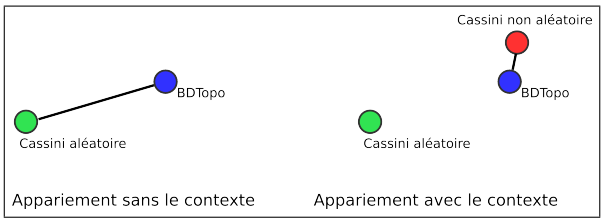


FIGURE 8 – Erreur d'appariement dû à un manque de contexte autour des données.

place dans le cas ou seules les données appariées manuellement sont utilisées. Un objet aléatoire a été apparié à un objet de la BDTopo. Or, si l'objet Cassini en rouge avait également était pris en compte, l'appariement aurait peut-être été différent et l'objet Cassini aléatoire en vert aurait pu soit ne pas être apparié, soit être apparié avec un autre objet de la BDTopo. Nous voyons bien ici la nécessité de considérer à la fois les objets Cassini appariés manuellement, mais également les autres objets Cassini qui en sont proche.
Pour l'évaluation de notre algorithme, nous utilisons donc pour l'appariement :
- les objets Cassini appariés manuellement
- les objets Cassini présents dans un buffer autour des objets Cassini aléatoires, de rayon supérieur au double du rayon de recherche des candidats (un rayon égal au rayon de recherche ne suffirait pas, car la situation de la figure 8 peut encore être envisagée si un objet BDTopo se trouve proche de la bordure du buffer, et qu'un objet Cassini se trouve juste au delà de ce buffer)
- les objets de la BDTopo présents dans ce même buffer.

Grâce à ces optimisation, l'algorithme d'appariement s'exécute en 15ms en moyenne, soit environ 5h pour une paramétrisation complète pour 13 paramètres avec 3 valeurs testées pour chacun.

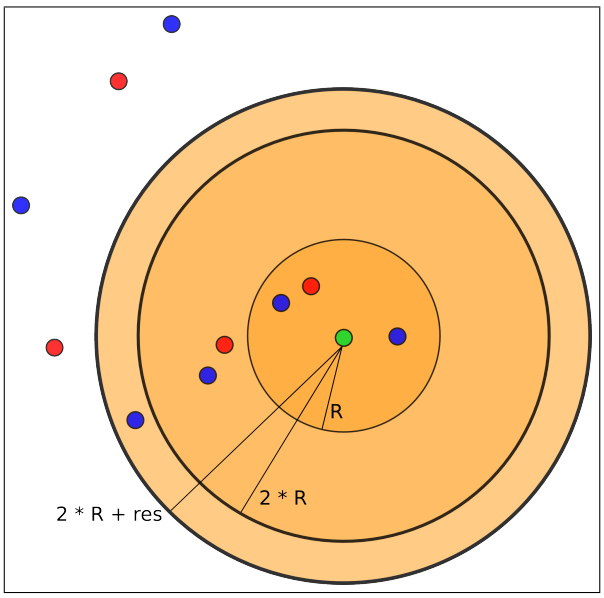


FIGURE 9 – Sélection des données pour l'appariement.

0.3.2.2.4 Méthode d'analyse des résultats La notion de cardinalité des liens d'appariement est délicate à traiter. Il existe des liens *1:1* – un objet Cassini apparié avec un objet de la BDTOPO –, *1:n* – un objet Cassini apparié avec plusieurs objets de la BDTOPO – et *n:m* – plusieurs objets Cassini appariés avec plusieurs objets de la BDTOPO –. Nous appelons lien *1:1* (ou *1:n* ou *n:m*) par

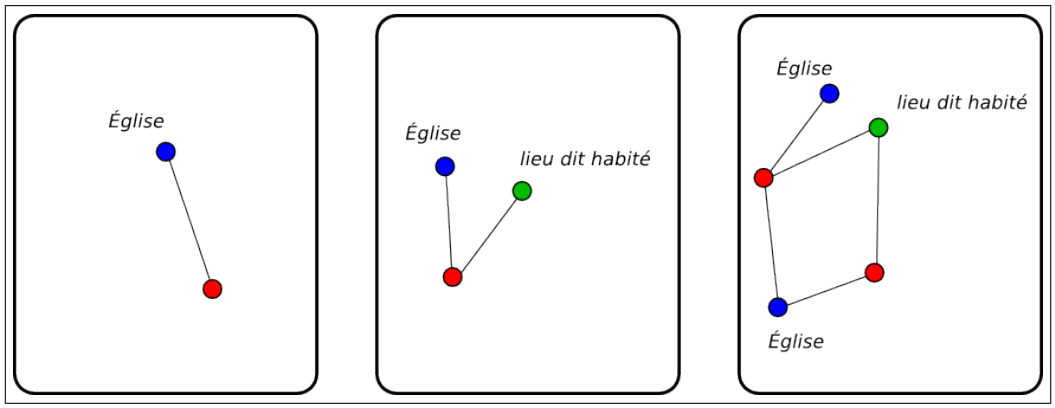


FIGURE 10 – Cardinalité des liens d'appariement.

nature un lien *1:1* (ou *1:n*, *n:m*) entre objets de même nature. Nous proposons une méthode d'étude locale des liens d'appariement par nature d'un objet en construisant le graphe d'appariement de l'objet relatif à une nature N donnée de candidat de proche en proche, tel qu'illustré figure 11.

Ce *graphe d'appariement* correspond au graphe formé par l'ensemble des liens d'appariement (les arcs), des objets et des candidats de nature N (les nœuds) tel qu'il existe un chemin entre l'objet initial et les autres objets ou candidats. Ce graphe permet d'étudier le contexte local d'un objet étudié en visualisant les objets et candidats en conflit.

Privilégier les liens *1:1* par nature peut s'avérer nécessaire. Il est par exemple courant qu'une église Cassini soit appariée avec plusieurs lieux-dits différents au terme de l'étape de prise de décision. Nous proposons de simplifier ces graphes d'appariement conflictuels en choisissant comme meilleur candidat d'un objet Cassini de référence, le plus proche voisin du ponctuel de la BDTOPO portant un toponyme proche de celui de l'objet étudié (cf. figure 12.c). Dans les cas où la toponymie ne peut être utilisée, les arêtes de chaque graphe sont pondérées par la valeur du poids global fusionné associée à ce lien. Nous ne conservons alors que les n_{min} liens dont la somme des pondérations est minimale, où $n_{min} =$

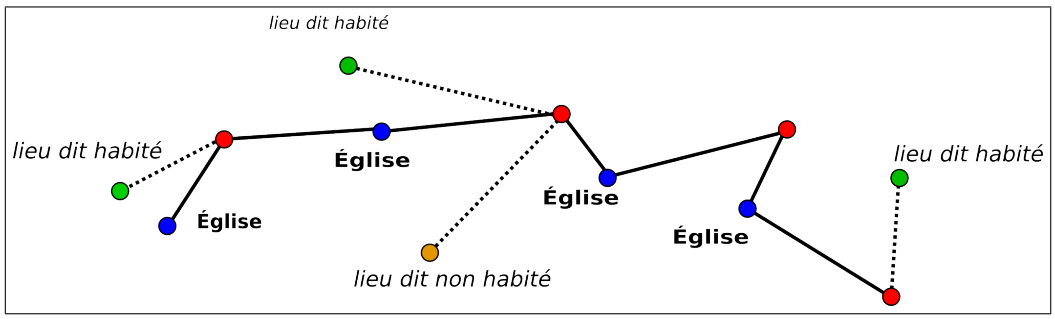
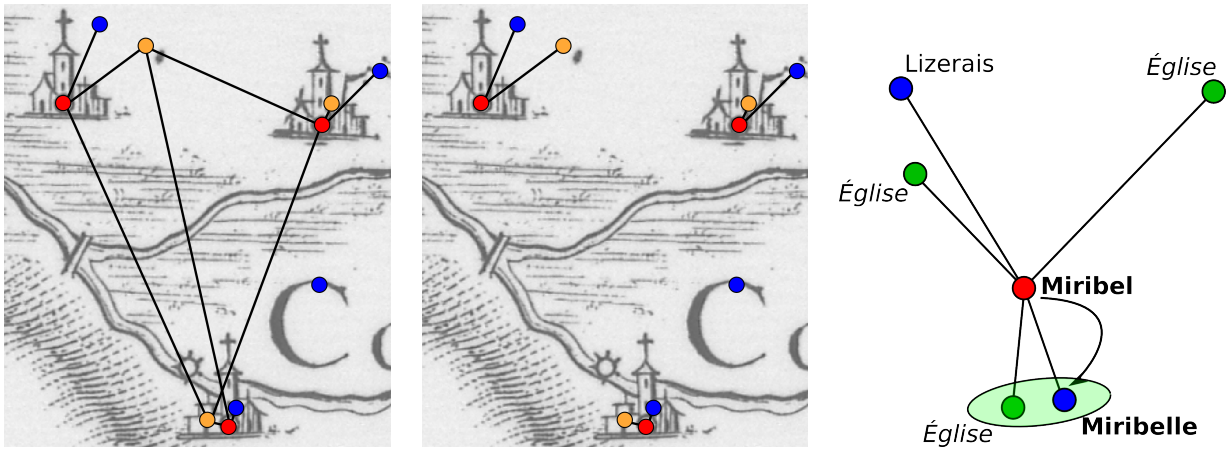


FIGURE 11 – Exemple de graphe d'appariement (en trait plein) pour une nature "Église". Les objets Cassini sont en rouge.

$min(|Objets|, |Candidats|)$ avec $|E|$ représentant le cardinal de l'ensemble E.



(a) (b) (c)

FIGURE 12 – Graphe d'appariement pour la nature "Église" avant (a) et après (b) simplification. Les églises Cassini sont en rouge, les églises de la BDTOPO en orange, les lieux-dits de la BDTOPO en bleu. La figure (c) montre la prédilection des candidats proches du meilleur candidat au sens de la distance toponymique.

0.3.3 Résultats et discussions

0.3.3.1 Résultats détaillés

Nous présentons ici les résultats obtenus pour les bases d'apprentissage et de validation, par le processus de paramétrisation automatique sur chaque classe de chaque zone, ainsi que les paramètres utilisés et le nombre d'objets appariés et non appariés. Les résultats sont globalement bons, le *F-score* moyen variant entre 91% et 100% sur les données de validation. Sur les trois zones, nous ne constatons pas de fort différentiel de précision et de rappel entre classes, malgré des proportions d'objets aujourd'hui disparus très variées : faible pour la classe "religieux" (entre 10% et 30%), moyenne pour la classe non "religieux" (entre 35% et 55%) et importante pour la classe "moulins" (entre 65% et 80%).

0.3.3.1.1 Reims

0.3.3.1.1.1 Religieux

	Apprentissage			Validation		
	Précision	**Rappel**	**F-Score moyen**	**Précision**	**Rappel**	**F-Score moyen**
Liens d'appariement	100%	97%	98%	100%	97%	99%
Objets non appariés	94%	100%		100%	100%	

	Apprentissage	**Validation**
Nombre d'objets appariés	54	28
Nombre d'objets non appariés	13	5

Paramètres	**Valeurs**
Seuil global	0.42
Fonction géométrie	Exponentielle, $\lambda = 4$
Valeur min. géométrie	0.5
Valeur max. géométrie	0.6
Seuil géométrie	1150
Fonction sémantique	Exponentielle, $\lambda = 1$
Valeur min. sémantique	0.3
Valeur max. sémantique	0.5
Seuil sémantique	0.6
Fonction toponymie	Linéaire
Valeur min. toponymie	0.0
Valeur max. toponymie	0.8
Seuil toponymie	0.8

0.3.3.1.1.2 Non religieux

	Apprentissage			Validation		
	Précision	**Rappel**	**F-Score moyen**	**Précision**	**Rappel**	**F-Score moyen**
Liens d'appariement	100%	95%	97%	91%	100%	98%
Objets non appariés	95%	100%		100%	100%	

	Apprentissage	**Validation**
Nombre d'objets appariés	36	20
Nombre d'objets non appariés	31	13

Paramètres	Valeurs
Seuil global	0.41
Fonction géométrie	Linéaire
Valeur min. géométrie	0.2
Valeur max. géométrie	0.5
Seuil géométrie	1000
Fonction sémantique	Logarithmique, $\lambda = 8$
Valeur min. sémantique	0.0
Valeur max. sémantique	0.6
Seuil sémantique	0.3
Fonction toponymie	Exponentielle, $\lambda = 1$
Valeur min. toponymie	0.1
Valeur max. toponymie	0.7
Seuil toponymie	0.9

0.3.3.1.1.3 Moulins

	Apprentissage			Validation		
	Précision	**Rappel**	**F-Score moyen**	**Précision**	**Rappel**	**F-Score moyen**
Liens d'appariement	83%	83%	91%	83%	100%	95%
Objets non appariés	98%	98%		100%	96%	

	Apprentissage	Validation
Nombre d'objets appariés	27	8
Nombre d'objets non appariés	40	25

Paramètres	Valeurs
Seuil global	0.41
Fonction géométrie	Linéaire
Valeur min. géométrie	0.3
Valeur max. géométrie	0.5
Seuil géométrie	1200
Fonction sémantique	Exponentielle , $\lambda = 1$
Valeur min. sémantique	0.2
Valeur max. sémantique	0.6
Seuil sémantique	0.5
Fonction toponymie	Linéaire
Valeur min. toponymie	0.1
Valeur max. toponymie	0.7
Seuil toponymie	0.9

0.3.3.1.2 Saint-Malo

0.3.3.1.2.1 Religieux

	Apprentissage			Validation		
	Précision	**Rappel**	**F-Score moyen**	**Précision**	**Rappel**	**F-Score moyen**
Liens d'appariement	99%	97%	92%	100%	98%	99%
Objets non appariés	75%	100%		100%	99%	

	Apprentissage	**Validation**
Nombre d'objets appariés	59	31
Nombre d'objets non appariés	6	2

Paramètres	**Valeurs**
Seuil global	0.43
Fonction géométrie	Exponentielle, $\lambda = 4$
Valeur min. géométrie	0.4
Valeur max. géométrie	0.7
Seuil géométrie	1750
Fonction sémantique	Logarithmique, $\lambda = 8$
Valeur min. sémantique	0.2
Valeur max. sémantique	0.5
Seuil sémantique	0.6
Fonction toponymie	Exponentielle, $\lambda = 1$
Valeur min. toponymie	0.0
Valeur max. toponymie	0.8
Seuil toponymie	0.8

0.3.3.1.2.2 Non religieux

	Apprentissage			Validation		
	Précision	**Rappel**	**F-Score moyen**	**Précision**	**Rappel**	**F-Score moyen**
Liens d'appariement	100%	92%	93%	100%	88%	90%
Objets non appariés	83%	100%		75%	100%	

	Apprentissage	**Validation**
Nombre d'objets appariés	48	24
Nombre d'objets non appariés	19	9

Paramètres	Valeurs
Seuil global	0.37
Fonction géométrie	Exponentielle, $\lambda = 1$
Valeur min. géométrie	0.2
Valeur max. géométrie	0.5
Seuil géométrie	1550
Fonction sémantique	Logarithmique, $\lambda = 8$
Valeur min. sémantique	0.0
Valeur max. sémantique	0.6
Seuil sémantique	0.3
Fonction toponymie	Exponentielle, $\lambda = 1$
Valeur min. toponymie	0.0
Valeur max. toponymie	0.7
Seuil toponymie	0.7

0.3.3.1.2.3 Moulins

	Apprentissage			Validation		
	Précision	**Rappel**	**F-Score moyen**	**Précision**	**Rappel**	**F-Score moyen**
Liens d'appariement	100%	100%	100%	100%	100%	100%
Objets non appariés	100%	100%		100%	100%	

	Apprentissage	Validation
Nombre d'objets appariés	13	13
Nombre d'objets non appariés	54	20

Paramètres	Valeurs
Seuil global	0.38
Fonction géométrie	Linéaire
Valeur min. géométrie	0.3
Valeur max. géométrie	0.5
Seuil géométrie	1900
Fonction sémantique	Exponentielle, $\lambda = 1$
Valeur min. sémantique	0.2
Valeur max. sémantique	0.6
Seuil sémantique	0.5
Fonction toponymie	Exponentielle, $\lambda = 1$
Valeur min. toponymie	0.2
Valeur max. toponymie	0.7
Seuil toponymie	0.6

0.3.3.1.3 Grenoble

0.3.3.1.3.1 Religieux

	Apprentissage			Validation		
	Précision	**Rappel**	**F-Score moyen**	**Précision**	**Rappel**	**F-Score moyen**
Liens d'appariement	99%	95%	94%	97%	99%	90%
Objets non appariés	82%	100%		100%	67%	

	Apprentissage	**Validation**
Nombre d'objets appariés	55	10
Nombre d'objets non appariés	31	2

Paramètres	**Valeurs**
Seuil global	0.37
Fonction géométrie	Linéaire
Valeur min. géométrie	0.2
Valeur max. géométrie	0.5
Seuil géométrie	1850
Fonction sémantique	Logarithmique , $\lambda = 8$
Valeur min. sémantique	0.2
Valeur max. sémantique	0.5
Seuil sémantique	0.7
Fonction toponymie	Exponentielle , $\lambda = 3$
Valeur min. toponymie	0.0
Valeur max. toponymie	0.8
Seuil toponymie	0.7

0.3.3.1.3.2 Non religieux

	Apprentissage			Validation		
	Précision	**Rappel**	**F-Score moyen**	**Précision**	**Rappel**	**F-Score moyen**
Liens d'appariement	98%	100%	97%	100%	100%	100%
Objets non appariés	93%	96%		100%	100%	

	Apprentissage	**Validation**
Nombre d'objets appariés	42	24
Nombre d'objets non appariés	23	9

Paramètres	Valeurs
Seuil global	0.39
Fonction géométrie	Exponentielle, $\lambda = 1$
Valeur min. géométrie	0.2
Valeur max. géométrie	0.5
Seuil géométrie	1700
Fonction sémantique	Logarithmique, $\lambda = 8$
Valeur min. sémantique	0.2
Valeur max. sémantique	0.5
Seuil sémantique	0.2
Fonction toponymie	Exponentielle, $\lambda = 1$
Valeur min. toponymie	0.1
Valeur max. toponymie	0.7
Seuil toponymie	0.9

0.3.3.1.3.3 Moulins

	Apprentissage			Validation		
	Précision	**Rappel**	**F-Score moyen**	**Précision**	**Rappel**	**F-Score moyen**
Liens d'appariement	100%	93%	98%	100%	100%	100%
Objets non appariés	98%	100%		100%	100%	

	Apprentissage	Validation
Nombre d'objets appariés	13	y 5
Nombre d'objets non appariés	50	27

Paramètres	Valeurs
Seuil global	0.41
Fonction géométrie	Logarithmique , $\lambda = 8$
Valeur min. géométrie	0.2
Valeur max. géométrie	0.5
Seuil géométrie	1800
Fonction sémantique	Linéaire
Valeur min. sémantique	0.1
Valeur max. sémantique	0.5
Seuil sémantique	0.5
Fonction toponymie	Exponentielle , $\lambda = 1$
Valeur min. toponymie	0.2
Valeur max. toponymie	0.8
Seuil toponymie	0.8

0.3.3.2 Résumé des résultats

Le tableau 2 montre les résultats obtenus pour chaque classe de chaque zone étudiée.

Zone	Classe	Apprentissage					Validation				
		Préc. LAP	**Rap. LAP**	**Préc. NAP**	**Rap. NAP**	**F-score mean**	**Préc. LAP**	**Rap. LAP**	**Préc. NAP**	**Rap. NAP**	**F-score mean**
Reims	Religieux	100%	97%	94%	100%	**98%**	100%	97%	100%	100%	**99%**
	Non religieux	100%	95%	95%	100%	**97%**	91%	100%	100%	100%	**98%**
	Moulins	83%	83%	98%	98%	**91%**	83%	100%	100%	96%	**95%**
St-Malo	Religieux	99%	97%	75%	100%	**92%**	100%	98%	100%	99%	**99%**
	Non religieux	100%	92%	83%	100%	**93%**	100%	88%	75%	100%	**90%**
	Moulins	100%	100%	100%	100%	**100%**	100%	100%	100%	100%	**100%**
Grenoble	Religieux	99%	95%	82%	100%	**93%**	97%	99%	100%	67%	**90%**
	Non religieux	98%	100%	93%	96%	**97%**	100%	100%	100%	100%	**100%**
	Moulins	100%	93%	98%	100%	**98%**	100%	100%	100%	100%	**100%**

TABLE 2 – Résumé des résultats d'appariement.

0.3.4 Résultats complémentaires

0.3.4.1 Sensibilité aux seuils

Nous étudions dans cette section la sensibilité du processus aux fluctuations des paramètres autour de leurs valeurs déterminées par l'algorithme de paramétrisation automatique, en analysant les variations du F-Score moyen. Pour notre étude, nous nous focalisons sur la classe "Religieux" de la feuille de Reims.

0.3.4.1.1 Étude de la sensibilité du processus aux seuils du critère géométrique Afin d'étudier la sensibilité du processus aux paramètres du critère géométrique, nous faisons varier les seuils geoMin, geoMax et geoS l'un après l'autre en conservant les autres seuils constants.

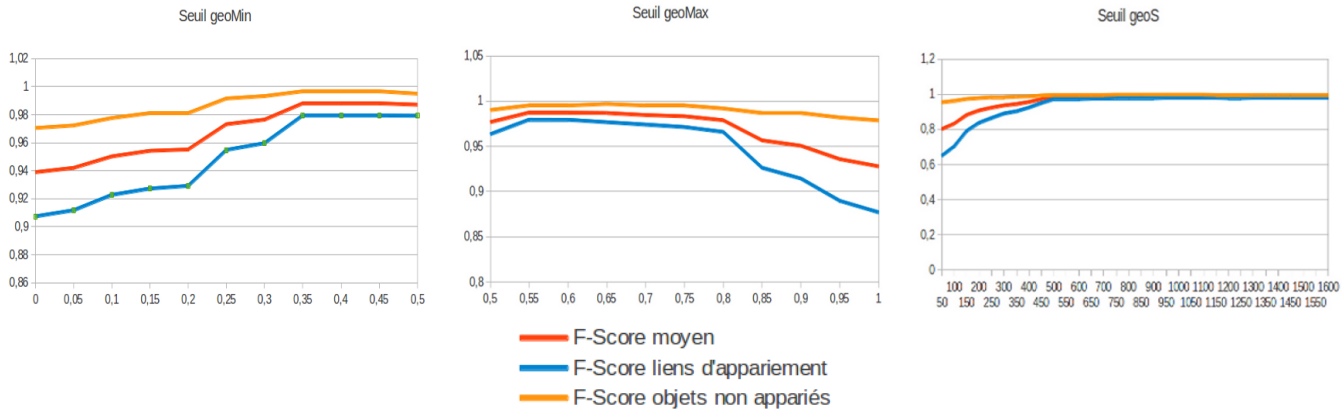


FIGURE 13 – Sensibilité du processus d'appariement aux paramètres du critères géométrique.

Nous constatons peu de sensibilité aux seuils geoMin et geoMax : le F-Score moyen varie dans un intervalle de valeurs d'amplitude 0.1, et reste supérieur à 0.9. Il augmente de 0.8 à près de 1.0 avec le paramètre geoS,atteint sa valeur maximale lorsque geoS vaut environ 450m et reste constant pour des valeurs de geoS supérieurs à 450. Cette évolution du F-Score moyen, et surtout du F-Score des liens d'appariement, pour des valeurs de geoS comprises entre 0 et environ 450, s'explique par le fait que beaucoup d'objets réellement en correspondance sont distants de plus de 450m, valeur pouvant s'interpréter comme étant relativement proche de la précision géométrique des données, et donc le nombre de faux positifs augmente pour des valeurs de geoS faibles. Cependant, le F-Score des liens d'appariement reste assez élevè (>60%) puisque le critère géométrique n'est pas le critère le plus

discriminant. En conclusion, choisir le seuil geoS supérieur à la précision géométrique des données nous semble pertinent, car il est difficile de privilégier deux candidats à l'appariement lorsque leur distance à l'objet Cassini étudié est inférieur à cette précision. Cette difficulté est prise en compte intelligemment par le processus qui privilégie souvent, lorsque le critère géométrique n'est pas le critère le plus important, l'utilisation d 'un courbe de pondération de type exponentielle. En effet, les distances géométriques de zéro à quelques centaines de mètres sont considérée comme proches, la discrimination entre candidats étant faite pour des distances supérieures à approximativement la précision géométrique des données.

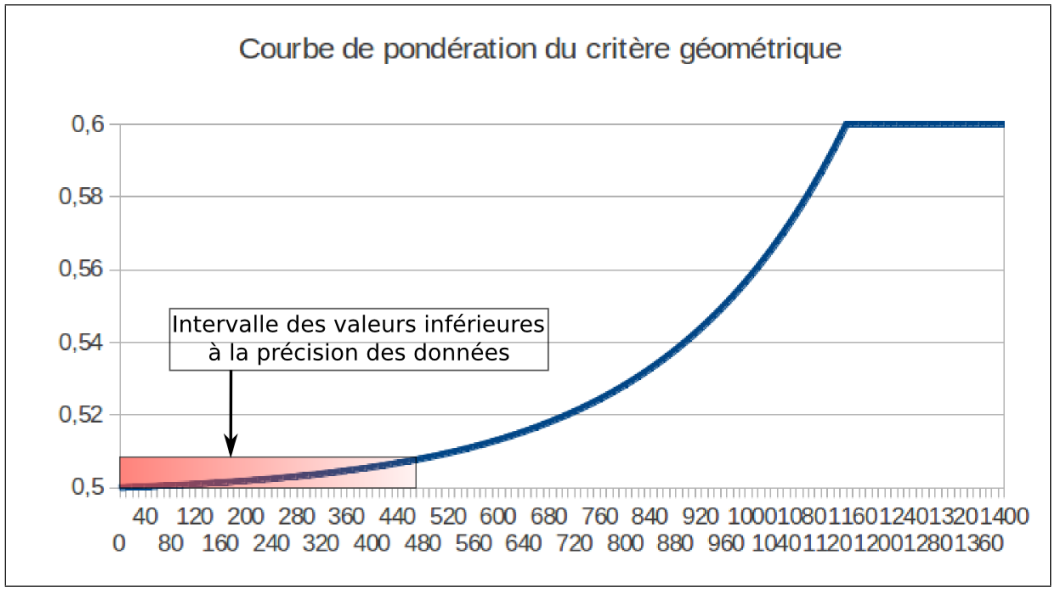


FIGURE 14 – La fonction de pondération du critère géométrique, de type exponentielle ($\lambda = 4$), varie peu (amplitude <0.01) sur l'intervalle des valeurs inférieures à la précision géométrique des données, représenté par le rectangle rouge. Le dégradé souligne le flou sur la valeur exacte de cette précision.

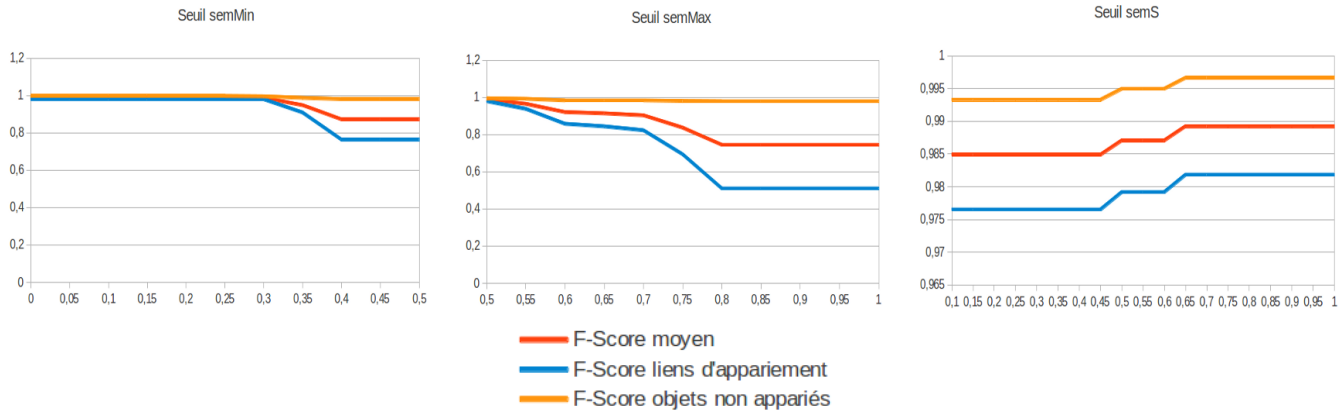


FIGURE 15 – Sensibilité du processus d'appariement aux paramètres du critère sémantique.

0.3.4.1.2 Étude de la sensibilité du processus aux seuils du critère sémantique Le processus est peu sensible aux variations des paramètres semMin et semMax du critère sémantique. De plus, le F-Score moyen est maximal et constant pour $semMin \leq 0.3$ et $semS \geq 0.6$. Lorsque le seuil semMax augmente, le critère sémantique devient plus permissif car les mesures pondérées de ce critère diminuent : il y a alors plus de candidats appariés n'ayant pas de rapport sémantique avec les objets Cassini. Une "bonne" valeur du seuil semMax semble être donc comprise en 0.5 et 0.7.

0.3.4.1.3 Étude de la sensibilité du processus aux seuils du critère toponymique Nous constatons peu de sensibilité aux seuils topMax et topS : le F-Score moyen varie dans un intervalle de valeurs d'amplitude inférieur à 0.2, et reste maximal et constant pour $topMax \leq 0.85$ et $topS \geq 0.7$. Le processus est néanmoins très sensible au paramètre topMin. Le F-Score moyen est maximal pour $topMin \leq 0.1$ mais diminue rapidement lorsque topMin augmente pour atteindre son minimum pour $topMin = 0.3$, le F-Score des liens d'appariement étant alors nul. Cette relation découle du fait que le critère toponymique est le critère le plus discriminant pour l'appariement d'objets de type "religieux". Au dessus de 0.3, la moyenne $\frac{topMin+semMin+geoMin}{3}$ est toujours supérieure au seuil global décision déterminé par l'algorithme de paramétrisation.

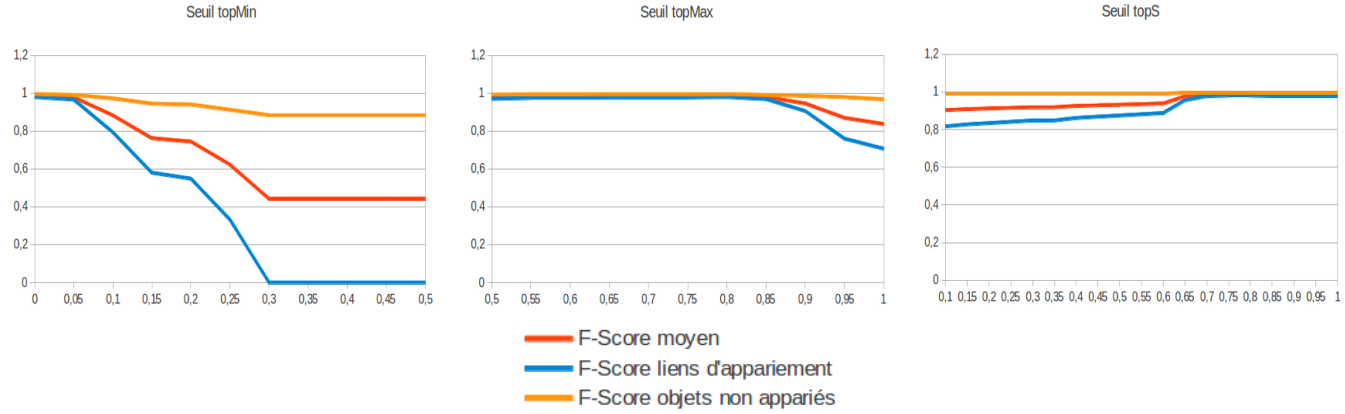


FIGURE 16 – Sensibilité du processus d'appariement aux paramètres du critère toponymique.

0.3.4.1.4 Étude de la sensibilité du processus aux seuil global de prise de décision Nous constatons que le F-Score moyen est minimal pour $seuil \leq 0.33$ (aucun objet n'est apparié, le seuil est trop bas), augmente entre 0.33 et 0.42 ou il atteint son maximum. Pour $seuil > 0.42$, le F-Score moyen diminue lentement, conséquence de l'apparition de sur-appariements : les vrai négatifs diminuent tandis que les faux positifs augmentent. Le poids global fusionné maximal vaut $\frac{topMax+semMax+geoMax}{3} = 0.63$, valeur au delà de laquelle le F-Score moyen est constant. La diminution du F-Score après avoir atteint

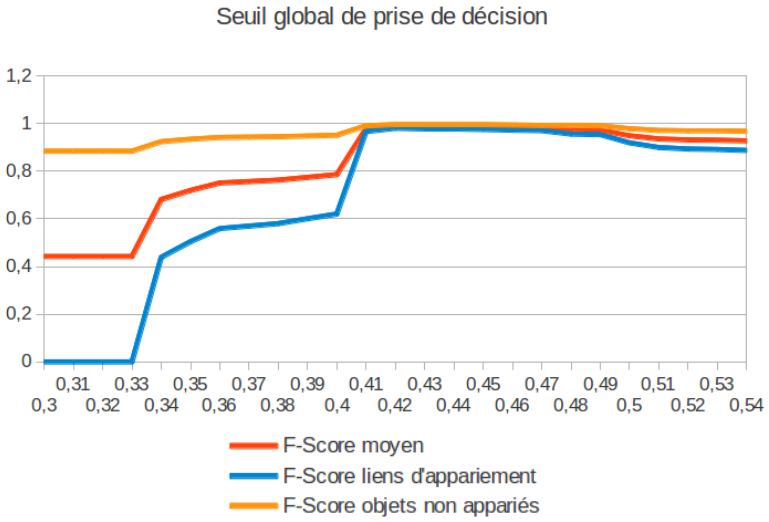


FIGURE 17 – Sensibilité du processus d'appariement au seuil global de prise de décision.

son maximum est ici très lente. Ce comportement est ici normal puisque le nombre d'objet réellement non appariés pour la classe "Religieux" est faible. Le nombre de faux positifs augmente donc peu. En revanche, pour une classe présentant peu d'objets appariés, cette diminution s'avère beaucoup plus

importante (faux positifs très nombreux), comme on peut le voir sur la figure 18, illustrant la sensibilité du processus d'appariement au seuil global de décision pour la classe "Moulins et activités industrielles".

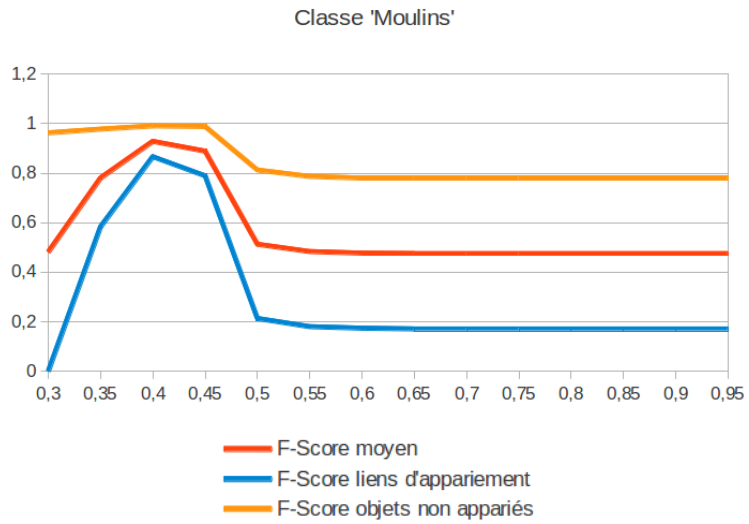


FIGURE 18 – Sensibilité du processus d'appariement au seuil global de prise de décision, pour la classe 'Moulins' de la feuille de Reims.

0.3.4.2 Vers une paramétrisation globale de l'appariement

Nous avons également créé, pour chaque classe, une base contenant l'union des objets des trois zones, et exécuté l'algorithme de paramétrisation sur ces données. Les bons résultats obtenus (cf. tableau 3) nous permettent d'envisager la définition d'une paramétrisation unique par classe, qui serait hypothétiquement applicable à l'ensemble du territoire français, au seuil de prise de décision près.

Classe	**Apprentissage**					**Validation**				
	Préc. LAP	**Rapp. LAP**	**Préc. NAP**	**Rapp. NAP**	**F-score mean**	**Préc. LAP**	**Rapp. LAP**	**Préc. NAP**	**Rapp. NAP**	**F-score mean**
Religieux	97%	97%	94%	97%	**96%**	97%	96%	88%	82%	**91%**
Non religieux	96%	95%	91%	97%	**95%**	93%	90%	87%	92%	**91%**
Moulins	93%	98%	99%	99%	**97%**	95%	90%	99%	100%	**96%**

TABLE 3 – Résultats d'appariement par classe pour la paramétrisation globale sur les trois zones.

0.3.4.3 Conclusion sur la paramétrisation

L'utilisation d'un échantillon de données appariées manuellement comme données d'apprentissage pour automatiser la paramétrisation donne de bons résultats. Les paramétrisation obtenues sur les mêmes classes de chaque zone sont relativement proches. Nous avons testé l'application d'une paramétrisation établie pour une classe sur la même classe d'une autre zone. Les résultats sont bons (entre 80 et 98% de F-Score moyen). Il est seulement parfois nécessaire de faire varier la valeur du seuil global de prise de décision.
Que nous utilisions une paramétrisation globale définie pour une classe sur n'importe quelle zone, ou une paramétrisation locale définie pour une classe sur une zone précise, nous serons amené à déterminer la valeur du seuil de prise de décision à utiliser. Cette estimation peut se faire par apprentissage comme nous l'avons vu dans notre approche. Cependant, l'appariement manuel est une tâche

Paramètres	Religieux	Non religieux	Moulins
Seuil global	0.35	0.4	0.42
Fonction géométrie	EXP, $\lambda = 1$	Linéaire	LOG, $\lambda = 8$
Valeur min. géométrie	0.2	0.3	0.2
Valeur max. géométrie	0.5	0.5	0.5
Seuil géométrie	1500	1200	1200
Fonction sémantique	LOG, $\lambda = 8$	Linéaire	,Linéaire
Valeur min. sémantique	0.2	0.3	0.1
Valeur max. sémantique	0.5	0.5	0.7
Seuil sémantique	0.4	0.6	0.5
Fonction toponymie	Linéaire	EXP, $\lambda = 1$	EXP, $\lambda = 1$
Valeur min. toponymie	0.0	0.0	0.2
Valeur max. toponymie	0.6	0.8	0.7
Seuil toponymie	0.8	0.8	0.8

TABLE 4 – Paramétrisations globale par classe, valable sur les trois zones.

relativement longue et délicate. Nous proposons une méthode alternative permettant de déterminer automatiquement une valeur de seuil global suivant une combinaison de fonctions donnée. Elle est basée sur l'étude de *rupture* dans la classification des natures des candidats appariés, c'est-à-dire à l'apparition de nouvelles natures d'objets appariés. Pour cela, nous analysons les valeurs du seuil pour lesquelles une nouvelle nature (valeur de l'attribut "nature") d'objet candidat est pris en compte et mis en correspondance avec un objet Cassini. Nous prenons comme hypothèse que la valeur de seuil optimal est celle pour laquelle une nature d'objet non désirée apparaît (*i.e.* une nature correspondant à une erreur d'appariement, par exemple avoir un moulin apparié avec un donjon). Des valeurs moyennes de rappel et de précision aux alentours de 80% sont ainsi obtenus avec cette méthode.

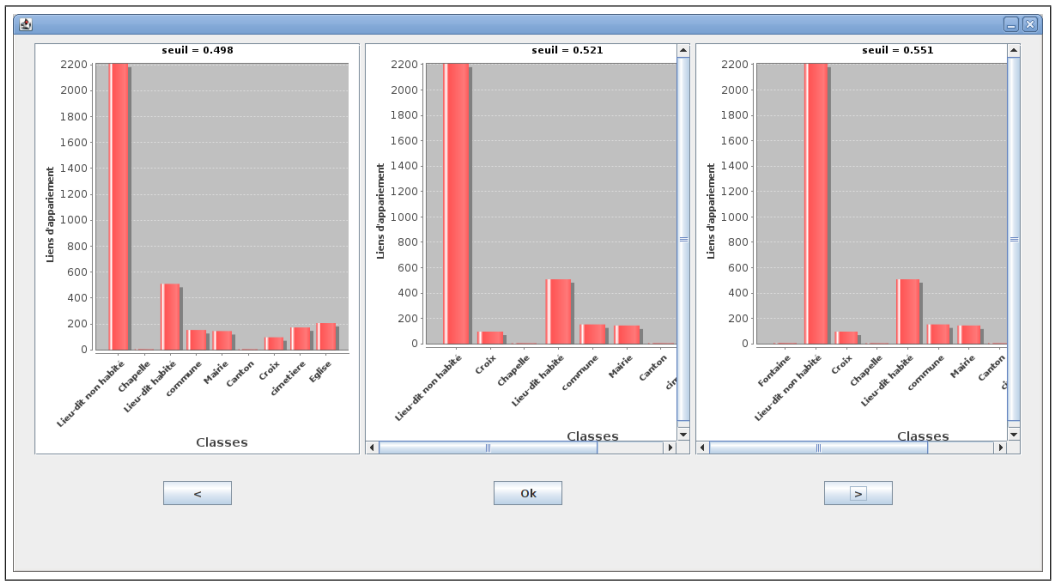


FIGURE 19 – Estimation du seuil de prise de décision par étude des ruptures dans la classification.

0.3.5 Conclusion sur l'appariement des données ponctuelles

L'intégration de données anciennes comme les cartes de Cassini, présentant des écarts d'actualité de plus de 250 ans avec le monde d'aujourd'hui, est un sujet difficile, du fait des fortes évolutions séman-

tique et toponymique des objets, et de l'imprécision de localisation des cartes utilisées. Pour répondre à cette problématique, que nous avons détaillé au long de l'article, nous avons proposé une méthode de géoréférencement adaptée, qui tient compte des déformations propres à chaque feuille de la carte. Nous avons par ailleurs démontré que si la transformée de Helmert utilisée peut induire de légers décalages, ceux-ci sont tout à fait comparables aux inexactitudes observées dans la carte. Ensuite, à partir de l'approche proposée par [Olteanu, 2008], nous avons proposé un processus d'appariement multi-critères, dont le paramétrage est facilité par un apprentissage sur données appariées manuellement, et qui permet l'analyse visuelle des liens de correspondance entre objets étudiés et leurs candidats.
L 'élaboration d'une ontologie de domaine spatio-temporelle devrait permettre d'améliorer l'évaluation de la distance sémantique, en tenant compte des époques des objets à apparier. Par ailleurs, certains toponymes d'objets homologues présentent d'importantes modifications orthographiques ("Le Ménil L'Epinpoy" devenant "Ménil-Lepinois"). Afin de réduire les erreurs dues à la toponymie nous envisageons d'exploiter une base de données recensant les évolutions des toponymes des communes de France [Motte et al., 2003]. A plus long terme, l'étude des correspondances avec des données intermédiaires comme les minutes d'État-major ou les fonds de carte de 1960 de l'IGN pourra permettre de valider ou d'infirmer les résultats obtenus par notre approche, mais aussi de détecter des erreurs de représentation et de contenu sur les cartes anciennes entrainant des incohérences lors de l'appariement.

Bibliographie

Nathalie Abadie and Sébastien Mustière. Création d' unetaxonomie géographique à partir des spécifications de bases de données. In *Actes de SAGEO 2008*, Montepellier, 2008.

C. Beeri, Y. Kanza, E. Safra, and Y. Sagiv. Object fusion in geographic information systems. In *Proceedings of the 30th VLDB Conference*, Toronto, 2004.

D. Blasby, M. Davis, D. Kim, and P. Ramsey. Gis conflation using open source tools. Rapport technique, 2004. URL http://www.jump-project.org/assets/JUMP_Conflation_Whitepaper.pdf.

A. D. Clodoveu and F. Fonseca. Assessing the certainty of locations produced by an address geocoding system. *GeoInformatica*, 11(1) :103–129, 2007.

Thomas Devogele. *Processus d'intégration et d'appariement de bases de données géographiques. Application à une base de données routières multi-échelles.* Thèse de doctorat, Université de Versailles, 1997.

J.H. Haunert. Link based conflation of geographic datasets. In *In Proceedings of the 8th ICA Workshop on Generalisation and Multiple Representation*, la Corogne, 7-8 juillet 2005.

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

P. Lüscher, D. Burghardt, and R. Weibel. Matching road data of scales with an order of magnitude difference. In *In Proceedings of the XXIII International Cartographic Conference*, Moscou, 4-10 août 2007.

Claude Motte, Isabelle Séguy, Christine Théré, and Dominique Tixier-Basse. *Communes d'hier, communes d'aujourd'hui. Les communes de la France métropolitaine, 1801-2001. Dictionnaire d'histoire administrative.* Institut National d'Études Démographiques, Paris, 2003.

Sébastien Mustière and Thomas Devogele. Matching networks with different levels of detail. *GeoInformatica*, 12(4) :435–453, December 2008.

A-M. Olteanu. *Fusion de connaissances imparfaites pour l'appariement de données géographiques. Proposition d'une approche s'appuyant sur la théorie des fonctions de croyance.* Thèse de doctorat, Université Paris-Est, 2008.

E. Safra, Y. Kanza, Y. Sagiv, and Y. Doytsher. Efficient integration of road maps. In ACM Press, editor, *In Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems*, pages 59–66, Arlington (Etats-Unis), 10-11 novembre 2006.

H. Sui, D. Li, and J. Gong. Automatic feature-level change detection (flcd) for road network. In *In Proceedings of the 20th ISPRS Congress*, Istanbul, 12-23 juillet 2004.

S. Voltz. An iterative approach for matching multiple representations of street data. In *In Proceedings of ISPRS Workshop, Multiple representation and interoperability of spatial data*, pages 101–110, Hanovre (Allemagne), 22-24 février 2006.

V. Walter and D. Fritsch. Matching spatial data sets : Statistical approach. *International Journal of Geographical Information Science*, 13(5) :445–473, 1999.

Z. Wu and M. Palmer. Verb semantics and lexical selection. In *In Proceedings of the 32nd Annual Meetings of the Associations for Computational Linguistics*, pages 133–138, 1994.

M. Zhang, W. Shi, and L. Meng. A generic matching algorithm for line networks of different resolutions. In *In Proceedings of ICA Workshop on Generalisation and Multiple Representation*, La Corogne, 7-8 juillet 2005.