Fusion de connaissances imparfaites pour l'appariement de données géographiques : proposition d'une approche s'appuyant sur la théorie des fonctions de croyance

Ana-Maria Olteanu

UNIVERSITÉ PARIS-EST

THÈSE

pour obtenir le grade de

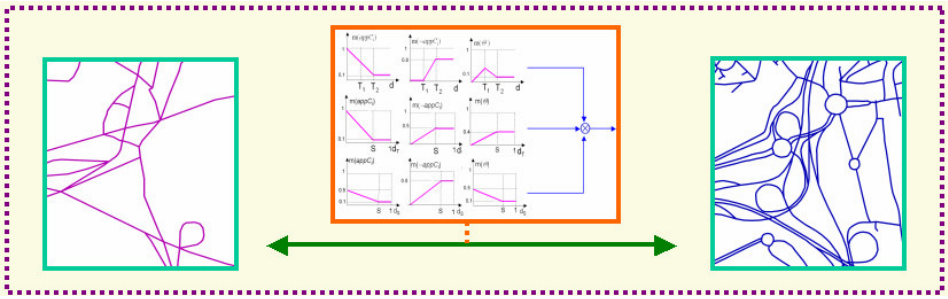
Docteur de l'Université Paris-Est

Spécialité : Sciences de l'Information Géographique

présentée et soutenue publiquement par

Ana-Maria OLTEANU

le 24 octobre 2008



Fusion de connaissances imparfaites pour l'appariement de données géographiques

Proposition d'une approche s'appuyant sur la théorie des fonctions de croyance

Jury :
Mme BOUCHON-MEUNIER Bernadette – Rapporteur, Directeur de recherche au CNRS
M. CLARAMUNT Christophe – Rapporteur, Professeur
M. MUSTIERE Sébastien – Co-Encadrant, Docteur
Mme REYNAUD Chantal – Examinateur, Professeur
Mme RUAS Anne – Directrice de thèse, HDR
Mme SERVIGNE Sylvie – Examinateur, Maître de conférence

Université Paris-Est – Ecole doctorale ICMS
Institut Géographique National - Laboratoire COGIT

à ma mère, nouvelle étoile

à Olivier

Remerciements

Je tiens d'abord à remercier Anne RUAS, directrice du laboratoire COGIT, pour m'avoir accueillie au sein de son laboratoire. Je voudrais aussi la remercier pour avoir accepté de diriger cette thèse, pour ses conseils avisés, sa disponibilité et son soutien.

Je remercie les membres du jury et tout particulièrement mes deux rapporteurs Mme Bernadette BOUCHON-MEUNIER et M. Christophe CLARAMUNT, pour avoir accepté de juger cette thèse, pour leurs remarques pertinentes et leurs questions très intéressantes.

Je tiens aussi à adresser un grand merci à Sébastien MUSTIERE pour la qualité de son co-encadrement, ses multiples relectures de mes articles et bien évidemment de ce rapport de thèse, ses conseils avisés et le soutien qu'il m'a apporté. Je lui témoigne tout mon respect et ma gratitude.

Je souhaite aussi remercier l'ensemble des collègues du laboratoire COGIT que j'ai côtoyés pendant trois ans. J'ai passé de très bons moments et j'ai beaucoup apprécié les divers échanges tant au niveau professionnel qu'au niveau personnel, leur gentillesse, leur bonne humeur. J'ai une pensée particulière pour mes collègues de bureau Cécile, Elisabeth et Laurence. Je leurs remercie pour leur gentillesse, leurs mots d'encouragement et leurs conseils.

Je ne peux pas oublier mes collègues sportifs et amis Dominique, Corina, Elodie, Marie-Lise, Maryse et Patrick, avec qui j'ai partagé de bons moments en faisant le tour du bois de Vincennes ou de la gymnastique. Ces parenthèses sportives ont été souvent une source d'inspiration pour la thèse.

Je voudrais adresser un grand merci à mes relecteurs Olivier, Christian, Elodie, Laurence et Sidonie.

Je remercie également ma famille et ma belle-famille pour leur aide et leur soutien.

Enfin, MERCI à toi Olivier, pour ta patience, ton soutien sans faille, ton aide, ton écoute et surtout pour tout ce que tu m'apportes au quotidien.

Résumé

De nos jours, il existe de nombreuses bases de données géographiques (BDG) couvrant le même territoire. Les données géographiques sont modélisées différemment (par exemple une rivière peut être modélisée par une ligne ou bien par une surface), elles sont destinées à répondre à plusieurs applications (visualisation, analyse) et elles sont créées suivant des modes d'acquisition divers (sources, processus). Tous ces facteurs créent une indépendance entre les BDG, qui pose certains problèmes à la fois aux producteurs et aux utilisateurs.

Ainsi, une solution est d'expliciter les relations entre les divers objets des bases de données, c'est-à-dire de mettre en correspondance des objets homologues représentant la même réalité. Ce processus est connu sous le nom d'appariement de données géographiques.

La complexité du processus d'appariement fait que les approches existantes varient en fonction des besoins auxquels l'appariement répond, et dépendent des types de données à apparier (points, lignes ou surfaces) et du niveau de détail. Nous avons remarqué que la plupart des approches sont basées sur la géométrie et les relations topologiques des objets géographiques et très peu sont celles qui prennent en compte l'information descriptive des objets géographiques. De plus, pour la plupart des approches, les critères sont enchaînés et les connaissances sont à l'intérieur du processus.

Suite à cette analyse, nous proposons une approche d'appariement de données qui est guidée par des connaissances et qui prend en compte tous les critères simultanément en exploitant à la fois la géométrie, l'information descriptive et les relations entre eux. Afin de formaliser les connaissances et de modéliser leurs imperfections (imprécision, incertitude et incomplétude), nous avons utilisé la théorie des fonctions de croyance [Shafer, 1976].

Notre approche d'appariement de données est composée de cinq étapes : après une sélection des candidats, nous initialisons les masses de croyance en analysant chaque candidat indépendamment des autres au moyen des différentes connaissances exprimées par divers critères d'appariement. Ensuite, nous fusionnons les critères d'appariement et les candidats. Enfin, une décision est prise.

Nous avons testé notre approche sur des données réelles ayant des niveaux de détail différents représentant le relief (données ponctuelles) et les réseaux routiers (données linéaires).

Mots-clés : appariement, données géographiques, fusion, connaissances, imperfection

Abstract

Nowadays, there are many geographic databases, (GDB), covering the same reality. The geographical data are represented differently (for example a river can be represented by a line or a polygon), they are used in different applications (visualisation, analysis) and they are created using various modes of acquisition (sources, processes). All these factors create independence between GDB, which causes problems for both producers and users.

Thus, a solution is to clarify the relationships between various database objects, i.e. to match homologous objects, which represent the same reality. This process is known as spatial data matching.

Because of the complexity of the matching process, the existing approaches depend on the types of data (points, lines or polygons) and the level of detail of the GDB. We realised, that most of the approaches are based on the geometry and the topology of the geographical objects, and very few approaches take into account the descriptive information of geographical objects. Besides, for most approaches, the criteria are applied one after the other and knowledge is contained within the process.

Following this analysis, we proposed a matching approach that is guided by knowledge and takes into account all criteria at the same time exploiting the geometry, descriptive information and relations between geographical objects. In order to formalise knowledge and model their imperfections (imprecision, uncertainty and incompleteness), we used the Belief Theory [Shafer, 1976].

Our approach of the data matching is composed of five steps. After a selection of candidates, the masses of beliefs are initialised by analysing each candidate separately from the others using different knowledge expressed by various matching criteria. Then, the matching criteria and candidates are fusioned. Finally, a decision is taken.

Our approach has been tested on real data having different levels of detail and representing relief (data points) and road networks (linear data).

Keywords : matching, geographical data, fusion, knowledge, imperfection

Table des matières

INTRODUCTION 24

A APPARIEMENT DE DONNEES GEOGRAPHIQUES 32

A.1 L'APPARIEMENT, UN PROBLEME COMPLEXE 32

A.2 L'APPARIEMENT, UN OUTIL POUR REPONDRE A PLUSIEURS BESOINS 34

A.2.1 APPARIER POUR EVALUER LA QUALITE DES DONNEES GEOGRAPHIQUES 34

A.2.2 APPARIER POUR RECALER DES DONNEES GEOGRAPHIQUES 36

A.2.3 APPARIER POUR METTRE A JOUR LES DONNEES GEOGRAPHIQUES 38

A.2.4 INTEGRATION DE BASES DE DONNEES GEOGRAPHIQUES HETEROGENES 40

A.2.4.1 Pré-intégration 41

A.2.4.2 Appariement des schémas et des données 41

A.2.4.3 Intégration 42

A.3 L'APPARIEMENT, UN OUTIL QUI DEPEND DES DONNEES GEOGRAPHIQUES 44

A.3.1 APPARIEMENT DE RESEAUX AU MEME NIVEAU DE DETAIL, APPROCHE DE [WALTER ET FRITCH, 1999] 44

A.3.2 APPARIEMENT DE RESEAUX A DES NIVEAUX DE DETAIL DIFFERENTS, APPROCHE DE [MUSTIERE ET DEVOGELE, 2008] 46

A.3.3 APPARIEMENT DES JEUX DE DONNEES SURFACIQUES, APPROCHE DE [BEL HADJ ALI, 2001] 47

A.3.4 APPARIEMENT DE PLUS DE DEUX JEUX DE DONNEES, APPROCHE DE [SAMAL ET AL., 2004] 48

A.4 CRITERES D'APPARIEMENT DE DONNEES ET LEUR COMBINAISON 49

A.4.1 DIFFERENTS CRITERES D'APPARIEMENT 50

A.4.1.1 Critères géométriques 50

A.4.1.2 Critères topologiques et de voisinage 51

A.4.1.3 Critères attributaires 52
A.4.1.4 Bilan sur les critères d'appariement 53
A.4.2 DIFFERENTES MESURES UTILISEES DANS LE PROCESSUS D'APPARIEMENT 54
A.4.2.1 Mesures comparant les géométries 54
A.4.2.2 Evaluation de la ressemblance entre les toponymes 61
A.4.2.3 Evaluation de la ressemblance des concepts à travers la sémantique 63
A.4.2.4 Mesures utilisant les relations topologiques 64
A.4.3 LES ETAPES GENERALES DU PROCESSUS D'APPARIEMENT DE DONNEES GEOGRAPHIQUES 65
A.5 APPARIEMENT ET IMPERFECTION DANS LES DONNEES GEOGRAPHIQUES 70
A.6 CONCLUSION ET DEFINITION DU SUJET 73

B IMPERFECTION, REPRESENTATION ET FUSION DES CONNAISSANCES...78

B.1 ANALYSE DES IMPERFECTIONS DES DONNEES GEOGRAPHIQUES 80
B.2 REPRESENTATION DE L'IMPERFECTION EN UTILISANT LA THEORIE DES FONCTIONS DE CROYANCE 82
B.2.1 MOTIVATION DU CHOIX DE LA THEORIE DES FONCTIONS DE CROYANCE 83
B.2.2 APERÇU SUR LES THEORIES DE L'INCERTAIN 87
B.3 QUELQUES APPLICATIONS DE LA THEORIE DES FONCTIONS DE CROYANCE 88
B.3.1 ANALYSE DE DONNEES 89
B.3.2 TRAITEMENT D'IMAGE 90
B.3.3 GEOMATIQUE 91
B.4 CADRE GENERAL DE LA THEORIE DES FONCTIONS DE CROYANCE 91
B.4.1 REPRESENTATION EXPLICITE DES CONNAISSANCES 92
B.4.2 INITIALISATION DES MASSES DE CROYANCE 95
B.4.3 COMBINAISON DES SOURCES D'INFORMATION 96
B.4.4 ANALYSE ET REDISTRIBUTION DU CONFLIT 98
B.4.5 AFFAIBLISSEMENT DES SOURCES 99
B.4.6 DECISION 100
B.5 APPROCHES DES SOURCES SPECIALISEES 101
B.6 CONCLUSION 103

C PROCESSUS D'APPARIEMENT DE DONNEES GEOGRAPHIQUES BASE SUR LA THEORIE DES FONCTIONS DE CROYANCE 106

C.1 LE PROCESSUS D'APPARIEMENT DE DONNEES GEOGRAPHIQUES 106
C.1.1 SELECTION DES CANDIDATS 107
C.1.2 INITIALISATION DES MASSES DE CROYANCE 109
C.1.3 FUSION DES CRITERES 111
C.1.4 FUSION DES CANDIDATS 113
C.1.5 DECISION 115
C.1.6 DISCUSSION SUR LA STABILITE DU PROCESSUS 118
C.2 MODELISATION DES CRITERES D'APPARIEMENT 120
C.2.1 CONNAISSANCES SUR LA GEOMETRIE 121
C.2.1.1 Initialisation des masses de croyance pour le critère d'écart de position 122
C.2.1.2 Initialisation des masses de croyance pour le critère écart d'orientation 123
C.2.2 CONNAISSANCES SUR LA SEMANTIQUE 124
C.2.2.1 Problématique 125
C.2.2.2 Initialisation des masses de croyance 127
C.2.3 CONNAISSANCES SUR LA TOPONYMIE OU SUR LES NOMS DES OBJETS GEOGRAPHIQUES 129
C.2.3.1 Problématique 129
C.2.3.2 Initialisation des masses de croyance 129
C.2.4 CONNAISSANCES SUR LE VOISINAGE 133
C.2.4.1 Problématique 133
C.2.4.2 Initialisation des masses de croyance 134
C.3 CONCLUSION 139

D EXPERIMENTATIONS ET EVALUATION 142

D.1 MISE EN OEUVRE 142
D.1.1 PROTOTYPE: PLATE-FORME GEOXYGENE 142
D.1.2 INTERFACE D'APPARIEMENT DE DONNEES 144
D.1.3 SCHEMA CONCEPTUEL DE DONNEES 146
D.1.3.1 Classes de GeOxygene utilisées 146

D.1.3.2 Schéma des données proposé pour l'appariement 148
D.2 ETUDE DES POINTS REMARQUABLES DU RELIEF 154
D.2.1 PRESENTATION DES DONNEES 154
D.2.2 TESTS 157
D.2.2.1 Paramétrage des courbes pour l'initialisation des masses de croyance 158
D.2.2.2 Résultats des expérimentations: exemples 163
D.2.3 EVALUATION QUANTITATIVE 173
D.3 ETUDE DES RESEAUX ROUTIERS 179
D.3.1 PRESENTATION DES DONNEES 179
D.3.2 TESTS 182
D.3.2.1 Paramétrage des courbes pour l'initialisation des masses de croyance 182
D.3.2.2 Résultats des expérimentations : exemples 192
D.3.2.3 Evaluation quantitative et discussion 200
D.3.2.4 Etude de la convergence du processus 206
D.3.2.5 Bilan des expérimentations sur les réseaux routiers 208
D.4 CONCLUSION 209

E VERS LA CONCEPTION D'UN SYSTEME GENERIQUE D'APPARIEMENT DE DONNEES GEOGRAPHIQUES 212

E.1 INTRODUCTION 212
E.2 DESCRIPTION DU PROCESSUS D'APPARIEMENT DE DONNEES (PROPOSITION) 213
E.3 PRE-TRAITEMENT DES DONNEES 215
E.4 APPARIEMENT AUTOMATIQUE 215
E.4.1 CHOIX DU SENS DE L'APPARIEMENT 215
E.4.2 CHOIX DU SEUIL DE SELECTION 217
E.4.3 INITIALISATION DES MASSES DE CROYANCE 218
E.4.4 FUSION DES CRITERES ET DES CANDIDATS 226
E.4.5 DECISION 227
E.4.6 AUTO-EVALUATION 227
E.5 APPARIEMENT INTERACTIF 227
E.6 QUALIFICATION DES RESULTATS D'APPARIEMENT 229
E.7 CONCLUSION 229

CONCLUSION ET PERSPECTIVES.....232

BIBLIOGRAPHIE 239

ANNEXES 257

Liste des tableaux

Tableau 1. Décision finale en fonction de la redistribution du conflit 120
Tableau 2. Représentation des connaissances pour le critère d'écart de position..... 122
Tableau 3. Représentation des connaissances pour le critère d'orientation 124
Tableau 4. Représentation des connaissances pour le critère sémantique..... 128
Tableau 5. Représentation des connaissances pour le critère toponymique 131
Tableau 6. Représentation des connaissances pour le critère nom d'objet..... 132
Tableau 7. Représentation des connaissances pour le critère voisinage-objets appariés (à gauche) dans les quatre cas définis à droite. 136
Tableau 8. Représentation des connaissances pour le critère voisinage-objets non-appariés (à gauche) dans les quatre cas définis à droite. 138
Tableau 9. Nombre d'objets dans les jeux de données utilisés dans les expérimentations, et les caractéristiques des départements couverts par les jeux de données..... 155
Tableau 10. Représentation des connaissances pour le critère d'écart de position..... 158
Tableau 11. Représentation des connaissances pour le critère sémantique..... 161
Tableau 12. Matrice « mot-mot » pour deux toponymes et les mesures de similarité entre les mots..... 162
Tableau 13. Représentation des connaissances pour le critère toponymique 163
Tableau 14. Matrice de confusion pour les données appariées pour le département 66 en utilisant deux critères 169
Tableau 15. Matrice de confusion pour les données appariées du département 66 en utilisant trois critères 170
Tableau 16. Evaluation des résultats pour les départements 11, 31, 64, 66 et 69..... 174
Tableau 17. Evaluation des résultats pour le département 64 en utilisant trois approches d'appariement 177
Tableau 18. Temps de calcul par département pour le processus d'appariement..... 178
Tableau 19. Représentation des connaissances pour le critère orientation..... 185
Tableau 20. Correspondances entre le type de classement et les instances MultiNet..... 188

Tableau 21. Représentation des connaissances pour le critère nom d'objet (à gauche) et les quatre cas définis (à droite) 189
Tableau 22. Représentation des connaissances pour le critère voisinage-objets appariés (à gauche) pour les quatre cas définis à droite (voir la partie C.2.4) 191
Tableau 23. Représentation des connaissances pour le critère voisinage-objets non-appariés (à gauche) pour les quatre cas définis à droite (voir la partie C.2.4). 192
Tableau 24. Evaluation des résultats d'appariement pour les réseaux 201
Tableau 25. Evaluation des résultats d'appariement pour les réseaux en utilisant notre approche et l'approche de [Mustière et Devogele, 2008] 204
Tableau 26. Tableau à remplir par les experts avec les distances sémantiques entre la nature des objets issus de la BDCARTO et de la BDTOPO 264
Tableau 27. Valeurs moyennes des distances sémantiques fournies par les experts 265
Tableau 28. Valeurs des distances sémantiques calculées automatiquement à partir de la taxonomie de domaine [Abadie et Mustière, 2008]..... 266
Tableau 29. Distances sémantiques entre les types de route de la BDCARTO et de MultiNet 267

Liste des figures

Figure 1. Modélisation des entités géographiques : monde réel et représentation vectorielle.24
Figure 2. Le monde réel (image a) et trois représentations différentes issues de BDCARTO (image b), BDTOPO (image c) et MultiNet(image d).....26
Figure 3. Contexte de l'appariement de données 27
Figure 4. Types de relation entre les jeux de données à apparier.....33
Figure 5. Exemples de liens d'appariement de cardinalités différentes en raison du niveau de détail différent 1 – n, (à gauche) et du découpage du monde réel différent n – m, (à droite).....33
Figure 6. Cardinalité des liens d'appariement.....34
Figure 7. Concept de qualité d'une base de données géographiques 35
Figure 8. Recalage de deux jeux de données vecteur-vecteur (jeux de données, initiaux à gauche et recalés à droite).....36
Figure 9. Recalage de deux jeux de données vecteur-raster (jeux de données, initiaux à gauche et recalés à droite).....36
Figure 10. Processus général de mise à jour 38
Figure 11. Processus d'intégration de bases de données géographiques.....40
Figure 12. Explicitation de l'information implicite : cas d'une patte d'oie.....41
Figure 13. Appariement des schémas dérivé de l'appariement des données, d'après [Voltz, 2005] 42
Figure 14. Le processus de conflation proposé par [Yaun et Tao, 1999].....43
Figure 15. Détermination du seuil optimal de filtrage au moyen d'une courbe de distribution de fréquence 45
Figure 16. Pré-appariement des nœuds (à gauche) et pré-appariement des arcs (à droite), d'après [Mustière et Devogele, 2008].....46
Figure 17. Appariement des nœuds (à gauche) et appariement des arcs (à droite), d'après [Mustière et Devogele, 2008].....47
Figure 19. Graphes de proximité et vecteurs de décalage, d'après [Samal *et al.*, 2004].....49
Figure 20. Exemples d'appariement utilisant la géométrie 51

Figure 21. Exemples de relations spatiales entre les objets géographiques.....52
Figure 22. Les caractéristiques d'un objet géographique 53
Figure 23. Distance euclidienne entre deux objets géographiques ponctuels.....54
Figure 24. Distance de Hausdorff entre deux lignes 55
Figure 25. Distance de Hausdorff et de Fréchet, d'après [Badard et Lemarié, 2002].....56
Figure 26. Distance moyenne entre deux lignes 57
Figure 27. Degré de co-linéarité local θ de deux polylignes L_1 et L_2 58
Figure 28. Appariements issus de la mesure basée sur la bande epsilon.....58
Figure 29. Analyse de deux objets linéaires en utilisant la bande epsilon, [Sui et al., 2004]..59
Figure 30. Distance surfacique entre deux objets surfaciques 59
Figure 31. Fonction à distance radiale d'un objet surfacique, [Bel Hadj Ali, 2001].....60
Figure 32. Fonction angulaire d'un objet surfacique.....61
Figure 33. Exemple de calcul de profondeur des concepts dans le cadre de la mesure Wu-Palmer [Wu et Palmer, 1994] 64
Figure 34. Relations topologiques binaires entre deux surfaces 65
Figure 35. Relations topologiques binaires entre une ligne et une surface et leurs relations de proximité 65
Figure 36. Sélection des candidats basée sur un buffer grandissant, [Zhang *et al.*, 2005].....67
Figure 37. Modèle conceptuel de l'incertitude selon [Fisher, 2003].....71
Figure 38. L'imperfection au cours des processus manipulant des données géographiques...72
Figure 39. Appariement par enchaînement des critères.....74
Figure 40. Appariement par combinaison des critères en parallèle 75
Figure 41. Grands types d'imperfection 78
Figure 42. Taxonomie des imperfections selon [Niskanen, 1989] traduite de l'anglais 79
Figure 43. Imperfection de la localisation des données géographiques 81
Figure 44. Imperfection dans les attributs des données géographiques (la nature de l'objet sur l'image à gauche et le toponyme sur l'image à droite).....82
Figure 45. Sélection des candidats à l'appariement 85
Figure 46. Résultat d'appariement basé sur la théorie des fonctions de croyance.....87
Figure 47. Position de la théorie des possibilités et de la théorie des probabilités dans le cadre de la théorie de l'évidence, d'après [Bouchon-Meunier, 1995], page 89. 88
Figure 48. Sélection des objets candidats à l'appariement 92

Figure 49. La fonction de crédibilité définie dans le cadre de discernement 94
Figure 50. La fonction de plausibilité définie dans le cadre de discernement 95
Figure 51. Processus d'appariement de données géographiques détaillé 107
Figure 52. Etape de sélection des candidats à l'appariement.....107
Figure 53. Candidats à l'appariement pour l'objet obj_1 sélectionnés selon un critère de distance géométrique 108
Figure 54. Exemple typique d'appariement de données géographiques 109
Figure 55. Deuxième étape du processus : initialisation des masses de croyance 109
Figure 56. Troisième étape du processus : fusion des critères.....111
Figure 57. Fusion des critères pour le candidat C_i 112
Figure 58. Jeux de masses après la combinaison des critères pour chaque candidat 113
Figure 59. Quatrième étape du processus : fusion des candidats.....113
Figure 60. Combinaison des masses de croyance pour chacun des candidats.....114
Figure 61. Fusion des candidats 115
Figure 62. Cinquième étape du processus : décision.....115
Figure 63. Décision selon le maximum de probabilité pignistique.....116
Figure 64. Jeu de masses de croyance après normalisation 117
Figure 65. Variation du coefficient de normalisation en fonction du conflit, [Colot, 2000].118
Figure 66. Exemple de jeu de masses après la fusion des quatre candidats 119
Figure 67. Redistribution du conflit selon plusieurs opérateurs.....120
Figure 68. Extrait de la taxonomie réalisée pour les points remarquables du relief par [Abadie et Mustière, 2008].....126
Figure 69. Processus d'appariement basé sur des résultats d'appariement 134
Figure 70. L'architecture de la plate-forme GeOxygene d'après [Badard et Braun, 2003] ..143
Figure 71. Interface de l'appariement de données géographiques, en mode multi-fenêtres..144
Figure 72. Exemple d'affichage des résultats d'appariement de données entre deux lignes.145
Figure 73. Exemple d'analyse des résultats en utilisant la comparaison des attributs.....146
Figure 74. Les classes d'objets relatives à la géométrie des objets géographiques 147
Figure 75. Classes d'objets de base de la plate-forme GeOxygene [Guide utilisateur] 147
Figure 76. Schéma conceptuel de la Carte Topologique défini dans la plate-forme GeOxygene [David, 1988] 148
Figure 77. Extrait du modèle conceptuel de données : les classes géométriques 149

Figure 78. Extrait du modèle conceptuel de données : analyse individuelle de chaque candidat 150
Figure 79. Extrait du modèle conceptuel de données : définition des hypothèses.....151
Figure 80. Modèle conceptuel de données du processus d'appariement basé sur la théorie des fonctions de croyance 153
Figure 81. Représentation des points remarquables du relief dans la BDCARTO (à gauche) et la BDTOPO (à droite).....154
Figure 82. Classes d'objets de la BDCARTO et de la BDTOPO représentant les points remarquables du relief.....156
Figure 83. Taxonomie pour la classe entité du relief terrestre d'après [Abadie et Mustière, 2008] 160
Figure 84. Taxonomie pour la classe entité maritime d'après [Abadie et Mustière, 2008]...161
Figure 85. Exemple typique de résultat d'appariement des points remarquables du relief...164
Figure 86. Résultat d'appariement illustrant le fait que le processus n'apparie pas au plus proche objet (à gauche) et résultat d'appariement basé sur l'approche géométrique de [Beeri *et al.*, 2004] (à droite).....165
Figure 87. Résultat d'appariement illustrant le fait que le processus apparie deux objets ayant des toponymes différents (à gauche), et résultat d'appariement basé sur l'approche de [BDCARTO-BDTOPO, 2005] (à droite).....165
Figure 89. Exemple de résultat d'appariement lorsque l'objet n'est pas apparié 167
Figure 90. Exemple de résultat d'appariement obtenu en utilisant deux critères (à gauche) et trois critères (à droite).....168
Figure 91. Exemple de résultat d'appariement obtenu en utilisant deux critères (à gauche) et trois critères (à droite).....171
Figure 92. Exemple de sur-appariement (à gauche) et carte au 1 : 25 000 de la zone représentant notre exemple (à droite) 171
Figure 93. Exemple de sous-appariement.....172
Figure 94. Exemple de conflit de fusion.....173
Figure 95. Histogrammes illustrant l'évaluation des liens pour les départements 11 (en haut à gauche), 31 (en haut à droite) 64 (au milieu à gauche), 66 (au milieu à droite) et 69 (en bas).....176
Figure 96. La zone d'étude utilisée : la BDCARTO (à gauche) et MultiNet (à droite) 180

Figure 97. Représentation des autoroutes dans la BDCARTO (à gauche) et MultiNet (à droite).....181
Figure 98. Représentation d'un rond-point dans la BDCARTO (à gauche) et dans MultiNet à (droite).....181
Figure 99. Exemple de demi-distance de Hausdorff entre deux arcs ayant les extrémités décalées 184
Figure 100. Taxonomie pour la classe « voie de communication carrossable » [Abadie et Mustière, 2008].....186
Figure 101. Taxonomie pour la classe « route » [Abadie et Mustière, 2008].....186
Figure 102. Taxonomie pour la classe « voie de communication non carrossable » [Abadie et Mustière, 2008].....187
Figure 103. Exemple typique d'appariement d'arcs.....193
Figure 104. Résultats d'appariement sans le critère nom d'objet 194
Figure 105. Résultats d'appariement avec le critère nom d'objet.....194
Figure 106. Arcs de MultiNet appariés en utilisant le critère voisinage 195
Figure 107. Exemple d'arc de MultiNet apparié après la deuxième passe du processus 196
Figure 108. Résultat d'appariement d'arcs obtenu sans le critère voisinage.....197
Figure 109. Résultat d'appariement d'arcs obtenu avec le critère voisinage.....197
Figure 110. Exemple de résultat d'appariement obtenu dans le cas d'un rond-point 198
Figure 111. Exemple de conflit total signalé par notre processus.....199
Figure 112. Résultats d'appariement obtenus avec l'approche de [Mustière et Devogele, 2008] 199
Figure 113. Histogramme de l'auto-évaluation des liens d'appariement pour les réseaux routiers 202
Figure 114. Exemple de résultats différents obtenus avec notre approche et avec l'approche de [Mustière et Devogèle, 2008] 203
Figure 115. Sensibilité au seuil d'écart de position.....205
Figure 116. Sensibilité au seuil d'écart sémantique 206
Figure 117. Non-convergence de notre processus : état du processus après une première passe 207
Figure 118. Non-convergence de notre processus : état du processus après une deuxième passe.....208

Figure 119. Les étapes du processus d'appariement de données géographiques.....212
Figure 120. Interface générale d'appariement de données géographiques.....213
Figure 121. Interface pour le chargement des jeux de données. 214
Figure 122. Interface pour la saisie des spécifications 214
Figure 123. Exemple de type de correspondance 1 : n.....216
Figure 124. Interface pour la sélection des candidats à l'appariement.....217
Figure 125. Interface pour définir des critères d'appariement pertinents.....218
Figure 126. Exemple d'aide pour le critère d'écart de position.....219
Figure 127. Exemple d'appariement où un candidat a été choisi par deux objets.....220
Figure 128. Nouveau critère : analyse du contexte spatial et du vecteur d'erreur systématique 220
Figure 129. Exemple d'interface pour la détermination des seuils pour le critère d'écart de position.....222
Figure 130. Exemple d'interface pour l'étape de décision finale 227
Figure 131. Interface pour l'appariement interactif 228
Figure 132. Interface de contrôle des couples d'objets appariés, d'après [Sheeren, 2005]...229
Figure 133. Processus d'appariement de données géographiques proposé 233

Liste des annexes

Annexe 1. Aperçu des théories mathématiques pour modéliser les imperfections.....258
Annexe 2. Distances sémantiques entre différents concepts pour les points remarquables du relief.....262
Annexe 3. Distances sémantiques entre différents concepts pour les réseaux routiers.....267

Introduction

Introduction

Jusqu'à la fin des années 70, les cartes papier étaient le seul moyen de manipuler et de représenter des données ayant une composante spatiale. L'arrivée de l'informatique a accompagné le passage de l'information papier à l'information numérique, et a permis de répondre à des besoins croissants en cartographie, tels que la réalisation des cartographies automatiques, l'analyse des risques ou la navigation. Ainsi, on parle de données géographiques numériques, c'est-à-dire des données qui possèdent à la fois une localisation géographique et des informations attributaires. Cette information géographique est stockée sous la forme de bases de données géographiques.

De nos jours l'information géographique a pris une place importante dans la société, de nombreuses applications nécessitant des données géographiques. Ce dernier aspect, ainsi que la facilité actuelle d'acquérir des données géographiques, ont eu pour conséquence l'existence de nombreuses bases de données géographiques couvrant le même territoire du monde réel.

Une base de données géographiques est une abstraction du monde réel. Chaque abstraction du monde réel est dépendante des besoins, des points de vue des producteurs de bases de données ainsi que des points de vue des opérateurs de saisie.

Afin de minimiser les erreurs issues des points de vue de différents opérateurs, le passage du monde réel à une base de données géographiques est accompagné de spécifications décrivant le contenu théorique de la base de données géographiques, appelé également *le terrain nominal*.

De nos jours, il y a une multiplicité de données géographiques. Cette multiplicité est due, parmi d'autres facteurs, à la dérivation et à l'acquisition des données géographiques. Cette dernière peut être différente d'une base de données géographiques à une autre et peut se faire par des sources et des processus distincts. Ainsi, les processus et les sources qui fournissent des données sont la table à numériser, le levé terrain, le levé des points GPS, le processus de généralisation, les images satellitaires, les photos aériennes, la saisie par scanner.

Un phénomène du monde réel, que ce soit un objet localisé (par exemple une maison) ou un champ continu (par exemple l'altitude) peut être représenté dans une base de données géographiques en mode vecteur ou en mode raster. La Figure 1 illustre la représentation en mode vecteur du monde réel.

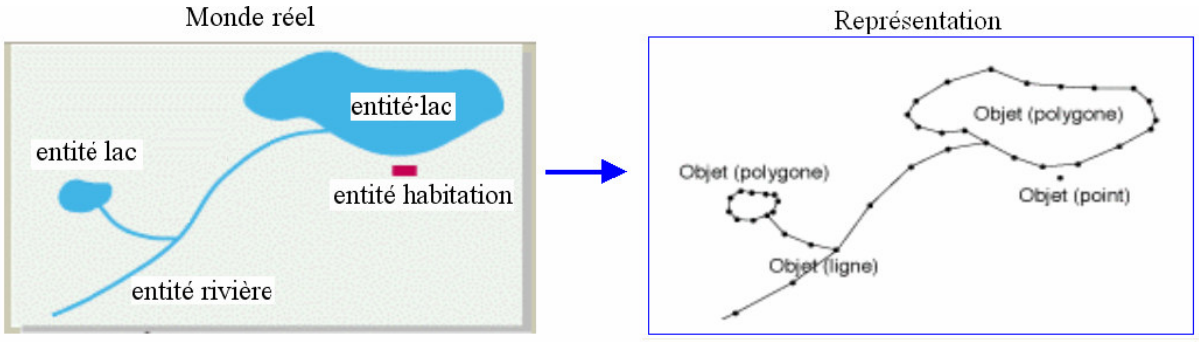


Figure 1. Modélisation des entités géographiques : monde réel et représentation vectorielle

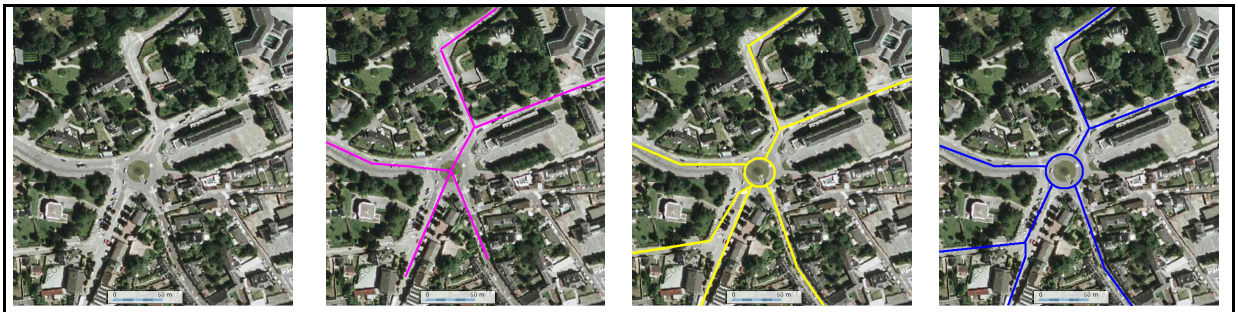
Un phénomène du monde réel représenté en mode vecteur possède une géométrie qui décrit sa localisation et implicitement sa forme, et des informations attributaires telles que le nom, la nature, la longueur, la superficie. D'une manière générale, la géométrie d'un objet géographique peut être décrite à travers trois primitives : le point défini par ses coordonnées (x, y), la ligne définie par un ensemble de points et de segments, et la surface, défini par son contour qui est une ligne fermée.

Dans la suite de ce mémoire de thèse, nous appelons *entité géographique* un phénomène ou objet du monde réel, représenté en mode vecteur par un objet géographique ou un ensemble d'objets géographiques dans la base de données géographiques. Le mot *objet géographique* désigne une instance de la base de données géographiques. Par exemple, une entité *route* est composée d'un ensemble d'objets « tronçon de route ».

Le monde réel peut être représenté à différents niveaux de détail, ce qui implique aussi l'existence de plusieurs bases de données géographiques couvrant le même territoire du monde réel. En fonction du niveau de détail de la base de données géographiques, une entité du monde réel peut être représentée par des géométries différentes. Par exemple, une rivière peut être modélisée par une géométrie linéaire ou bien par une géométrie surfacique ; un bâtiment peut être modélisé par un point représentant son centroïde ou par une surface. La base de données plus détaillée peut contenir par exemple une route moins importante qui n'est pas représentée dans une autre base de données moins détaillée, justement à cause de son importance moindre.

Suivant le niveau de détail, il existe des bases de données géographiques de résolution métrique ou de résolution décamétrique, par exemple. Plus la résolution de la base est fine, plus la saisie devient fastidieuse et longue et plus le volume de données devient important. En raison des contraintes de délais de production, les producteurs ne suivent pas toujours un processus idéal qui consisterait d'après [Ruas, 2002] « à saisir la base de données précise puis la généraliser pour produire des bases de données et des cartes ». Le plus souvent, les producteurs produisent d'abord la base de données la moins détaillée. Cet aspect génère une coexistence de plusieurs bases de données et même plus, une indépendance entre elles et entre les objets les composant, contrairement au processus idéal dans lequel il y aurait une source commune et des bases de données dérivées en fonction du besoin, ayant des liens avec cette source commune.

Nous illustrons en Figure 2 à titre d'exemple trois types de représentation différentes issues de la BDCARTO et de la BDTOPO de l'IGN (images b et c) et de MultiNet produite par TéléAtlas (image d), du même territoire du monde réel illustré sur l'image a (source : ©Géoportail). Nous remarquons que les bases BDTOPO et MultiNet ont des niveaux de détail comparables, tandis que la BDCARTO est moins détaillée que les deux autres.



a b c d

Figure 2. Le monde réel (image a) et trois représentations différentes issues de BDCARTO (image b), BDTOPO (image c) et MultiNet(image d).

Si nous abordons la situation du côté de la production, une autre source de différences entre les bases de données est le point de vue. En effet, deux bases de données peuvent répondre à des besoins différents, ou servir à des applications différentes. De ce fait, les points de vue peuvent être différents d'une base à une autre. Prenons par exemple les trois bases de données mentionnées ci-dessus : BDCARTO, BDTOPO et MultiNet. La BDCARTO a été produite par l'IGN pour faire de la cartographie à l'échelle de 1 : 100 000 ou 1 : 250 000 et pour réaliser des analyses à des niveaux régionaux et départementaux, ayant une résolution décamétrique. Elle a été également créée pour des études de réseaux, et elle s'intéresse donc à la manière dont les objets sont connectés. La BDTOPO, le référentiel à grande échelle, de précision métrique, prend en compte les détails topographiques et donne une position précise des objets. Elle est utilisée pour réaliser des cartes à l'échelle de 1 : 25 000 ou pour faire des études à moyenne et grande échelle. La base de données MultiNet, qui a une précision similaire à la BDTOPO est, quant à elle, destinée à la navigation routière. Elle s'intéresse surtout aux connexions entre les objets géographiques, étant plus riche au niveau attributaire.

Tous les facteurs que nous venons d'énumérer, liés à l'abstraction, à la représentation, aux contraintes de production et aux besoins des données géographiques, créent des différences de représentation entre les bases de données géographiques.

Ces différences entre les bases de données géographiques posent certains problèmes à la fois aux producteurs et aux utilisateurs.

En raison de l'absence de lien entre les différentes représentations, par exemple lorsque une mise à jour doit être faite, les producteurs doivent répéter les opérations de mise à jour dans toutes les bases de données géographiques, ce qui nécessite un coût supplémentaire d'une part pour la maintenance de leurs propres produits et d'autre part pour propager les mise à jours aux utilisateurs. L'existence de relations entre les différentes représentations pourrait faciliter la mise à jour, en saisissant le nouvel objet dans la base la plus détaillée par exemple, puis en propageant la mise à jour automatiquement dans les autres bases de données ainsi qu'aux utilisateurs [Badard, 2000 ; Kilpelaïnen, 2000].

Comme nous l'avons dit précédemment, le monde réel est très complexe et donc une représentation idéale de celui-ci est illusoire. Des erreurs de représentation, de géométrie, de connexions ou des erreurs au niveau de l'information attributaire par rapport au terrain nominal sont présentes dans les bases de données. Une étude de la qualité de chaque base de données s'impose afin de détecter et de corriger les éventuelles incohérences. Or, cette étape

pourrait également être facilitée si les relations entre les différentes représentations existaient [Egenhofer *et al.,* 1994 ; Sheeren *et al.*, 2008].

A leur tour, à cause de cette indépendance, les utilisateurs peuvent rencontrer des difficultés pour étudier plusieurs zones adjacentes ou pour faire des analyses multi-niveaux. La représentation multiple permettrait à l'utilisateur de passer très facilement d'un niveau de détail à un autre afin d'avoir la meilleure information qui répond à son besoin [Devogèle, 1997 ; Mustière et van Smaalen, 2007]. Un exemple typique d'analyse multi-niveaux est la navigation en utilisant les systèmes embarqués GPS, qui utilisent par exemple la représentation la moins détaillée si on est en milieu rural et la plus détaillée si on est en milieu urbain.

En raison de tous ces besoins liés à la mise à jour des données, à la détection des incohérences, au recalage des données, à l'étude de zones adjacentes ou à l'analyse multi-niveaux, de nombreux travaux de recherche ont été menés et ont permis d'améliorer le processus de mise en correspondance des différentes bases de données. Plus concrètement, définir les relations entre les différentes représentations consiste à expliciter les relations entre les objets homologues des bases de données géographiques existantes. Ce processus est connu sous le nom d'appariement de données (voir Figure 3). L'appariement de données est donc un outil qui permet de mettre en correspondance des objets homologues, c'est-à-dire des objets qui représentent la même réalité [Walter et Fritsch, 1999].

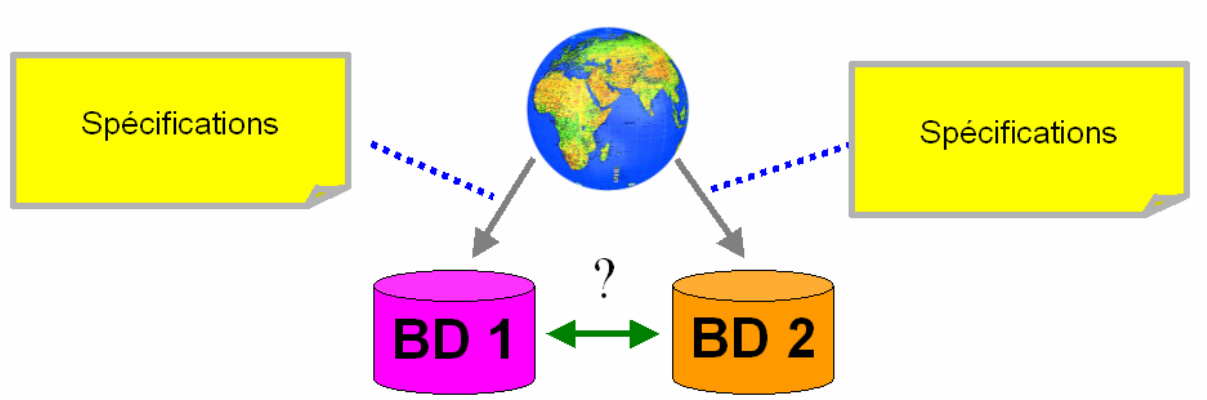


Figure 3. Contexte de l'appariement de données

Cette thèse est concentrée sur la problématique de l'appariement de données vecteur.

L'appariement de données géographiques est loin d'être immédiat. Nombreuses sont les causes qui font que l'appariement de données géographiques est une tâche difficile. Par exemple, les objets géographiques ne possèdent pas toujours d'identifiant, le niveau d'abstraction peut différer d'une base à une autre ou encore même si deux jeux de données ont le même niveau de détail, deux objets géographiques représentant le même objet du monde réel n'ont pas les mêmes coordonnées géographiques à cause des imperfections sur la localisation.

Dans ce contexte, l'objectif de la thèse est de proposer un processus d'appariement de données géographiques *flexible*, *adaptable* à toutes les données, que ce soit des données représentées par des points, des lignes ou des surfaces, ou des données provenant de bases de données ayant des niveaux de détails différents, et qui soit efficace même si les données ne

sont pas idéales, c'est-à-dire qui prenne en compte les erreurs au niveau de la géométrie, des informations attributaires ou des relations spatiales.

Notre approche s'appuie sur plusieurs idées. Premièrement, nous considérons que le processus d'appariement doit être basé sur les différentes caractéristiques des objets géographiques telles que la géométrie, l'information attributaire et sémantique ou les relations spatiales. En conséquence il sera adaptable en fonction des informations présentes dans les données. Deuxièmement, le processus d'appariement doit être guidé par des connaissances. Troisièmement, nous considérons que toutes les imperfections doivent être prises en compte autant à l'intérieur du processus qu'au niveau de la décision. Par conséquent il sera moins sensible aux éventuelles erreurs.

Ainsi, nous avons défini une taxonomie de l'imperfection (imprécision, incertitude et incomplétude) et nous avons modélisé l'imperfection dans les données et dans les connaissances. Pour y parvenir, nous avons choisi la théorie des fonctions de croyance [Shafer, 1976]. Elle sera à la base de notre nouvelle approche d'appariement de données géographiques.

Le processus que nous proposons est basé sur des critères d'appariement, chaque critère d'appariement s'appuyant sur une caractéristique des données géographiques : la localisation, l'orientation, le voisinage, le nom, la sémantique, etc.

Organisation du mémoire de thèse

Le mémoire de thèse est composé de cinq chapitres.

Le chapitre A contient un état de l'art sur l'appariement de données en fonction des besoins. Nous avons distingué quatre besoins : l'évaluation de la qualité, le recalage, la mise à jour et l'intégration des données géographiques. L'intérêt de ce chapitre est de donner une vision globale des caractéristiques de l'appariement ainsi que de situer notre travail par rapport aux travaux existants sur l'appariement de données. Après avoir défini la problématique, nous exposons les objectifs de ce travail de thèse.

Le but du chapitre B est de présenter les types d'imperfection présents dans les données et dans les connaissances issues de divers travaux dans le domaine de l'intelligence artificielle, et d'exposer la théorie des fonctions de croyance, qui fait partie de la famille des théories de l'incertain. Après avoir justifié le choix de la théorie des fonctions de croyance à travers quelques besoins et exemples, nous donnons un aperçu des applications de la théorie des fonctions de croyance. Enfin, nous présentons le contexte mathématique de la théorie des fonctions de croyance.

Le chapitre C représente le cœur de ce mémoire de thèse, car il contient la description détaillée de notre approche d'appariement de données. Nous présentons d'abord les étapes de notre processus qui sont au nombre de cinq : sélection des candidats, initialisation des masses de croyance, fusion des critères par candidat, fusion des candidats et décision. Ensuite, nous détaillons la modélisation des critères d'appariement typiques qui peuvent être utilisés dans un processus d'appariement.

Le chapitre D présente l'expérimentation et l'évaluation de notre nouvelle approche d'appariement. Nous décrivons d'abord le modèle de données que nous avons défini et le prototype du processus qui a été conçu sur la plate-forme GeOxygene. Ensuite, nous présentons les deux études que nous avons mises en place afin de valider notre approche

basée sur la théorie des fonctions de croyance. Premièrement, nous montrons les résultats que nous avons obtenus en testant notre approche sur cinq jeux de données ayant des niveaux de détail différents et représentant les points remarquables du relief. Les jeux de données couvrent cinq départements français et ils sont issus de la BDCARTO et de la BDTOPO de l'IGN. Deuxièmement, nous présentons les résultats obtenus en effectuant des tests sur des jeux de données à différents niveaux de détail représentant les réseaux routiers, et issus de la BDCARTO de l'IGN et de MultiNet de TéléAtlas. Les résultats sont ensuite évalués en termes de précision et de rappel.

Le chapitre E est consacré à la présentation d'une proposition d'un processus complet d'appariement de données géographiques destiné à l'utilisateur. Le processus décrit toutes les étapes nécessaires pour apparier deux jeux de données, en commençant avec le chargement des jeux de données et en finissant avec la qualification et la sauvegarde des résultats d'appariement. Des améliorations à notre processus décrites dans le chapitre C sont également apportées dans le chapitre E.

Enfin, nous concluons en résumant notre approche d'appariement de données en présentant les apports de la thèse, et nous donnons également des perspectives de travail.

CHAPITRE A
Appariement de données géographiques

A Appariement de données géographiques

De nombreuses approches d'appariement de données géographiques existent dans la littérature. Dans ce chapitre nous présentons d'abord un état de l'art sur l'appariement de données géographiques afin de nous positionner par rapport aux travaux existants. Les approches sont présentées d'une part en fonction du besoin auquel elles répondent et d'autre part en fonction des données géographiques traitées. Ensuite, nous présentons quelques aspects liés à l'imperfection des données géographiques et à la nécessité de la prise en compte de cette imperfection d'une manière explicite à l'intérieur du processus d'appariement de données géographiques. Enfin, nous concluons et nous présentons notre sujet de recherche.

A.1 L'appariement, un problème complexe

Le processus d'appariement de données géographiques est un outil qui permet de mettre en correspondance des objets homologues, c'est-à-dire des objets qui représentent la même réalité [Walter et Fritsch, 1999]. Il s'appuie sur la notion de ressemblance, c'est-à-dire que deux objets géographiques A et B appartenant à des bases de données géographiques différentes sont appariés s'ils se ressemblent. Il consiste donc à mettre en valeur des ressemblances de lieu, de nature, de relation spatiale ou de forme.

A première vue, la ressemblance semble à être facile à définir à travers des règles basées sur les différentes propriétés des objets géographiques. Par exemple :

- lieu : « les objets A et B se ressemblent s'ils sont positionnés sur le même lieu »,
- forme : « les objets A et B se ressemblent s'ils ont la même forme »,
- type/nature : « les objets A et B se ressemblent s'ils ont le même type/nature »,
- relations topologiques : « les objets A et B se ressemblent s'ils ont les mêmes relations topologiques avec leurs voisins ».

Cependant, la spécificité des données géographiques fait que la notion de ressemblance est très complexe à définir dans la pratique et qu'elle n'est pas immédiate puisqu'elle s'appuie sur des connaissances qui ne sont pas toujours cohérentes entre elles.

Des objets géographiques se ressemblent s'ils sont positionnés sur le même lieu. Cependant, dans la réalité il existe de différentes simplifications géométriques de ce lieu, plus ou moins précises. Une entité du monde réel, peut être représentée dans une base de données par une ligne, qui modélise son axe principal ou même par un point, qui modélise son centre. De la même manière, il est possible qu'un bâtiment soit représenté par une surface signifiant sa forme ou un point signifiant son centroïde. Prenons un exemple : supposons que nous voulons représenter une crête. La localisation de la crête dans la base de données peut ne pas refléter parfaitement la réalité et de plus, la précision de la localisation peut être différente. Par exemple elle est plus précise lorsque l'entité est représentée par une ligne, que lorsqu'elle est représentée par un point.

De plus, le degré de cohérence entre les données géographiques de différentes bases de données est lié à la notion de niveau de détail de chaque base de données. Ainsi, plus les bases de données à apparier ont des niveaux de détail différents, plus le degré de cohérence est faible. Les règles de ressemblance peuvent être différentes en fonction des niveaux de détail

des bases de données à apparier. Globalement, nous pouvons distinguer trois types de relation entre les jeux de données à apparier (voir Figure 4) :

- chevauchement : les deux bases de données ont le même niveau de détail et le même contenu [Badard, 2000],
- inclusion : les deux bases de données ont des niveaux de détail différents, et la base de données moins détaillée est incluse dans la base de données plus détaillée [Devogèle, 1997 ; Mustière et Devogele, 2008],
- chevauchement partiel : les deux bases de données sont sensiblement différentes en contenu et en niveau de détail [Olteanu-Raimond, 2008].

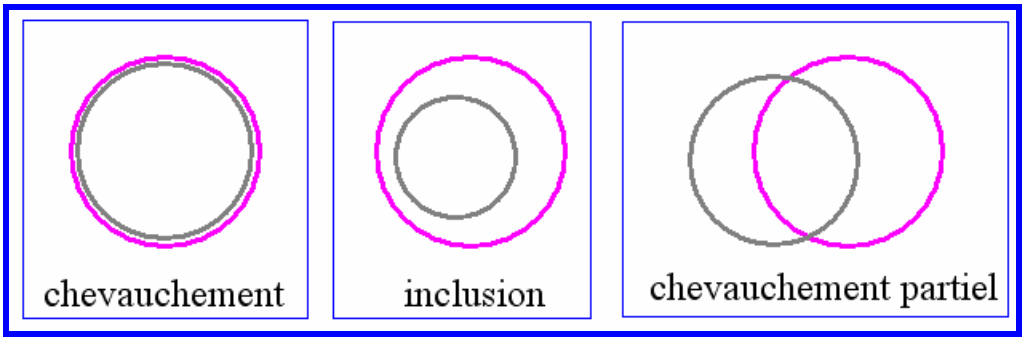


Figure 4. Types de relation entre les jeux de données à apparier

Le premier type de relation est rencontré dans le contexte de la mise à jour ou dans le cas du contrôle de la qualité d'une base de données, tandis que les deux derniers types de relation sont plutôt spécifiques au contexte de l'intégration de base de données ou de la mise à jour de bases de données multi-représentations.

Une autre problématique qui rend l'appariement de données géographiques complexe est liée à la notion de cardinalité des liens d'appariement, c'est-à-dire le nombre d'objets en correspondance. Dans le processus d'appariement, la cardinalité des liens d'appariement peut être différente, d'une part en raison des niveaux de détail différents des bases de données à apparier (voir la Figure 5 à gauche), et d'autre part en raison des règles de découpage du monde réel (voir la Figure 5 à droite). Par exemple, les routes peuvent être découpées en tronçons de route selon le nombre de voies dans une base de données et selon le revêtement dans la seconde base, ou encore les rivières peuvent être découpées en tronçons de rivière selon leur largeur ou leur débit.

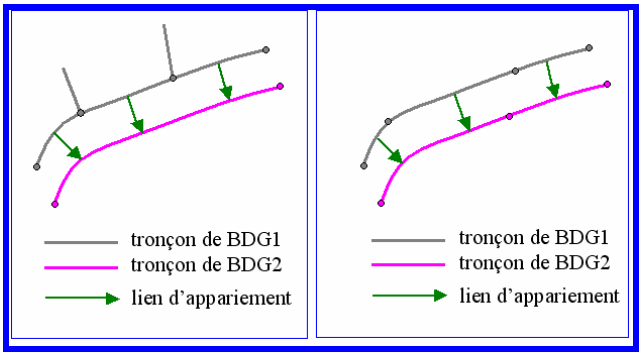


Figure 5. Exemples de liens d'appariement de cardinalités différentes en raison du niveau de détail différent 1 – n, (à gauche) et du découpage du monde réel différent n – m, (à droite)

Ainsi, en fonction des bases de données à apparier, la cardinalité des liens d'appariement peut être : 1 : 1 (un objet d'une base de données est mis en correspondance avec un seul objet dans l'autre base de données), 1 : n (un objet d'une base de données est mis en correspondance avec n objets dans l'autre base de données) ou n : m (n objets d'une base de données sont mis en correspondance avec m objets dans l'autre base de données), comme le montre la Figure 6.

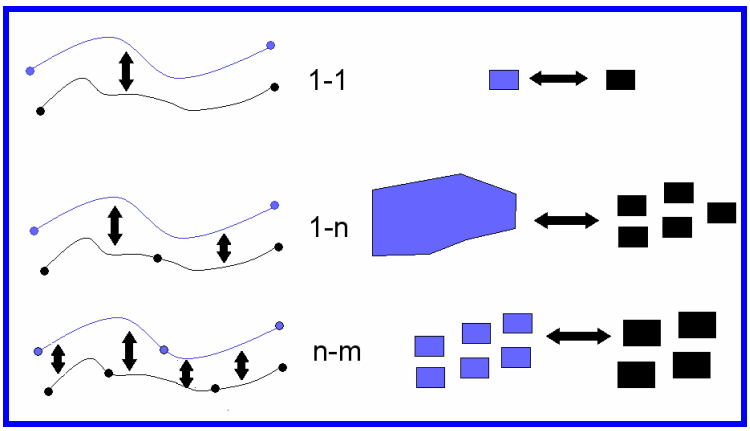


Figure 6. Cardinalité des liens d'appariement

L'appariement 1 : 1 apparaît surtout lorsque les bases de données à apparier ont des points de vue proches ou des niveaux de détail proches, tandis que l'appariement 1 : n est dû à la différence de segmentation entre les deux bases (c'est-à-dire que d'une base à l'autre les routes ne sont pas découpées en tronçons de route de la même façon) et des niveaux de détail différents. La différence de segmentation et les différents points de vue sont la cause de l'appariement n : m.

A.2 L'appariement, un outil pour répondre à plusieurs besoins

L'appariement est utilisé dans de nombreuses applications manipulant les données géographiques. Dans cette sous-partie les approches sont présentées en fonction du besoin : l'évaluation des données géographiques, le recalage, la mise à jour et l'intégration.

A.2.1 Apparier pour évaluer la qualité des données géographiques

La production d'une base de données géographiques est un long processus qui consiste en un enchaînement de nombreuses étapes et qui nécessite l'intervention de plusieurs acteurs. La complexité de l'information géographique et sa chaîne de production (acquisition d'une source de données, modélisation des données, discrétisation des objets géographiques) sont à l'origine d'erreurs et d'imprécisions.

Actuellement, de plus en plus d'utilisateurs exploitent l'information géographique et prennent des décisions dans différentes applications telles que la navigation, l'aménagement de territoire ou les études de risques. De ce fait, la qualité des données géographiques est un enjeu important à la fois pour les utilisateurs et pour les producteurs de données. Elle suscite depuis plus de dix ans un vif intérêt qui se manifeste à travers des conférences, et des ouvrages dédiés à ce sujet ainsi que des nombreux articles [Guptill et Morrison, 1995 ; David et Fasquel, 1997 ; Dassonville *et al.*, 2002].

La qualité des données concerne à la fois le producteur et l'utilisateur, comme le montre la Figure 7. Du point de vue de l'utilisateur, le concept de qualité implique l'adéquation des données à ses besoins, et il est connu dans la littérature sous le nom de *qualité externe* [Devillers, 2004 ; Devillers et Jeansoulin, 2005]. Du point de vue du producteur, évaluer la qualité d'une base de données géographiques consiste à vérifier la conformité des données géographiques aux spécifications : il s'agit de la *qualité interne* [Devillers et Jeansoulin, 2005].

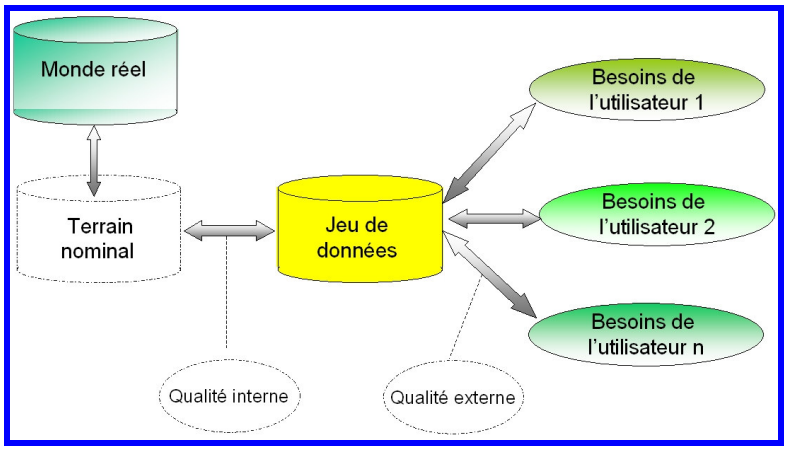


Figure 7. Concept de qualité d'une base de données géographiques

Vu son importance, la qualité interne, que nous appelons par abus de langage « qualité », a fait l'objet de nombreux travaux de recherche qui font intervenir l'appariement et que nous présentons ci-après.

L'étude de la qualité consiste à contrôler la qualité géométrique [Vauglin, 1997 ; Goodchild, 2005], la sémantique et l'exhaustivité d'un jeu de données géographiques par rapport à un autre jeu de données géographiques représentant le monde réel, appelé jeu de référence. Afin de réaliser ce contrôle, une première étape est l'appariement des données, c'est-à-dire la mise en correspondance des objets issus de la base de référence avec leurs homologues dans la base de données à analyser.

Dès lors que les données sont appariées, on s'intéresse par exemple à évaluer la qualité de leur forme et de leur position en s'appuyant sur différentes mesures telles que la distance de Hausdorff ou la distance surfacique [Bel Hadj Ali, 2001].

La spécificité des données dans le cadre de l'évaluation de la qualité réside dans le fait qu'elles sont relativement proches tant au niveau géométrique qu'au niveau sémantique. L'appariement est donc relativement simple et il est basé sur des mesures de proximité [Bel Hadj Ali, 2001 ; Sheeren, 2005]. De nombreux travaux liés à l'étude de la qualité des données géographiques ont d'ailleurs été réalisés sans aborder le sujet de l'appariement de données, ce dernier étant supposé évident [Hunter et Goodchild, 1993 ; Egenhofer *et al.*, 1994].

Afin d'assurer la cohérence de deux bases de données géographiques lors de leur intégration, [Sheeren, 2005] s'intéresse à l'évaluation de la cohérence de représentation des objets appariés en se basant sur un processus d'apprentissage automatique.

L'appariement de données peut également être utilisé pour évaluer le processus de généralisation ou les données généralisées [Ruas, 2001]. La généralisation des données géographiques a fait l'objet de nombreux travaux de recherche [Ruas, 1999 ; Mustière, 2001 ;

Duchêne, 2004]. La généralisation est « un processus de transformation de données géographiques utilisé pour produire des données qui répondent à des besoins précis » [Ruas, 1999]. Après le processus automatique de généralisation, des questions liées à l'adéquation des données généralisées aux besoins se posent [Bard *et al.,* 2003 ; Bard, 2004].

A.2.2 Apparier pour recaler des données géographiques

L'appariement de données géographiques est également un outil pour recaler des données géographiques sur un référentiel dans le but d'améliorer la qualité géométrique des données. Le recalage de données est un processus qui consiste à superposer deux jeux de données représentant potentiellement la même réalité sur une même zone géographique. Il s'agit de réaliser une transformation géométrique à partir de points identifiés comme homologues par appariement. Le recalage de données concerne les données vectorielles et les données raster (image). Ainsi la Figure 8 montre un exemple de recalage vecteur-vecteur et la Figure 9 montre un exemple de recalage vecteur-raster.

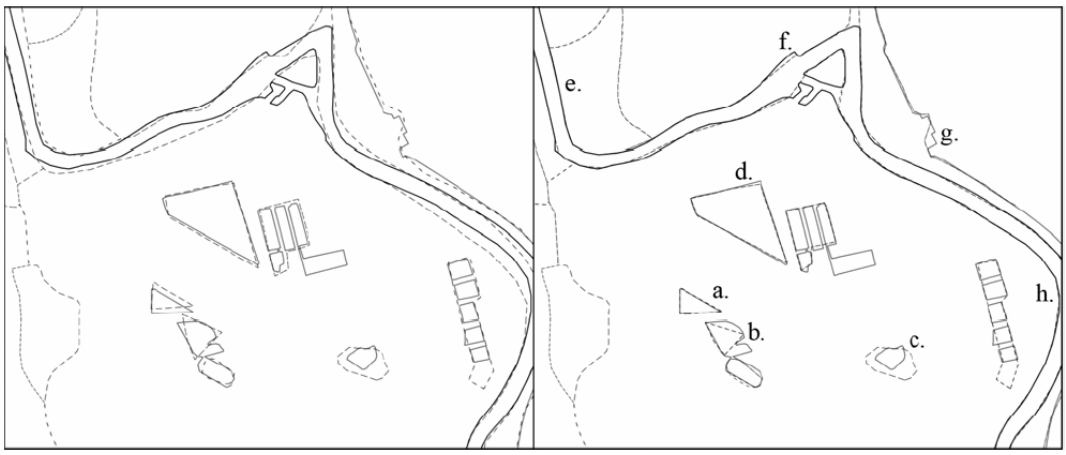


Figure 8. Recalage de deux jeux de données vecteur-vecteur (jeux de données, initiaux à gauche et recalés à droite)

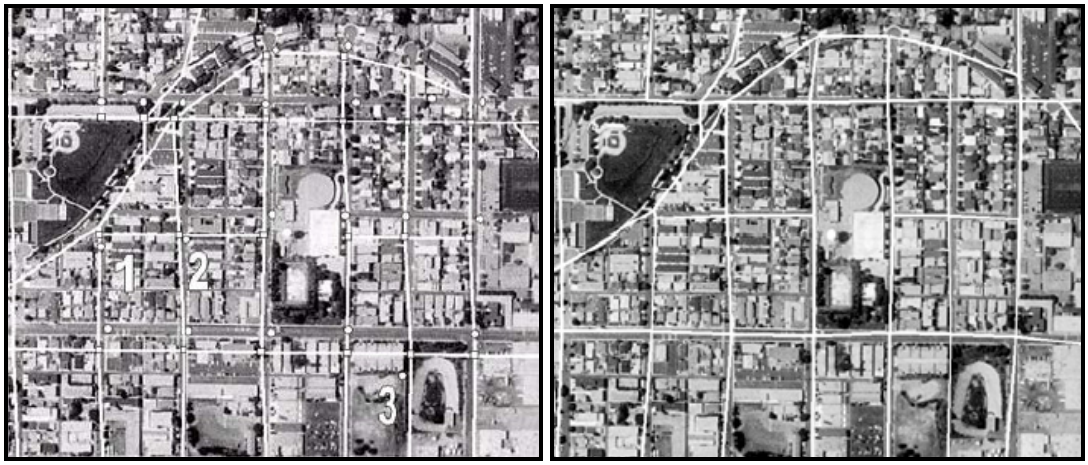


Figure 9. Recalage de deux jeux de données vecteur-raster (jeux de données, initiaux à gauche et recalés à droite)

[Chen *et al.*, 2006] proposent une approche d'appariement vecteur-raster, afin de recaler un réseau routier et une orthophoto. L'approche de mise en correspondance des points homologues, nommés également points de contrôle, est la suivante. La première étape consiste à détecter des points de contrôle dans les données vecteur, c'est-à-dire des points représentant l'intersection de plus de deux tronçons de route. La deuxième étape consiste à trouver les points homologues des points de contrôle détectés lors de la première étape. Pour cela, ils détectent d'abord les intersections dans l'image en utilisant des outils de traitement d'image ainsi que des connaissances issues des données vecteur, telles que les directions des segments ou la largeur des tronçons de route. Une fois que les points de contrôle dans les données vecteur et dans l'image ont été identifiés, l'appariement des points homologues est réalisé en utilisant les coordonnées géographiques des données vecteur et celles de l'image géoréférencée. Un filtrage des points de contrôle est ensuite appliqué pour éliminer les mauvaises détections ou les détections incertaines. Les filtres détectent les couples de points de contrôle qui ont un vecteur d'écart entre les points de contrôle très différent par rapport aux couples voisins, dans ce cas il s'agit d'une erreur systématique. Enfin, à partir des couples de points de contrôle, le recalage de données est réalisé par une transformation élastique [Saalfeld, 1988].

Afin de recaler des cartes scannées sur des orthoimages, [Chen *et al.*, 2008] proposent une approche similaire que celle que nous venons de détailler, à la différence près qu'ils utilisent des points de contrôle extraits automatiquement d'une base de données vecteur.

[Clodoveu *et al.*, 2007] s'intéressent aux recalages des « adresses » sur une base d'adresses ayant une référence spatiale. Le concept d'adresses dans cette approche correspond à la fois aux adresses postales et à d'autres informations qui font référence à un lieu telles que le nom du bâtiment, le code postal, le code téléphonique, etc. Une fois que les adresses sont structurées grâce à un traducteur, c'est-à-dire que chaque adresse est un n-uplet contenant des attributs tels que le nom et le type de la route, le numéro du bâtiment, le code postal, le nom de la ville, etc., l'appariement de données consiste à trouver les objets appartenant à la base de données d'adresses qui correspondent aux données fournies par le traducteur. L'appariement recherché est de cardinalité 1 : 1. Afin de trouver le couple d'objets homologues, [Clodoveu *et al.*, 2007] comparent les attributs de chaque n-uplet avec les attributs des objets de la base de données postale en utilisant des mesures de similarité entre les chaînes de caractères telles que la distance de Levenshtein [Levenshtein, 1965] et l'algorithme *Shift-And* [Zobel et Dart, 1995]. L'évaluation des résultats est automatique ; elle se base sur un indicateur qui peut prendre des valeurs entre 0 et 1 (0 signifiant complètement incertain et 1 signifiant complètement sûr). Cet indicateur final est calculé à partir des indicateurs déterminés pendant l'étape d'appariement (l'indicateur obtenu à partir du nombre d'attributs appariés pondérés par leur importance) et de localisation (l'indicateur obtenu à partir de la géométrie des objets : point, ligne, surface).

Dans le but de réaliser une carte de données géoréférencées des forêts anciennes, [Dupouey *et al.*, 2007] s'intéressent au recalage des cartes anciennes afin de pouvoir les représenter dans un référentiel géographique. Le processus de recalage s'appuie sur des points de contrôle représentant des objets remarquables tels que les églises, les châteaux, les carrefours, les moulins, etc. Ces points de contrôle, choisis à la fois sur les cartes anciennes et sur la carte de référence actuelle, sont ensuite mis en correspondance manuellement. A partir de ces points de contrôle, plusieurs transformations mathématiques (translation, rotation, homothétie, etc.) sont appliquées à la carte d'origine afin de pouvoir la superposer sur la carte actuelle. Enfin, une étape de raccordement des contours des forêts aux limites entre deux cartes est réalisée.

D'autres travaux de recalage utilisant l'appariement de données ont été réalisés [Besl et McKay, 1992 ; Haunert, 2005; Doytsher *et al.*, 2001 ; Gösseln et Sester, 2003]. Le processus d'appariement étant un outil développé à partir des approches décrites auparavant, ces travaux ne sont pas détaillés davantage dans ce mémoire de thèse.

A.2.3 Apparier pour mettre à jour les données géographiques

Le processus de mise à jour concerne à la fois les producteurs et les utilisateurs. Les producteurs doivent d'une part détecter les évolutions et les intégrer dans leur bases de données, appelées bases de données de référence, et d'autre part les propager aux utilisateurs en leur fournissant des outils capables de réaliser une mise à jour, la plus automatique possible, tout en leur permettant de garder les modifications qu'ils ont faites et qui ont conduit aux bases de données dérivées.

Un processus complet de mise à jour est illustré sur la Figure 10. Nous montrons le processus de mise à jour en fonction du temps : du côté du producteur (en rose) et du côté de l'utilisateur (en bleu).

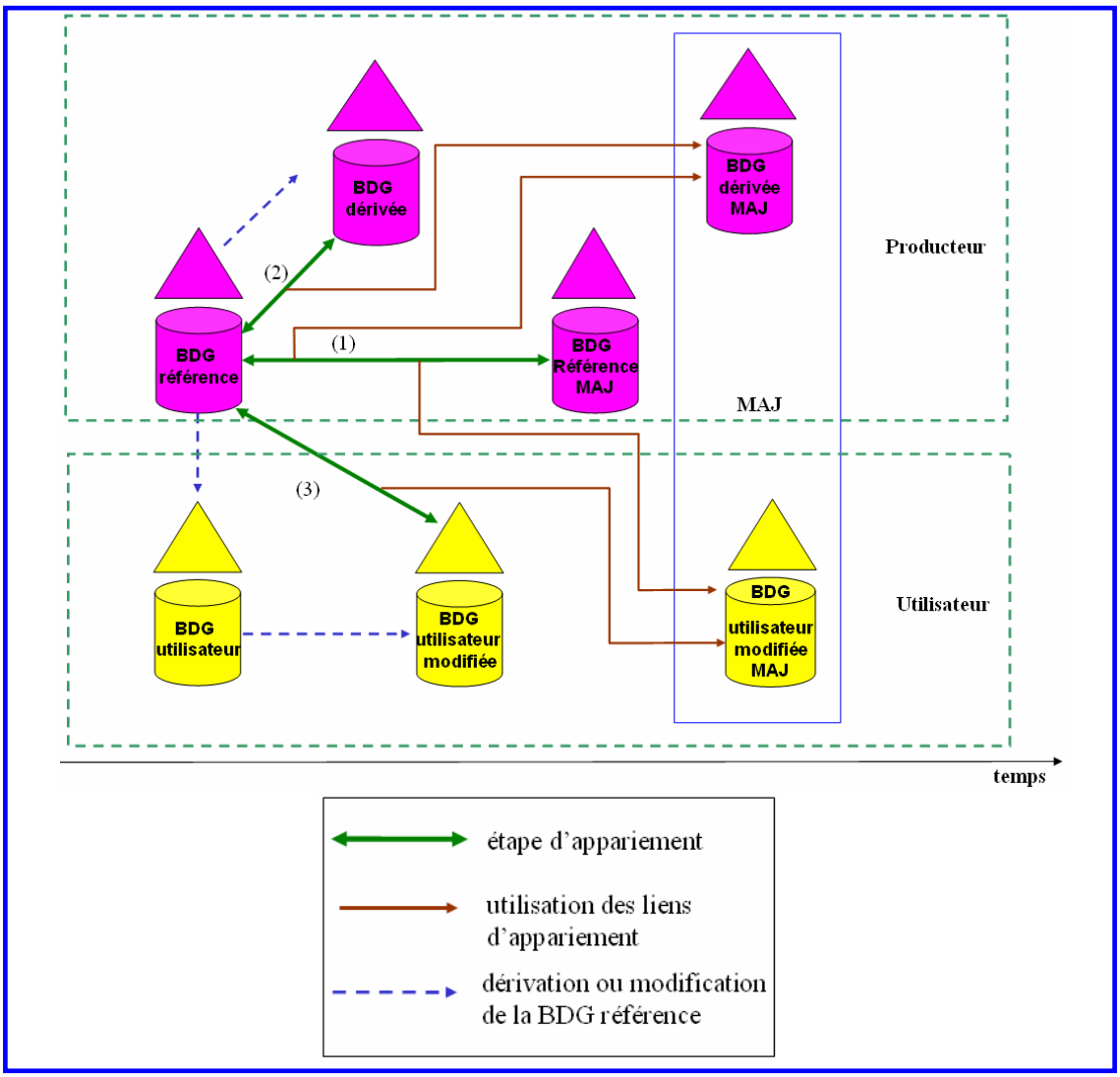


Figure 10. Processus général de mise à jour

A partir d'une base de données de référence (BDG référence), plusieurs bases de données peuvent être dérivées : les bases de données dérivées du côté du producteur (BDG dérivée) ou les bases de données utilisateur du côté de l'utilisateur (BDG utilisateur). Les bases de données dérivées peuvent être réalisées à travers des processus, par exemple la généralisation. Les bases de données utilisateur peuvent être toute la base de référence ou uniquement un extrait de celle-ci, et peuvent être obtenues par un transfert.

Supposons que la BDG de référence soit mise à jour. Soit BDG référence MAJ la nouvelle version de la base de référence. La problématique qui se pose est : comment mettre à jour les bases de données dérivées chez le producteur et les bases de données utilisateur chez les utilisateurs ?

Afin de répondre à cette problématique, il existe plusieurs approches. Par exemple, selon [Badard, 2000] la mise à jour des bases de données dérivées chez le producteur nécessite d'effectuer les étapes (1) et (2), et la mise à jour des bases de données utilisateur nécessite d'effectuer les étapes (1) et (3). Ces trois étapes sont illustrées sur la Figure 10 par des doubles flèches vertes. L'étape (1) consiste à apparier la base de données de référence (BDG référence) et la base de données de référence mise à jour (BDG référence MAJ), dans le but d'extraire les évolutions. L'étape (2) consiste à apparier la base de données dérivée (BDG dérivée) et BDG référence, et l'étape (3) à apparier la base de données utilisateur modifiée (BDG utilisateur modifiée) et BDG référence afin d'extraire les relations implicites entre les données de référence et les données dérivées ou rajoutées.

[Uitermark *et al.*, 1998], quant à eux, estime que seule l'étape (1) est nécessaire pour mettre à jour la base de données utilisateur ou dérivée.

Pour identifier les mises à jour faites sur une base, il est nécessaire de détecter les évolutions entre deux actualités d'une même base de données géographiques, lorsqu'il n'y a pas de gestion temporelle des identifiants. L'appariement de données est alors un outil qui permet de détecter les éléments homologues entre deux actualités. La cardinalité des liens d'appariement est interprétée différemment dans le contexte de la mise à jour et dans les autres contextes. Ainsi, lorsque deux bases de données BDG_0 à l'instant T_0 et BDG_1 à l'instant T_1 sont évaluées dans le but de la mise à jour, la cardinalité est interprétée de la façon suivante :

- 1 : 1 : modification possible de l'objet de la base BDG_0, des tests d'égalité géométrique et attributaire sont faits,
- 1 : 0 : l'objet de la base BDG_0 a été supprimé,
- 1 : n : l'objet de la base BDG_0 a été modifié, il a subi une scission
- n : m : modifications dans les deux bases,
- n : 1 : il y a eu une fusion d'objets de la base BDG_0,
- 0 :1 : un objet a été créé dans la base BDG_1.

Les deux bases de données géographiques, celle qui contient les données à jour et celle qui contient les données à mettre à jour, sont alors très comparables. En effet, elles présentent le même niveau de détail, elles ont une hétérogénéité réduite ou nulle au niveau de la représentation des objets de la base, de la sémantique et de la géométrie, et enfin les objets qui n'ont pas subi de changements ont gardé leur position.

Ce dernier aspect est exploité dans le processus d'appariement. On détecte, en utilisant la géométrie et la topologie, d'abord les objets homologues qui n'ont pas changé, puis les objets qui ont changé [Badard, 1998]. L'interprétation des évolutions entre les deux bases de données géographiques est réalisée en utilisant la cardinalité des liens.

Ainsi, le fait que l'appariement de données utilisé dans le contexte de la mise à jour ne nécessite pas d'outils très complexes [Bouziani, 2003 ; Lemarié, 1997] fait qu'on s'intéresse plutôt à la rapidité/complexité du processus [Gomboši *et al.,* 2003]. Ce dernier propose une approche qui recherche les segments exactement identiques, puis ceux qui restent sont qualifiés d'issus d'une fusion, d'une scission ou d'une création.

Un autre cas de figure souvent rencontré consiste à mettre à jour une base de données à partir d'une autre base de données différente. Il existe des producteurs de données, comme l'IGN par exemple, qui possèdent plusieurs bases de données à des niveaux de détails différents. Pour faciliter la mise à jour de leurs bases et ainsi réduire les coûts, un processus de mise à jour de la base de données la plus détaillée est mis en oeuvre, puis les évolutions sont propagées automatiquement vers les autres bases de données. Ainsi, pour pouvoir propager les mises à jour, un appariement entre des bases de données multi-représentation doit être réalisé. Dans ce cas, l'appariement de données devient plus complexe en raison de l'hétérogénéité des bases de données.

A.2.4 Intégration de bases de données géographiques hétérogènes

L'intégration de bases de données géographiques a comme objectif d'unifier la sémantique de deux ou plusieurs bases de données, d'éliminer les objets redondants et de regrouper les objets similaires [Devogele *et al.*, 1998]. Le processus d'intégration de bases de données géographiques se décompose en trois étapes : pré-intégration, mise en correspondance des schémas et des données et intégration des schémas et des données (voir Figure 11).

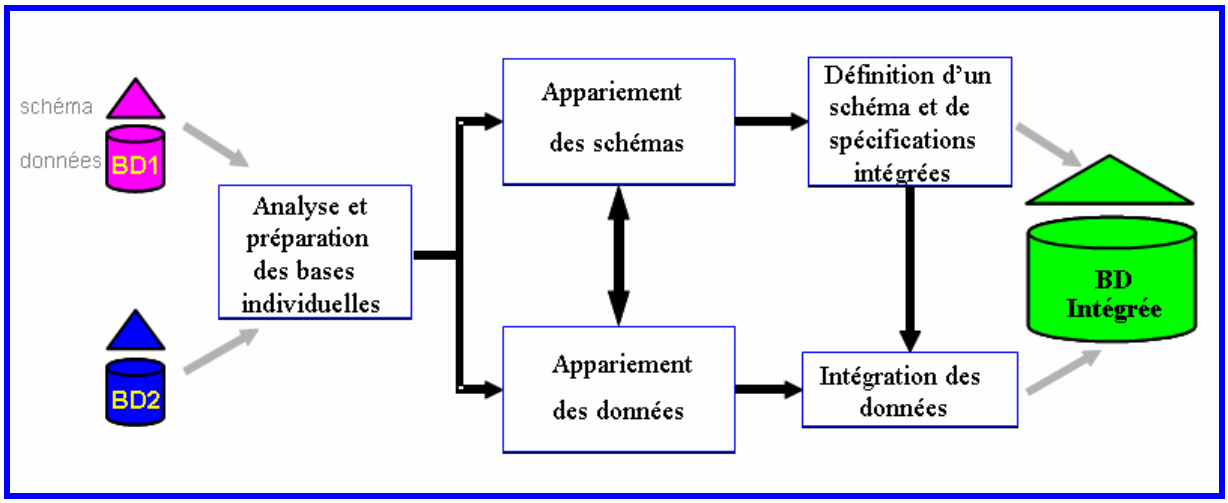


Figure 11. Processus d'intégration de bases de données géographiques

A.2.4.1 Pré-intégration

L'étape d'analyse et de préparation des bases individuelles, dite aussi de pré-intégration, a comme objectif de préparer les bases de données afin qu'elles soient intégrées le plus automatiquement possible [Devogele, 1997 ; Sheeren, 2005]. Elle consiste à analyser chaque base individuellement, à les transformer dans des schémas plus proches au moyen de règles de normalisation, à expliciter l'information implicite et à transformer les géométries dans une projection commune. L'enrichissement des schémas est une sous-étape primordiale lorsque l'information implicite est présente dans une base de données et qu'elle s'appuie sur les spécifications.

Par exemple, sur la Figure 12 nous pouvons remarquer la représentation d'une patte d'oie. Cependant, cette représentation est implicite, la patte d'oie étant composée de tronçons de route. Afin d'expliciter cette information, une solution est de créer une nouvelle classe appelée par exemple *Patte d'oie*. Chaque objet, ou instance de la classe, aura une géométrie et des attributs.

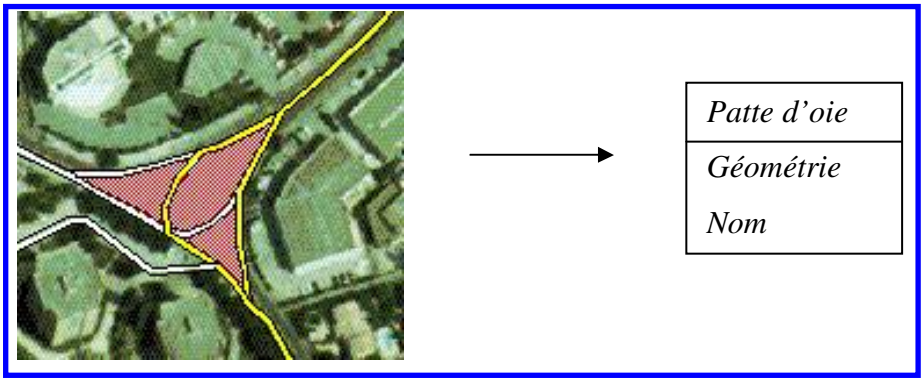


Figure 12. Explicitation de l'information implicite : cas d'une patte d'oie

A.2.4.2 Appariement des schémas et des données

La deuxième étape consiste à définir les correspondances entre les schémas, processus appelé « appariement des schémas » et entre les données, processus appelé « appariement des données ». Les deux processus ne sont pas complètement indépendants, en principe ils interagissent entre eux. Ainsi, l'appariement des schémas peut s'appuyer sur les données et réciproquement l'appariement des données peut s'appuyer sur les schémas.

Concernant l'appariement des schémas, plusieurs techniques existent dans la littérature. Il y a d'une part les techniques dites simples qui s'appuient sur des informations issues directement des schémas, telles que le nom d'une classe, les attributs, le type de l'objet, les relations entre les schémas (par exemple «est un », « est composé de ») [Madhavan *et al.*, 2001 ; Rahm et Bernstein, 2001 ; Do *et al.*, 2002], et d'autre part les techniques fondées sur la déclaration d'Assertions de Correspondance Inter-schémas (ACI), définies initialement pour les bases de données classiques [Parent et Spaccapietra, 1996 ; Devogele, 1997 ; Sheeren *et al.*, 2008]. Une approche d'appariement de schémas de plus en plus privilégiée s'appuie sur une ontologie[1] de domaine [Fonseca *et al.*, 2002 ; Comber *et al.*, 2004 ; Rodriguez et

[1] D'après [Gruber, 1993], une ontologie est la spécification d'une conceptualisation d'un domaine de connaissance.

Egenhofer, 2004 ; Gesbert, 2005 ; Abadie *et al.*, 2006 ; Mostafavi, 2006 ; Mustière *et al.*, 2007].

Comme nous pouvons le constater en Figure 11, dans le contexte de l'intégration des bases de données géographiques, l'appariement de données est utilisé pour apparier des schémas ou bien pour intégrer des données géographiques.

Dans le contexte de l'appariement de schémas, l'appariement de données est utilisé soit pour améliorer la qualité de l'appariement de schémas basé sur une ontologie de domaine [Uitermark, 2001], soit dans le but d'apparier les schémas (voir la Figure 13) [Voltz, 2005 ; Kieler, 2007].

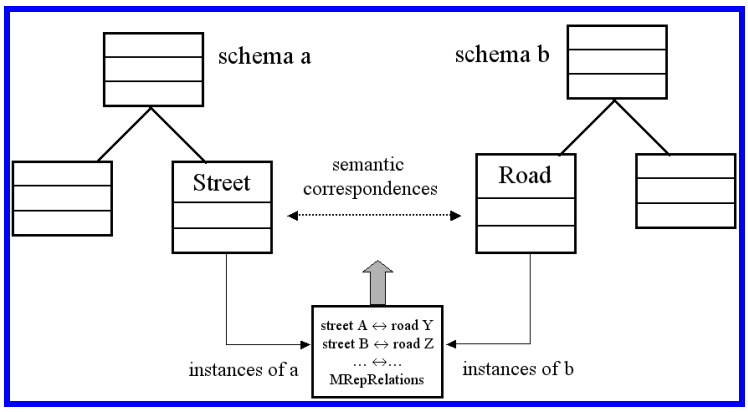


Figure 13. Appariement des schémas dérivé de l'appariement des données, d'après [Voltz, 2005]

L'appariement de données peut être un outil pour comparer des classifications de l'occupation du sol. Dans ce cas, le terme le plus employé est celui de comparaison. L'objectif principal est d'une part de comparer classe par classe deux jeux de données raster pour lesquels il existe une classification préalablement définie, et d'autre part d'évaluer les résultats en utilisant une matrice de confusion dont les lignes sont les différentes classes du jeu de données de référence et les colonnes sont les classes du jeu de données à comparer [Pontius et Cheuk, 2006 ; Hagen-Zanker *et al.*, 2004 ; Fritz et See, 2004 ; Duckham et Worboys, 2005 ; Foody, 2006 ; Vasco et Caetano, 2006 ; Comber *et al.*, 2004].

A.2.4.3 Intégration

La troisième étape du processus est l'étape d'intégration des schémas et des données. Cette dernière étape consiste à définir les stratégies d'intégration des schémas et des données nécessaires, entre autres, pour résoudre les conflits entre les schémas et les données. Le choix de la stratégie dépend des bases de données géographiques à intégrer et des besoins auxquels la base de données géographiques intégrée doit répondre. D'une manière générale, dans la littérature, le nombre de stratégies d'intégration des schémas possibles a été réduit à deux stratégies, à savoir la stratégie multi-représentations et la stratégie mono-représentation [Devogele, 1997].

Dans le cadre de la stratégie multi-représentations, les différentes représentations du monde réel sont préservées et les éléments homologues (schémas et données) sont reliés entre

eux. Cette stratégie a de nombreux avantages, tels que la possibilité de faire des analyses multi-représentations (par exemple la simulation des phénomènes urbains ou la navigation embarquée), la réutilisation des représentations existantes, la possibilité d'intégrer les mises à jour d'une manière automatique et dans toutes les bases en même temps, etc.

La stratégie mono-représentation consiste à fusionner les informations les plus riches et à éliminer les redondances. Dans la littérature, la stratégie mono-représentation est connue également sous le nom de « conflation » [Yuan et Tao, 1999 ; Blasby *et al.*, 2004 ; Doyster *et al.*, 2001]. D'une manière générale, le terme de « conflation » définit l'ensemble des opérations qui consistent, à partir de deux bases de données géographiques, à créer une nouvelle base de données géographiques rassemblant les informations contenues dans les deux autres. [Yuan et Tao, 1999] identifient deux types de conflation : la conflation verticale, c'est-à-dire la fusion de deux jeux de données qui couvrent le même territoire du monde réel, et la conflation horizontale, c'est-à-dire la conflation entre deux jeux de données adjacents. Dans la suite de cette partie, nous nous intéressons seulement à la conflation verticale, et nous l'appelons pour simplifier « conflation ».

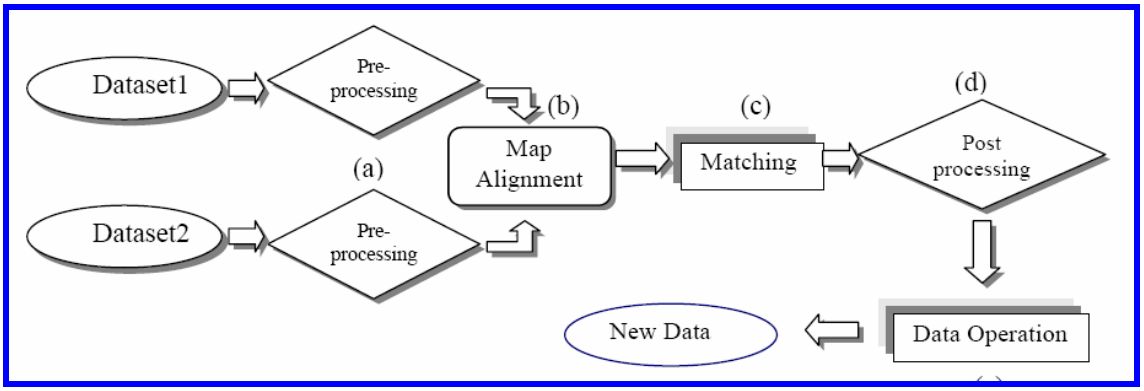


Figure 14. Le processus de conflation proposé par [Yaun et Tao, 1999]

Le processus de conflation proposé par [Yuan et Tao, 1999], à la différence du schéma général d'intégration de données illustré sur la Figure 11, consiste d'abord à aligner[2] les jeux de données (étape (b) sur la Figure 14), puis de réaliser l'appariement de données (étape (c) sur la Figure 14). L'objectif de l'alignement des deux jeux de données est de les rendre comparables pour faciliter le processus d'appariement.

L'appariement de données géographiques utilisé pour intégrer deux bases de données géographiques différentes (produites différemment) s'avère plus complexe en raison des différences qu'il peut y avoir entre les deux bases de données géographiques. En effet, les bases de données géographiques ayant été créées pour répondre à des besoins différents, la représentation des objets est différente. Par exemple, le niveau de détail de chaque base de données géographiques joue un rôle important dans la stratégie d'appariement de données. De nombreuses approches existent dans la littérature et sont souvent spécifiques aux données à apparier. Ainsi, il existe des approches qui s'appliquent aux bases de données géographiques représentant une même réalité à des niveaux de détail différents [Devogele *et al.*, 1998 ;

[2] Le processus d'alignement consiste à superposer deux jeux de données et à réaliser des transformations géométriques afin que les jeux de données deviennent plus comparables du point de vue de la géométrie.

Zhang *et al.*, 2005 ; Mustière, 2006] ou au même niveau de détail [Haunert, 2005 ; Voltz, 2006].

La complexité de ce type d'appariement nécessite d'une part la définition de plusieurs critères géométriques et topologiques basés sur des mesures de distance, d'orientation, d'angle, et sur des relations topologiques [Walter et Fritsch, 1999 ; Lüscher, 2007 ; Mustière et Devogele, 2008] et d'autre part la conception d'outils spécifiques aux données qui rend difficile la mise en place d'un processus d'appariement générique [Mustière, 2006]. L'évaluation de la qualité des liens d'appariement a fait l'objet de divers travaux [Mustière et Devogele, 2008 ; Beeri *et al*, 2004 ; Safra *et al.*, 2006].

A.3 L'appariement, un outil qui dépend des données géographiques

Nous avons vu dans la partie A.2 que les outils d'appariement développés ne sont pas génériques, c'est-à-dire qu'ils dépendent des objectifs du processus d'appariement (étude de la qualité, mise à jour, intégration, etc.). Une autre raison de leur diversité, que nous détaillons dans cette partie, est qu'ils dépendent aussi des données géographiques, c'est-à-dire du type de représentation (point, ligne ou surface) et de leur niveau de détail. Généralement, les auteurs s'intéressent à un type de représentation, ce qui fait qu'il existe des approches d'appariement spécifiques aux objets ponctuels, linéaires et surfaciques.

En ce qui concerne la dépendance du niveau de détail, nous avons remarqué qu'en général, pour les objets ponctuels et surfaciques, les approches d'appariement peuvent être appliquées à la fois aux jeux de données ayant le même niveau de détail et aux jeux de données ayant des niveaux de détail différents. Par contre, les approches destinées à apparier des lignes sont très différentes, selon que les jeux de données à apparier ont ou non le même niveau de détail. Par exemple, la comparaison du nombre d'arcs entrants et sortants de deux nœuds est pertinente lorsque les jeux de données ont le même niveau de détail, alors qu'elle n'a aucune signification lorsque les jeux de données ont des niveaux de détail différents.

De plus, nous avons vu que toutes ces différences entre les bases de données géographiques entraînent une cardinalité des liens différente dans le processus d'appariement.

Dans la suite de cette partie, nous illustrons à titre d'exemple quelques approches d'appariement de données.

A.3.1 Appariement de réseaux au même niveau de détail, approche de [Walter et Fritch, 1999]

Afin d'apparier des réseaux au même niveau de détail, [Walter et Fritch, 1999] proposent une approche basée sur des analyses statistiques en utilisant la géométrie et la topologie. Afin d'obtenir des mesures d'appariement pour faciliter les décisions ils utilisent la théorie de l'information [Shannon, 1948]. La décision finale est prise en fonction des caractéristiques des couples formés et aussi du voisinage. Le couple d'arcs dont l'information mutuelle est maximale est choisi.

L'approche d'appariement est composée de cinq étapes :

- pré-traitement : consiste à recaler les données afin de corriger les biais systématiques. Il est réalisé à partir de points de contrôle choisis manuellement,

- recherche des candidats : consiste à sélectionner les candidats. La sélection est basée sur un critère de distance utilisant un buffer grandissant et sur des considérations sur les angles,
- filtrage des couples très improbables : cette étape consiste à éliminer les couples pour lesquels les différences de longueur ou d'orientation sont trop importantes. Les seuils qui permettent le filtrage sont déterminés grâce à des analyses statistiques, par exemple une courbe de distribution de fréquence. La Figure 15 illustre un exemple de courbe de distribution de fréquence pour les différences de longueur entre des arcs appariés manuellement. D'après cette figure, les arcs potentiellement homologues sont ceux pour lesquels la différence de longueur est inférieure à 30 m.

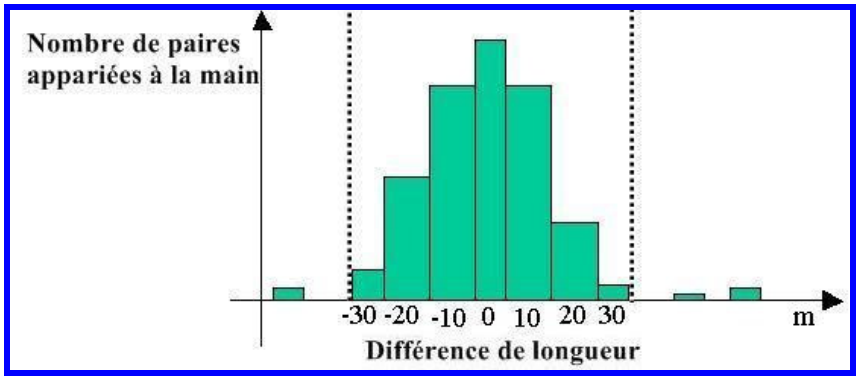


Figure 15. Détermination du seuil optimal de filtrage au moyen d'une courbe de distribution de fréquence

- choix des couples à apparier : quatre critères d'appariement basés sur la géométrie (localisation, longueur, forme et angle) sont combinés. La combinaison consiste à calculer pour chaque couple d'arcs potentiellement homologues une mesure finale représentant la somme de toutes les informations mutuelles[3] calculées pour chacun des critères. Les probabilités qui sont utilisées dans le calcul de l'information mutuelle sont estimées à partir des données déjà appariées. Enfin, le couple d'objets homologues choisi est celui pour lequel la mesure finale a la valeur maximale.
- évaluation des résultats : chaque couple d'objets homologues est évalué automatiquement à travers l'information mutuelle qui est définie dans ce cas comme étant la différence entre l'auto-information et l'information conditionnelle.

L'avantage de cette approche, entre autres, est l'automatisation complète du processus, y compris l'évaluation des résultats et le calcul des liens n : m. Afin de prendre en compte les relations de voisinage et de gagner en temps de calcul, l'appariement est fait d'abord en local, puis par secteur, puis pour le réseau entier. L'inconvénient réside dans le fait que le processus a besoin de connaissances a priori pour définir les paramètres nécessaires à l'appariement optimal, et qu'il nécessite le recalage des données, donc la recherche des points de contrôle.

[3] L'information mutuelle est une mesure issue de la théorie de l'information. L'information mutuelle d'un couple de deux variables aléatoires a et b, notée I (a, b), représente le degré de dépendance statistique de ces variables [Shannon, 1948].

A.3.2 Appariement de réseaux à des niveaux de détail différents, approche de [Mustière et Devogele, 2008]

[Mustière et Devogele, 2008] s'intéressent à l'appariement de réseaux à des niveaux de détail différents, en exploitant la géométrie des données, la topologie et occasionnellement la sémantique. La nouveauté de cette approche réside dans le fait que pour apparier les arcs ils utilisent les nœuds, et plus précisément la topologie aux nœuds. Le processus d'appariement proposé consiste en un enchaînement de six étapes :

- préparation des réseaux : cette étape consiste à passer des objets géographiques (points, lignes, surfaces) à une structure de graphe (nœuds, arcs, faces),
- pré-appariement des nœuds : pour chaque nœud du réseau le moins détaillé R_1, on sélectionne des nœuds candidats à l'appariement dans le réseau le plus détaillé R_2 qui se trouvent à une distance inférieure à un seuil (Figure 16, à gauche). La sélection des candidats est réalisée en utilisant la distance euclidienne,
- pré-appariement des arcs : pour chaque arc du réseau R_2, on sélectionne des arcs candidats dans le réseau R_1, qui se trouvent à une distance inférieure à un seuil (Figure 16, à droite). Dans cette étape, la distance utilisée est la demi-distance de Hausdorff (cette distance sera définie dans la partie A.4.2.1),

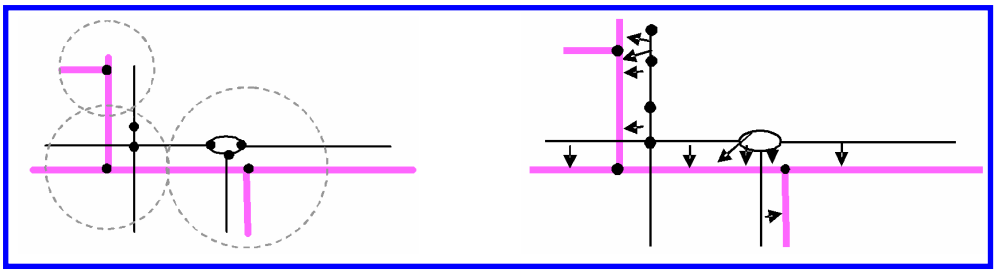


Figure 16. Pré-appariement des nœuds (à gauche) et pré-appariement des arcs (à droite), d'après [Mustière et Devogele, 2008]

- appariement des nœuds : cette étape analyse les nœuds et les arcs pré-appariés. Ainsi, pour chaque nœud du réseau R_1, on analyse chaque nœud candidat du réseau R_2 et en fonction des pré-appariements des arcs entrants et sortants du nœud, le nœud est qualifié de nœud « complet », « incomplet » ou « impossible ». En fonction de cette classification, l'appariement des nœuds est qualifié de certain ou incertain. Afin de définir des liens 1 : n, signifiant qu'un nœud du réseau R_1 est apparié à plusieurs nœuds du réseau R_2, un regroupement des nœuds et de ses arcs connectés est réalisé, composant ainsi un nouveau groupe connexe. La Figure 17 à gauche montre deux types d'appariement des nœuds : un appariement de type 1 : 1 (les nœuds entourés en vert) et un appariement de type 1 : n (les nœuds entourés en rouge),
- appariement des arcs : cette étape est basée sur l'appariement des nœuds. Ainsi, chaque arc du réseau R_1 est apparié, avec le plus court chemin, à un arc du réseau R_2 qui relie les nœuds appariés, extrémités de l'arc de R_1, comme le montre la Figure 17. En fonction de l'évaluation locale des nœuds appariés, l'appariement des arcs est qualifié de certain ou d'incertain. La Figure 17 à droite montre un appariement des arcs de type 1 : 1, c'est-à-dire qu'un arc du réseau R_2 est apparié avec un arc du réseau R_1 (en bleu),

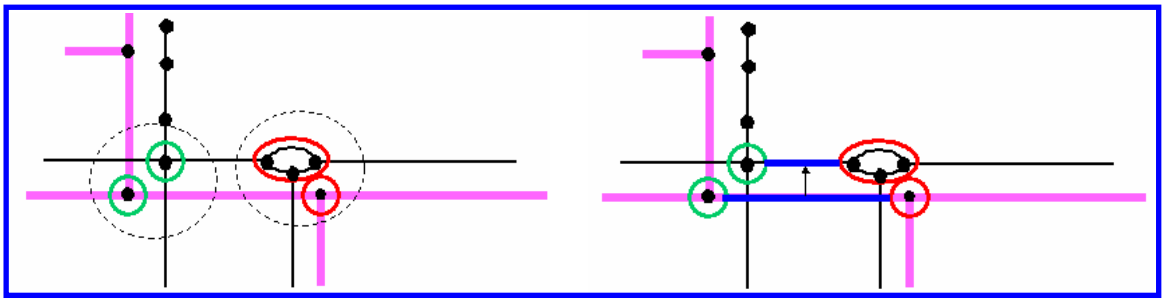


Figure 17. Appariement des nœuds (à gauche) et appariement des arcs (à droite), d'après [Mustière et Devogele, 2008]

- évaluation globale des résultats : cette étape consiste à sélectionner les arcs ou les nœuds appartenant au réseau R_2 appariés à plusieurs éléments du réseau R_1. Normalement cette situation ne doit pas arriver. Si c'est le cas, l'appariement est qualifié d'incertain, dans le cas contraire il est qualifié de certain.

L'originalité de cette approche provient d'une part du fait qu'elle permet d'apparier des réseaux de niveaux de détail différents en définissant des liens n : m, et d'autre part du fait que l'évaluation locale et globale du processus est automatique, ne nécessitant pas l'intervention de l'opérateur.

La qualité de l'étape de sélection des nœuds et des arcs et donc la qualité des résultats peut être améliorée grâce à des critères sémantiques, spécifiques aux données à apparier, qui réduisent l'aspect générique du processus. L'approche donne de bons résultats lorsqu'un jeu de données est inclus dans l'autre jeu de données et de moins bons résultats lorsque le chevauchement des deux jeux de données est faible.

Cette approche permet de prendre en compte l'imprécision de la localisation des arcs à travers le seuil de sélection des candidats lors de l'étape de pré-appariement des arcs (le seuil est défini comme étant la somme de la distance de Hausdorff entre l'arc du réseau le plus détaillé en cours d'analyse et l'arc le plus proche, et d'une distance minimale qui est de l'ordre de grandeur de la précision géométrique du réseau le moins précis). Cependant, ce processus trouve ses limites lorsque la topologie des réseaux présente des imperfections et en bordure de la zone couverte par chaque jeu de données.

A.3.3 Appariement des jeux de données surfaciques, approche de [Bel Hadj Ali, 2001]

L'appariement de deux jeux de données surfaciques proposé par [Bel Hadj Ali, 2001] est composé de quatre étapes, représentées sur la Figure 18 :

- détection des liens d'appariement probables grâce à l'intersection des deux jeux de données,
- suppression des liens parasites (un lien qui relie deux surfaces est supprimé si la distance surfacique entre les deux surfaces est inférieure à un seuil, par exemple le lien {5, E}). La distance surfacique sera définie dans la partie A.4.2.1,
- détection des liens multiples en calculant une matrice d'association. Dans la figure ci-dessous, nous remarquons que deux types de liens ont été définis : un lien de cardinalité n : m, {(A, B, C), (1, 2, 3, 4)}, et deux liens de cardinalité 1 : 1, {D, 5} et {E, 6},

- filtrage des liens n-m, c'est-à-dire que pour deux groupes G1 et G2, on recherche le meilleur groupe en minimisant la distance surfacique.

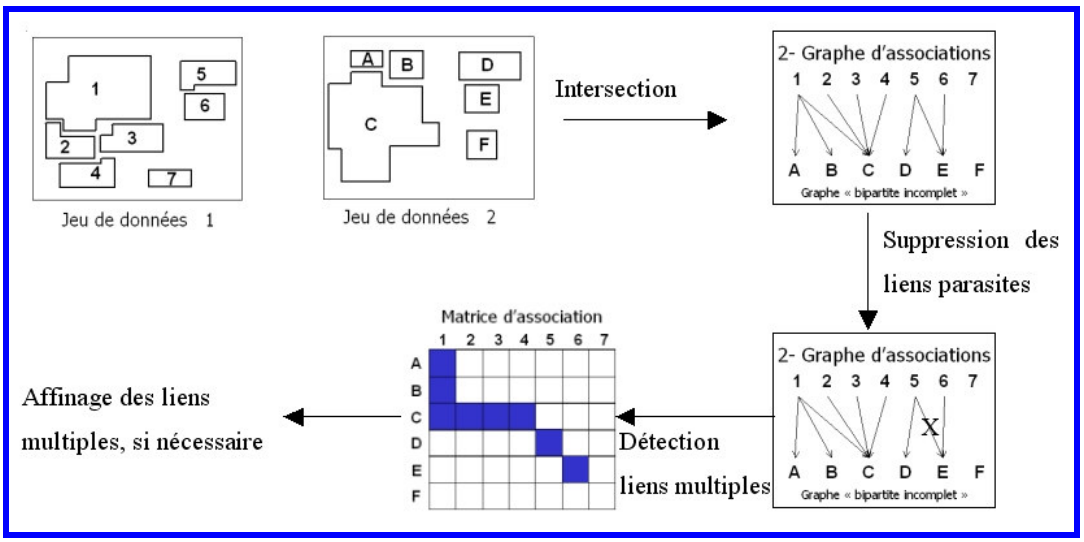


Figure 18. Approche d'appariement de données surfaciques proposée par [Bel Hadj Ali, 2001]

A.3.4 Appariement de plus de deux jeux de données, approche de [Samal et al., 2004]

Un autre aspect important est que la plupart des approches traitent uniquement de l'appariement de deux jeux de données géographiques. Nous illustrons à titre d'exemple une approche intéressante qui permet l'appariement de plus de deux jeux de données.

[Samal *et al.*, 2004] ont proposé une méthodologie fondée sur l'appariement de données pour intégrer plusieurs bases de données hétérogènes. La méthodologie est complexe parce qu'elle permet d'intégrer dans un seul processus des données raster et vecteur ayant des niveaux de détail différents et des représentations différentes. Afin de trouver les correspondances entre les objets homologues, trois étapes ont été mises en œuvre :

- la première étape consiste à apparier les objets selon leurs attributs ou selon leur géométrie, sans prendre en compte les relations entre les objets. Plusieurs critères d'appariement ont été définis, basés sur la géométrie, et sur l'information descriptive. Des méthodes de comparaison des attributs ont été ensuite définies en fonction du type d'attribut (chaîne de caractères, scalaire, localisation) pour chaque critère. Ensuite, pour chaque couple d'objets potentiellement homologues, les critères sont combinés, c'est-à-dire qu'une somme pondérée des mesures issues de chacun des critères est calculée. Le couple d'objets choisi est celui pour lequel la valeur de la somme pondérée est maximale,
- la deuxième étape consiste à analyser les relations entre les objets. On sélectionne d'abord les « objets signifiants », c'est-à-dire les objets appariés dans l'étape précédente. Ensuite, un graphe de proximité est construit pour chaque objet signifiant. Ce graphe de proximité relie un objet signifiant avec les objets de la même base. Ainsi, les nœuds du graphe représentent les objets d'un jeu de données et les arcs signifient les relations entre les objets. Chaque arc du graphe qui relie deux objets possède un poids déterminé par la distance et l'orientation des objets. De la même manière, des graphes de proximité sont

construits pour les autres jeux de données. Sur la Figure 19, nous avons illustré par des lignes roses pointillées le graphe de proximité pour un objet A du jeu de données JD1, et par des lignes grises pleines le graphe de proximité pour son homologue, l'objet B, dans le jeu de données JD2. Enfin, les vecteurs de décalage entre deux graphes de proximité sont obtenus par superposition. Les vecteurs de décalage sont utilisés ensuite pour définir une nouvelle mesure pour les couples d'objets appariés lors de l'étape précédente. Cette nouvelle mesure est utilisée dans la dernière étape,

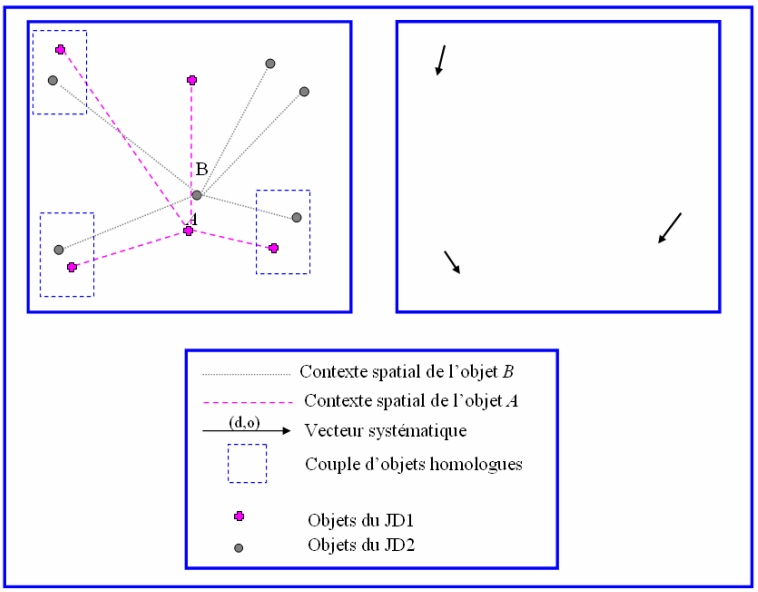


Figure 19. Graphes de proximité et vecteurs de décalage, d'après [Samal *et al.*, 2004]

- la dernière étape consiste à combiner les deux étapes précédentes, c'est-à-dire les mesures issues de la géométrie et de l'information descriptive et celles issues de l'analyse des relations spatiales entre les objets. La combinaison consiste en une analyse récursive, c'est-à-dire que les mesures issues du critère qui analyse les vecteurs de décalage sont révisées à la fin de chaque itération. Plus concrètement, pour chaque objet signifiant, une nouvelle mesure est obtenue en calculant la moyenne entre la mesure pondérée issue de la première phase et celle issue de la deuxième phase. Ensuite à partir de cette nouvelle mesure, les mesures issues du critère basé sur les vecteurs de décalage sont recalculées et le processus continue jusqu'au moment où les mesures se stabilisent. Enfin, pour un objet donné, son objet homologue choisi est celui qui correspond à la valeur maximale de la mesure parmi tous les candidats potentiels.

Un premier avantage de cette approche réside dans le fait qu'elle permet d'apparier en même temps plusieurs jeux de données, que ce soit des données vecteur ou raster. Cette approche permet de définir seulement des liens de cardinalité 1 : 1 en choisissant le lien qui correspond à une clique maximale.

A.4 Critères d'appariement de données et leur combinaison

Dans cette partie, nous présentons d'abord les différents critères d'appariement qui peuvent être utilisés dans le processus d'appariement de données ainsi que quelques mesures qui peuvent être calculées en fonction des critères d'appariement, puis la manière dont ces critères

d'appariement sont combinés dans les approches existantes, afin d'atteindre l'objectif final : apparier deux ou plusieurs jeux de données géographiques.

A.4.1 Différents critères d'appariement

En raison de la complexité du processus d'appariement de données géographiques, due à la complexité des données elles-mêmes, et à la différence des niveaux de détail des bases de données géographiques, il est nécessaire de s'appuyer sur l'évaluation de l'écart entre une ou plusieurs propriétés de deux objets potentiellement homologues. L'évaluation des écarts repose sur de nombreux critères, appelés critères d'appariement. Ces derniers peuvent s'appuyer sur la géométrie des objets, sur les attributs ou sur les relations spatiales entre les objets géographiques.

A.4.1.1 Critères géométriques

Les critères d'appariement géométriques s'appuient sur la géométrie des objets. Cette dernière représente la spécificité des données géographiques par rapport aux données classiques. C'est pour cela que la géométrie peut être considérée dans certaines conditions comme un identifiant commun des objets géographiques. L'hypothèse de base dans le contexte de l'appariement de données est que les objets homologues sont les objets qui sont les plus proches. Cependant, nous avons vu dans l'introduction que cette hypothèse n'est pas toujours vraie, pour de nombreuses raisons. Par exemple, sur la Figure 20 à gauche, nous remarquons que l'objet a_2 a comme homologue son plus proche voisin, l'objet b_3, tandis que l'homologue de l'objet a_1 n'est pas son plus proche voisin, l'objet b_2, mais l'objet b_1.

D'une manière générale, la géométrie des objets géographiques désigne à la fois leur localisation et des informations implicites sur leur forme (longueur, orientation…).

Pour les objets ponctuels, la localisation peut être exploitée pour comparer l'écart de position entre les objets à travers la distance euclidienne. A partir de la distance euclidienne calculée entre un objet et ses candidats à l'appariement, le candidat choisi est celui qui est le plus proche [Minami, 2000 ; Safra *et al.*, 2006]. Pour ne pas apparier au plus proche voisin, [Beeri *et al.*, 2004] proposent une méthode probabiliste qui consiste à analyser, pour chaque objet *obj1*, tous ses candidats à l'appariement $C_{i,\ i=1..N}$ au moyen d'une mesure de confiance. Cette mesure est basée sur la probabilité que l'objet *obj1* soit apparié avec le candidat C_i, et sur la probabilité que le candidat C_i soit apparié avec l'objet *obj1*. La probabilité est définie en utilisant la distance euclidienne entre o*bj1* et le candidat C_i ainsi que les distances euclidiennes entre l'objet o*bj1* et tous ses candidats à l'appariement.

Si pour les objets ponctuels le seul critère géométrique qui puisse être défini est basé sur la localisation, pour les objets linéaires ou surfaciques, en plus de la localisation, des critères d'appariement basés sur des informations implicites issues de la géométrie peuvent être définis et leur exploitation peut améliorer le processus d'appariement. Ainsi, des informations telles que la longueur, l'orientation et la sinuosité d'un objet linéaire, la forme ou l'aire d'un objet surfacique peuvent être utilisées pour comparer deux ou plusieurs objets géographiques.

Deux lignes homologues doivent, par exemple, avoir la même longueur à une tolérance près si le découpage est réalisé de la même manière, et leurs orientations doivent être comparables. De la même manière, deux surfaces homologues doivent être proches du point de vue de la localisation, leur forme doit être similaire, etc. Sur la Figure 20 à droite, nous remarquons l'importance de l'information implicite (la forme des objets) dans le processus

d'appariement. Si seulement la localisation des objets est utilisée, les couples d'objets appariés sont (A_1, B_1) et (A_2, B_3) en raison de leur proximité et de leur pourcentage d'intersection. Or, les objets A_2 et B_3 ne sont pas homologues en raison de leur forme. Si la forme des objets est utilisée pour comparer les objets, cette erreur est réduite, parce que la forme porte aussi implicitement la sémantique.

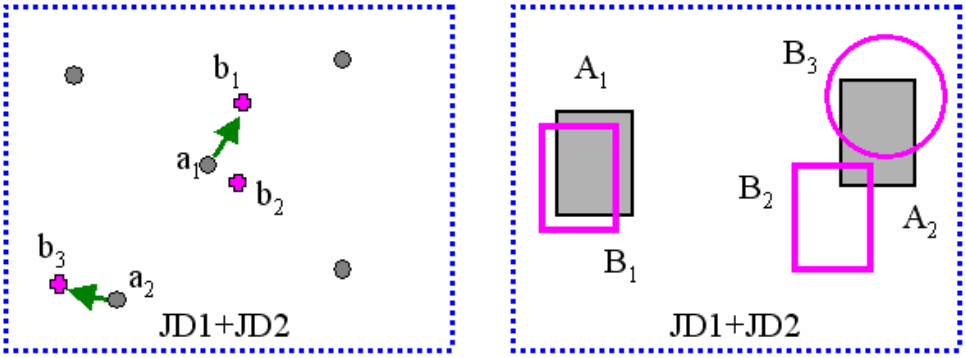


Figure 20. Exemples d'appariement utilisant la géométrie

Afin d'évaluer les critères basés sur la géométrie, de nombreuses mesures sont définies dans la littérature : des distances pour mesurer l'écart de localisation (distance euclidienne, distance de Hausdorff, distance de Fréchet, distance surfacique, etc.) et des mesures pour comparer les informations implicites (signature polygonale, fonction angulaire, etc.).

Nous reviendrons plus en détail sur les mesures de distance dans la partie A.4.2.

A.4.1.2 Critères topologiques et de voisinage

La topologie décrit les relations d'inclusion et d'adhérence entre les objets et elle utilise la notion de voisinage. Les relations topologiques se traduisent par des relations du type : la forêt borde la route, deux routes sont connectées, etc. Les relations topologiques sont construites à partir de la géométrie des objets géographiques initiaux.

Un appariement basé sur la topologie ou les relations de voisinage peut être décrit globalement de la façon suivante : deux objets géographiques A et B sont appariés, c'est-à-dire se ressemblent, si l'objet A possède des relations avec son voisinage comparables ou cohérentes avec les relations de l'objet B avec son voisinage.

Les objets d'une base de données géographiques ont des relations spatiales qui sont décrites par la topologie, les mesures de distance ou d'orientation, la densité, etc. Dans le contexte de l'appariement de données, les relations spatiales peuvent être une bonne source d'information. Des critères d'appariement sont donc définis pour comparer les relations spatiales entre les objets.

Il existe des cas où l'analyse du contexte spatial peut être bénéfique pour le processus d'appariement. La Figure 21 illustre deux jeux de données représentant des points remarquables du relief. Nous observons que les objets a_1, a_2, a_3 et les objets b_1, b_2, b_3 forment un arrangement structuré. Si après l'analyse des objets nous déduisons que les objets a_1, a_2, a_3 d'un jeu de données forment un arrangement structuré et que de la même manière les objets b_1, b_2, b_3 forment un autre arrangement structuré, alors nous pouvons comparer les deux arrangements pour voir s'ils se ressemblent. Cette comparaison peut donner un meilleur

appariement qu'une analyse objet par objet. En effet, l'analyse individuelle pourrait apparier l'objet a_3 avec l'objet b_4, le candidat b_4 étant l'objet le plus proche de l'objet a_3.

Concernant les réseaux géographiques, de nombreux critères d'appariement peuvent être définis en utilisant les relations topologiques entre les objets d'un même jeu de données [Walter et Fritch, 1999 ; Mustière et Devogèle, 2008] ou entre deux objets de deux jeux de données différents [Safra *et al.*, 2006], ou encore en utilisant les relations de voisinage [Stigmar, 2005]. Les relations topologiques permettent, d'une part de créer des critères d'appariement, exactement comme les autres propriétés le font. Par exemple, les propriétés géométriques sont comparées et filtrées par une condition topologique : pour les nœuds, le nombre d'arcs entrants et sortants d'un nœud et la valeur des angles entre les arcs incidents aux nœuds [Voltz, 2006 ; Blasby *et al.*, 2004], et pour les arcs, l'orientation, la longueur, la sinuosité [Zhang *et al.*, 2005 ; Lüscher *et al.*, 2007]. D'autre part, elles peuvent guider l'appariement, c'est-à-dire nous apparions d'abord les nœuds ensuite les arcs connectés aux nœuds appariés [Mustière et Devogele, 2008].

Nous illustrons sur la Figure 21 à droite deux réseaux routiers ayant des niveaux de détail différents. Si on utilise seulement la localisation des nœuds, le nœud a_1 est apparié au nœud le plus proche b_1. Une analyse des arcs incidents aux nœuds permettrait de bien apparier le nœud a_1 avec son homologue, le nœud b_2, puisque a_1 et b_2 ont quatre arcs incidents et b_1 a trois arcs incidents.

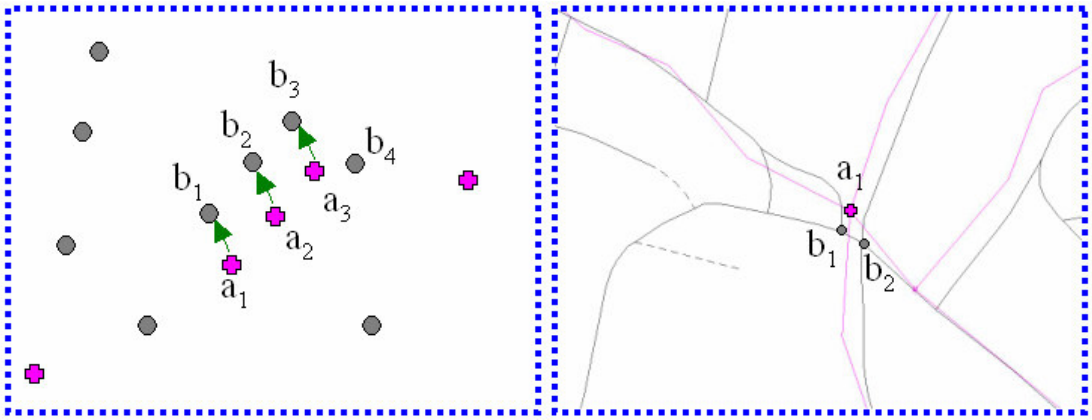


Figure 21. Exemples de relations spatiales entre les objets géographiques

A.4.1.3 Critères attributaires

Comme l'illustre la Figure 22, les objets géographiques possèdent des attributs tels que le nom, la largeur, le nombre de voies ou la nature. Ces attributs peuvent être quantitatifs (par exemple le nombre de voies, la largeur) ou qualitatifs (par exemple le nom, la nature).

Un attribut important à employer impérativement dans le processus d'appariement est la nature des objets géographiques, connue dans la littérature sous le nom d'information sémantique [Comber *et al.*, 2004 ; Abadie et Mustière, 2008]. La comparaison de la nature de deux objets géographiques ne consiste pas en une simple comparaison des chaînes de caractères, mais en un processus plus complexe qui analyse le sens du nom désignant la nature de l'objet à travers, par exemple, des thésaurus, des taxonomies ou des ontologies de domaine [Uitermark, 2001 ; Gesbert, 2005 ; Abadie *et al.*, 2007], par exemple le concept

« chemin » est plus proche du concept « route » que du concept « rivière ». Nous y reviendrons dans la partie A.4.2.3.

La comparaison des noms s'avère très utile lorsque ceux-ci sont présents. De nombreuses mesures d'écart entre les chaînes de caractères existent dans la littérature. Celles-ci sont calculées à partir de distances telles que la distance de Levenshtein [Levenshtein, 1965] ou la distance de Hamming qui compare les lettres communes de deux mots ou deux séries de mots [Hamming, 1950]. Plus de détails sur les distances entre les chaînes de caractères sont présentés dans la partie A.4.2.2.

Concernant les attributs quantitatifs, la comparaison est directe, étant basée sur une simple distance entre les chiffres.

A.4.1.4 Bilan sur les critères d'appariement

Les caractéristiques d'un objet géographique sont illustrées en Figure 22. Ainsi, un objet possède une géométrie qui peut être une des primitives géométriques : le point, la ligne ou la surface, des attributs (qualitatifs et quantitatifs) et des relations spatiales avec les autres objets géographiques. La localisation d'un objet géographique et les informations implicites (la forme, la largeur, l'angle, l'orientation) sont issues de la géométrie.

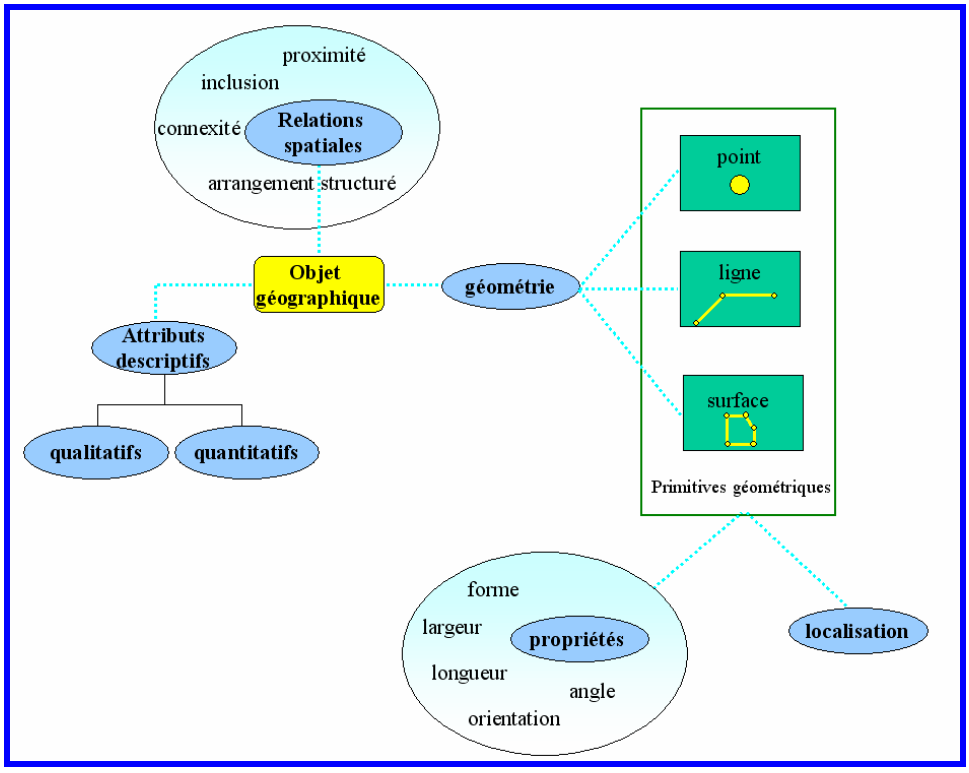


Figure 22. Les caractéristiques d'un objet géographique

Comme nous l'avons vu, en fonction du type de primitive géométrique (point, ligne ou surface), un critère d'appariement peut être défini pour chaque propriété de l'objet géographique. Nous avons identifié trois principaux groupes de critères : les critères géométriques basés sur la géométrie, les critères topologiques et de voisinage basés sur les

relations spatiales, et les critères attributaires fondés sur l'information descriptive, c'est-à-dire les attributs.

A.4.2 Différentes mesures utilisées dans le processus d'appariement

Dans cette partie, nous présentons à titre d'illustration quelques mesures qui peuvent être utilisées pour définir les critères d'appariement que nous avons mentionnés précédemment. Nous précisons qu'il ne s'agit pas d'une liste exhaustive de toutes les mesures existant dans la littérature.

Nous avons classé les mesures en fonction des propriétés des objets géographiques à savoir la géométrie, l'information descriptive et les relations topologiques.

A.4.2.1 Mesures comparant les géométries

Nous présentons dans cette sous-partie quelques mesures qui mettent en valeur l'écart de position ou de forme entre deux géométries en fonction du type de primitive géométrique. Afin de comparer deux géométries, nous pouvons raisonner soit en distance soit en zone d'influence (zone tampon, zone epsilon).

Primitives ponctuelles

Pour les objets géographiques représentés par des points, la principale distance qui permet de mesurer l'écart de position est la distance euclidienne.

Soit deux objets géographiques O_1 et O_2, de coordonnées géographiques respectives (x_1, y_1) et (x_2, y_2) comme illustré sur la Figure 23.

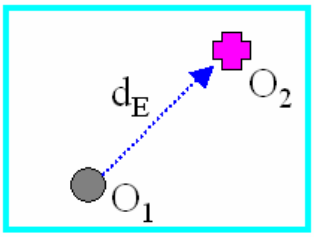


Figure 23. Distance euclidienne entre deux objets géographiques ponctuels

La distance euclidienne d_E entre les objets O_1 et O_2 est définie comme ci-après :

$$d_E = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (1)$$

Primitives linéaires

Afin d'étudier si deux polylignes sont homologues ou pas, nous devons mesurer leur degré de ressemblance. Pour y parvenir nous pouvons comparer leur localisation, leur forme, leur longueur, etc. Contrairement aux objets ponctuels pour lesquels la distance euclidienne est suffisante pour mesurer la distance entre eux, les objets linéaires sont plus complexes et plusieurs distances existent dans la littérature. Évidemment, la première question que nous nous posons est : quelle mesure de distance devons-nous utiliser pour évaluer le plus précisément possible la ressemblance entre les polylignes ?

Nous allons décrire par la suite quelques mesures telles que la distance de Hausdorff, la distance de Fréchet et la distance moyenne.

- **Distance de Hausdorff**

Etant donnés deux objets géographiques représentés par deux lignes L_1 et L_2, la distance de Hausdorff représente l'écart maximal de position entre les deux lignes, voir l'équation suivante.

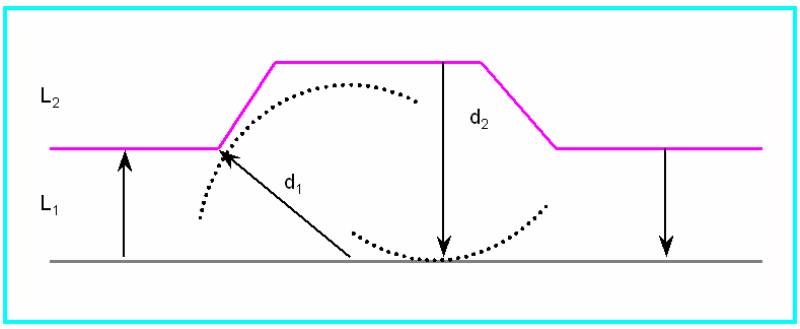


Figure 24. Distance de Hausdorff entre deux lignes

$$d_H = \max(d_1, d_2) \quad (2)$$

où d_1 et d_2 sont définis de la manière suivante :

$$d_1 = \max_{p_1 \in L_1}\left[\min_{p_2 \in L_2}\left[d_E(p_1, p_2)\right]\right]$$
$$d_2 = \max_{p_2 \in L_2}\left[\min_{p_1 \in L_1}\left[d_E(p_2, p_1)\right]\right] \quad (3)$$

où d_E représente la distance euclidienne.

Lorsque les deux lignes à comparer n'ont pas la même longueur, ce qui arrive dans le cas de deux jeux de données qui ont des niveaux de détail différents, la distance maximale porte sur les extrémités. Par conséquent, une solution est d'utiliser seulement la première composante de la distance de Hausdorff, appelée demi-distance de Hausdorff. Notons que cette dernière n'est pas une distance au sens mathématique du terme puisqu'elle n'est pas symétrique. Cette distance est en particulier utilisée pour apparier des données linéaires par [Devogele, 1997 ; Mustière et Devogele, 2008]. Cependant, elle n'est pas restreinte aux primitives linéaires, elle peut être calculée pour les primitives surfaciques.

Malgré le fait que la distance de Hausdorff soit utilisée dans de nombreuses applications, il existe des cas où elle est moins adaptée, par exemple lorsque les polylignes sont sinueuses.

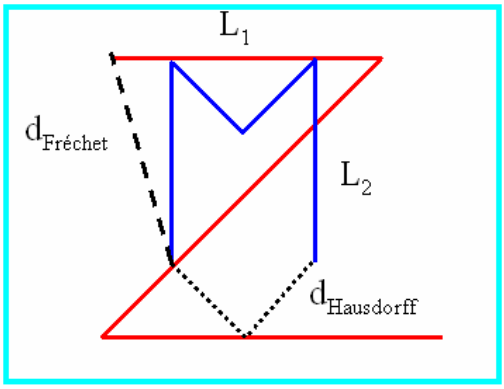


Figure 25. Distance de Hausdorff et de Fréchet, d'après [Badard et Lemarié, 2002]

Dans l'exemple illustré en Figure 25, la distance de Hausdorff est faible mais les polylignes ne se ressemblent pas. Ceci est dû au fait que la distance de Hausdorff considère les polylignes comme des ensembles de points, l'ordre n'étant pas pris en compte. Etant basée sur la distance entre deux points les plus proches et non entre deux points homologues, elle ne permet donc pas de mettre en évidence les différences de forme. Dans ce cas, une mesure plus adaptée est la distance de Fréchet. Nous remarquons dans cet exemple que la distance de Fréchet est supérieure à la distance de Haussdorf.

- **Distance de Fréchet**

La distance de Fréchet d_F [Alt et Godau, 1995] est basée sur la propriété que toute polyligne orientée est équivalente à une fonction continue. Elle est définie de la manière suivante.

Etant données deux polylignes f : [0, N]→V et g : [0, M]→V' et une distance d, une distance euclidienne par exemple, la distance de Fréchet est définie comme ci-après :

$$d_F(f,g) = \min_{\substack{\alpha:[0,1]\to[0,N] \\ \beta:[0,1]\to[0,M]}} \{ \max_{t\in[0,1]} [d(f(\alpha(t)), g(\beta(t)))] \} \quad (4)$$

où N, M∈ℜ représentent le nombre de segments composant les polylignes f et g, et où V et V' sont des espaces vectoriels.

α(t) et β(t) sont des fonctions continues et croissantes avec le temps, avec α(0)=0, β(0)=0, α(1)=N et β(0)=M.

Afin de mieux comprendre cette distance, citons l'exemple de [Devogele, 1997] d'un maître et de son chien se déplaçant chacun le long d'une ligne : « Ils avancent et ils s'arrêtent indépendamment à volonté. La distance de Fréchet entre les deux lignes est la longueur minimale de la laisse qui permet la progression simultanée ».

La distance de Fréchet est plus adaptée pour comparer les formes de deux polylignes [Mascret et Devogele, 2006 ;Bouziani et Pouliot, 2008]. Par contre elle est plus complexe à calculer donc, nécessite plus de temps de calcul que la distance de Hausdorff.

- **Distance moyenne**

La distance moyenne entre deux lignes a été introduite par [McMaster, 1986] afin de comparer la généralisation d'une polyligne avec la polyligne d'origine. Ainsi, l'écart moyen

entre les deux lignes est la surface formée par les deux lignes divisée par la moyenne des longueurs des polylignes (voir la Figure 26).

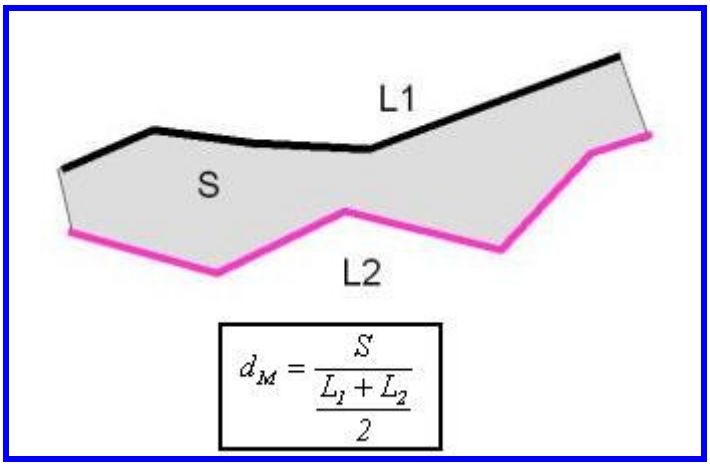


Figure 26. Distance moyenne entre deux lignes

La distance moyenne, par sa simplicité, est facile à mettre en oeuvre, elle permet de mettre en évidence le déplacement dû à la généralisation, mais elle est moins adaptée au processus d'appariement de données, qui consiste à comparer deux polylignes afin de décider si elles sont homologues ou non. La raison est que l'écart de position est exprimé par une valeur moyenne et non pas par une valeur maximale comme par exemple avec les distances de Haussdorf et de Fréchet. C'est pour cette raison que de nombreux auteurs ont affirmé que cette distance doit être absolument utilisée avec d'autres distances telles que la distance de Hausdorff ou la distance de Fréchet [Devogele, 1997 ; Bouziani et Pouliot, 2008].

Une extension de la distance de Hausdorff a été proposée par [Min *et al.*, 2007]. Cette distance est caractérisée par une description statistique complexe de la distance entre deux objets géographiques à travers les distances minimale (la plus petite distance parmi toutes les distances), maximale (la distance de Hausdorff classique) et médiane (la médiane de toutes les distances). Les deux premières sont utilisées pour mesurer la dispersion et la troisième est employée pour mesurer la tendance centrale de la distribution des distances entre deux objets géographiques.

- **Orientation**

Une autre mesure qui permet de comparer deux polylignes est l'écart d'orientation. Ce dernier consiste par exemple à évaluer le degré de co-linéarité local des polylignes.

Etant données deux polylignes L_1 et L_2, le degré de co-linéarité local est défini comme l'écart entre les orientations de la tangente T_1 à L_1 au point le plus proche de L_2, et de la tangente T_2 à L_2 au point le plus proche de L_1 (voir la Figure 27). Afin de calculer l'écart d'orientation entre deux polylignes, nous déterminons d'abord le point de la polyligne L_1 le plus proche de L_2, puis nous calculons l'angle avec l'horizontale de la tangente $\overrightarrow{T_1}$ à L_1 en ce point. De la même manière nous calculons pour la polyligne L_2 l'angle de la tangente $\overrightarrow{T_2}$ au point le plus proche de L_1. Enfin, nous déterminons l'angle θ entre les orientations des deux tangentes.

Si l'angle θ entre les deux polylignes est proche de 0, alors elles sont relativement parallèles et elles ont la même direction. Si la valeur de l'angle θ est proche de π, alors les polylignes sont parallèles et dans la direction opposée. Enfin, si la valeur de l'angle est proche de π/2, alors les polylignes sont perpendiculaires.

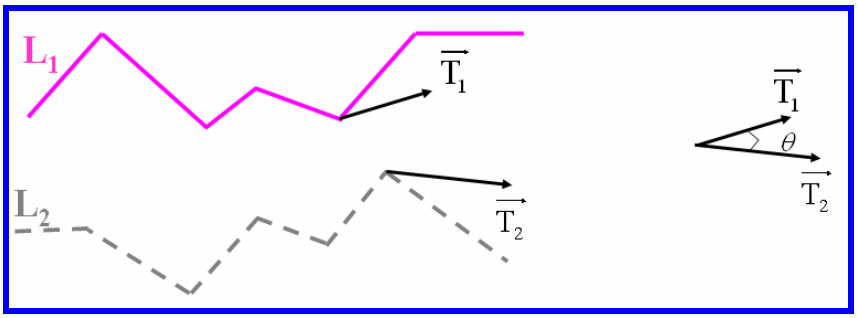


Figure 27. Degré de co-linéarité local θ de deux polylignes L_1 et L_2

Cette mesure est moins pertinente si les longueurs des polylignes sont très différentes, si les polylignes sont décalées ou si elles sont sinueuses. Une amélioration à cette méthode consiste à ne pas déterminer l'orientation localement mais sur toute la longueur des polylignes. Ainsi, nous pouvons calculer l'orientation par rapport à l'axe des x, de la droite orientée passant au mieux au milieu d'un nuage de points ordonnés des points de la polyligne, obtenue au moyen d'une régression par moindres carrés.

- **La bande epsilon**

Une autre mesure qui peut être appliquée à la fois aux objets ponctuels et aux objets linéaires s'appuie sur la bande epsilon. Cette mesure raisonne en termes de zone d'influence et non pas en termes de distance. La bande epsilon consiste à définir un buffer symétrique ou asymétrique autour des objets (ponctuels ou linéaires) afin de trouver leurs objets homologues.

Afin d'évaluer si deux objets sont appariés, il existe plusieurs possibilités d'utilisation de la bande epsilon. Par exemple, deux objets (point ou ligne) sont appariés si un objet se trouve à l'intérieur du buffer défini autour de l'autre objet, c'est-à-dire qu'il se trouve dans la zone de tolérance [Gabay, 1994]. Dans ce cas, la bande epsilon est une technique qui permet de définir trois types d'appariement : appariement point-point, appariement point-ligne et appariement ligne-ligne (voir la Figure 28).

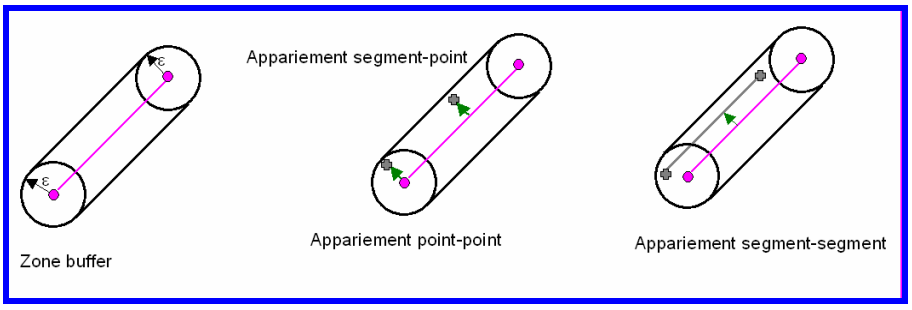


Figure 28. Appariements issus de la mesure basée sur la bande epsilon

[Sui *et al.*, 2004], quant à eux, suggèrent d'évaluer si deux objets objet1 et objet2 sont appariés ou non au moyen d'une mesure qui s'appuie sur la longueur de objet2 à l'intérieur du buffer construit autour de objet 1, comme le montre la Figure 29.

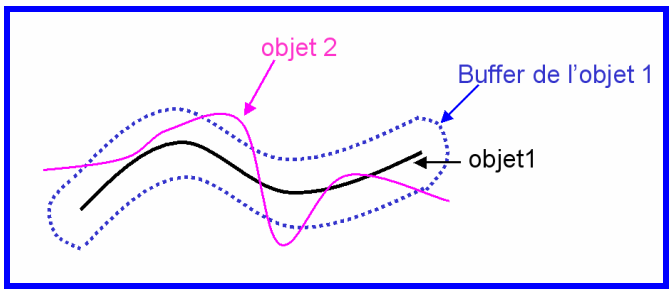


Figure 29. Analyse de deux objets linéaires en utilisant la bande epsilon, [Sui et al., 2004]

Les mesures que nous avons décrites précédemment mettent surtout en évidence l'écart de position entre les lignes. Afin de comparer deux lignes, nous pouvons également comparer leur forme. De nombreuses mesures de forme, que nous ne détaillons pas, existent dans la littérature [McMaster, 1983 ; Buttenfield, 1991 ; Mitropoulos *et al.*, 2005]. Dans le but d'automatiser le processus de généralisation, [Plazanet, 1996], par exemple, s'intéresse à la sinuosité des lignes en proposant de nombreuses mesures qui qualifient localement les formes (les virages des routes).

Primitives surfaciques

La comparaison des primitives surfaciques nécessite à la fois une comparaison des positions et une comparaison des formes. Les distances définies pour les polylignes peuvent être également utilisées pour comparer des surfaces. Cependant, des mesures spécifiques aux données ont été définies telles que la distance surfacique, la fonction à distance radiale ou la fonction angulaire.

- **Distance surfacique**

La distance surfacique permet de mesurer l'écart de position entre deux objets surfaciques. Elle a été initialement définie par [Vauglin, 1997] et utilisée entre autres par [Bel Hadj Ali, 2001 ; Sheeren, 2005] pour apparier des objets surfaciques.

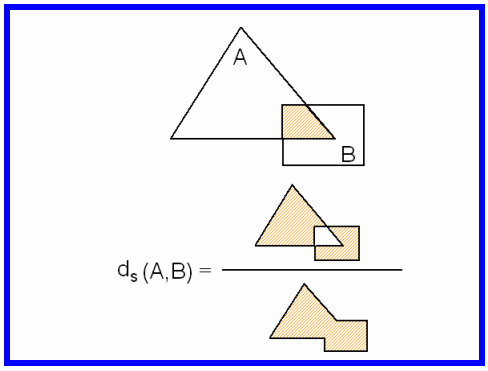


Figure 30. Distance surfacique entre deux objets surfaciques

Etant donnés deux objets surfaciques A et B, la distance surfacique est définie de la manière suivante :

$$d_S = 1 - \frac{S(A \cap B)}{S(A \cup B)} \quad (5)$$

où S(A∩B) représente l'aire d'intersection des objets A et B, et S(A∪B) représente l'aire d'union des deux objets. Notons que la distance surfacique est une distance au sens mathématique du terme, à valeurs dans l'intervalle [0, 1]. Si la distance est égale à 0, alors les deux objets se superposent totalement, c'est-à-dire qu'ils sont égaux, tandis que si la distance est égale à 1, les deux objets n'ont aucun point d'intersection, c'est-à-dire qu'ils sont disjoints.

La comparaison de l'écart de position entre deux objets surfaciques ne suffit pas, il est nécessaire aussi de comparer leur forme. Donnons à titre d'exemple deux mesures qui permettent de comparer l'écart de forme : la fonction à distance radiale et la fonction angulaire.

- **Fonction à distance radiale (ou signature polygonale)**

La fonction à distance radiale décrit un objet surfacique par les mesures des distances séparant le centre de masse du polygone aux points composant son contour en le parcourant dans le sens trigonométrique [Cohen et Guibas, 1997]. Les points ainsi obtenus sont ensuite représentés graphiquement en fonction de l'abscisse curviligne s, normalisée par le périmètre de l'objet surfacique.

La fonction à distance radiale, notée SP(s), est définie de la manière suivante :

$$SP : [0,\ 1] \rightarrow \Re^+,\quad SP(s) = \sqrt{(x_c - x(s))^2 + (y_c - y(s))^2} \qquad (6)$$

où : - x_c et y_c représentent les coordonnées du centre de masse de l'objet surfacique,

- x(s) et y(s) représentent les coordonnées du point courant du contour d'abscisse curviligne s.

Un exemple illustrant la fonction à distance radiale d'un objet surfacique est montré en Figure 31.

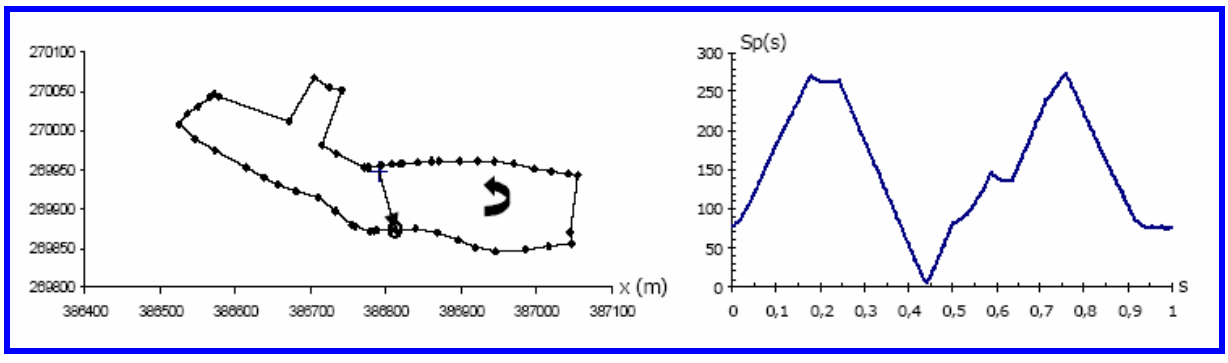


Figure 31. Fonction à distance radiale d'un objet surfacique, [Bel Hadj Ali, 2001]

- **Fonction angulaire (ou fonction Turning)**

La fonction angulaire décrit un objet surfacique par les mesures des angles formés par les segments composant son contour et une demi-droite horizontale orientée selon l'axe des abscisses, [Arkin *et al.*, 1991]. Les points ainsi obtenus sont ensuite représentés graphiquement en fonction de l'abscisse curviligne normalisée par le périmètre du polygone (voir la Figure 32). Afin de comparer deux fonctions angulaires, une correction du déphasage est nécessaire. Notons que la fonction angulaire ne permet pas de décrire des polygones complexes ou ayant des trous.

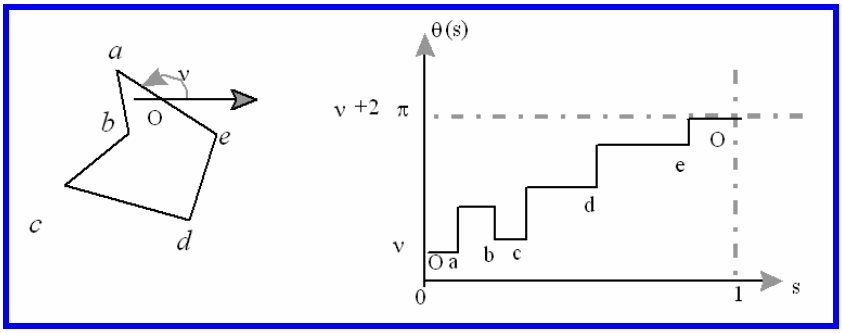


Figure 32. Fonction angulaire d'un objet surfacique

D'autres mesures de forme qui permettent de comparer la concavité, l'élongation, la forme du squelette, la compacité, etc., des objets surfaciques existent dans la littérature [Vauglin, 1997 ; Bel Hadj Ali, 2001].

Dans le contexte de la généralisation cartographique, de nombreux travaux de recherche ont été menés pour évaluer la qualité des données géographiques généralisées [McMaster et Shea, 1992 ; Ruas, 2000, Bard, 2004 ; Mackaness et Ruas, 2007]. Dans ce contexte de nombreuses mesures destinées à mesurer l'écart entre deux objets surfaciques ont été définies. Nous énumérons quelques mesures sans toutefois les détailler :

- la taille : elle compare les surfaces des objets surfaciques,
- l'orientation : elle est basée sur l'orientation des murs des bâtiments [Duchêne et Cambier, 2003],
- la position : elle est basée sur la localisation du centroïde de l'objet surfacique,
- la granularité : elle donne une information relative au niveau de détail de l'objet surfacique, basée sur la longueur du plus petit côté de l'objet.

A.4.2.2 Evaluation de la ressemblance entre les toponymes

Il existe dans la littérature de nombreuses mesures qui permettent de comparer des chaînes de caractères. [Cohen *et al.*, 2003] comparent expérimentalement des mesures de distance afin d'apparier des noms. Nous détaillons dans la suite quelques distances parmi les plus utilisées entre deux chaînes de caractères. Nous trouvons dans [Euzenat et Shvaiko, 2007] une description plus détaillée des mesures de distance existantes.

- **Distance de Hamming**

La distance de Hamming est surtout utilisée en télécommunication pour compter le nombre de bits altérés lors de la transmission d'un message d'une longueur donnée [Hamming, 1950]. Elle peut également être utilisée pour comparer deux chaînes de caractères et consiste à compter le nombre de positions pour lesquelles les deux chaînes de caractères à comparer sont différentes. Cependant elle est moins adaptée pour comparer deux chaînes de caractères de longueurs très différentes ou des chaînes de caractères pouvant subir des substitutions, des insertions ou des effacements.

Etant données deux chaînes de caractères ch_1 et ch_2 et $|ch_i|_{,i=1..2}$ la longueur de la chaîne de caractères ch_i, la distance de Hamming d_H est définie comme suit :

$$d_H = (\sum_{k=1}^{\min(|ch_1|,\ |ch_2|)} ch_1[k] \neq ch_2[k]) + \big\|ch_1| - |ch_2\big\| \quad (7)$$

Nous donnons ci-dessous quelques exemples de calcul de distance de Hamming.

- d_H (« temple » et « temples ») = 1,
- d_H (« william » et « willlaim ») = 3.

- **Distance de Levenshtein** ou **distance d'édition**

La distance de Levenshtein, notée d_L, représente le nombre minimal de suppressions, d'ajouts ou de remplacements pour passer d'une chaîne de caractère à une autre [Levenshtein, 1965].

Etant donné un ensemble d'opérations, Op, sur le type chaîne de caractères, $op : C \to C$, et une fonction de coût $c : Op \to \Re$, telle que pour chaque paire de chaînes de caractères (ch_1, ch_2) il y a une séquence d'opérations qui transforme ch_1 en ch_2, et réciproquement ch_2 en ch_1, la distance de Levenshtein est définie de la manière suivante :

$$d_L(ch_1, ch_2) = \min_{(op_i)_I ; op_n(...op_1(ch_1))=ch_2} (\sum_{i \in I} c_{op_i}) \quad (8)$$

Donnons ci-dessous quelques exemples de calcul de distance de Levenshtein.

- d_L (« temple » et « temples ») = 1, la seule opération à faire est de supprimer ou bien de rajouter le caractère s,
- d_L (« william » et « willlaim ») = 2, afin de transformer « willlaim » en « william », deux opérations sont nécessaires : une suppression de caractère et un remplacement.

A partir de la distance de Levenshtein, ou d'autres distances, il existe plusieurs mesures qui permettent de mesurer l'écart entre deux chaînes de caractères.

Afin de comparer deux chaînes de caractères, une possibilité est de découper les chaînes de caractères en mots (« unités lexicales ») selon un modèle de langue spécifique, processus connu sous le nom de « tokenization ». A partir de ce découpage, [Samal *et al.*, 2004] s'intéressent à l'appariement des mots, c'est-à-dire que la distance de Levenshtein entre deux mots appartenant aux deux chaînes de caractères est calculée et les valeurs sont stockées dans une matrice « mot-mot » [Samal *et al.*, 2004]. De manière à connaître le degré de similarité entre les deux chaînes de caractères, un coefficient de similarité est calculé à partir des valeurs de la matrice. Il est calculé de la manière suivante :

$$coefficient = \frac{\sum_{i=1}^{nombre\ Lignes} valeurMax_i}{(nombre\ de\ lignes + nombre\ de\ colonnes)/2} \qquad (9)$$

L'avantage de cette méthode est que, grâce au découpage en mots, elle permet de gérer des erreurs telles que l'omission (« Boulevard du Général Charles de Gaulle » - « Boulevard du Général de Gaulle »), la substitution (« Aéroport Charles de Gaulle » - « Aéroport de Paris »), la transposition (« Institut Géographique National » ou « Institut National Géographique ») ou l'abréviation (« Boulevard du Général de Gaulle » - « Bld du Gal de Gaulle).

Une autre possibilité pour comparer deux chaînes de caractères, utilisée en particulier dans le processus d'appariement, est de normaliser la distance de Levenshtein par la chaîne de caractères la plus longue :

$$d_T = \frac{d_L(nom_1, nom_2)}{\max(L_1, L_2)} \qquad (10)$$

où : - d_T signifie la distance toponymique,
- d_L représente la distance de Levenshtein,
- L_1 représente la longueur du nom nom_1,
- L_2 représente la longueur du nom nom_2.

Cette mesure autorise des erreurs dues à une faute de frappe, ou à une imprécision (par exemple un nom d'un jeu de données commence avec le nom de l'autre jeu de données : nom_1 est « col de peyrelue ou port vieux sallent », nom_2 est « col de peyrelue »).

A.4.2.3 Evaluation de la ressemblance des concepts à travers la sémantique

En raison de la complexité de l'information géographique, dans le contexte de l'intégration de bases de données, de nombreux auteurs s'intéressent de plus en plus à l'analyse de l'information sémantique, c'est-à-dire l'analyse du sens des concepts. Pour une meilleure exploitation de cette riche information, plusieurs travaux s'intéressent à l'organiser d'une manière hiérarchique. Ainsi, la construction hiérarchique s'appuie sur différentes structures telles que : une ontologie [Bruns et Egenhofer, 1996 ; Comber *et al.*, 2004 ; Rodriguez et Egenhofer, 2004 ; Gesbert, 2005 ; Abadie et Mustière, 2008], ou un treillis d'information [Duckham et Worboys, 2005 ; Pham, 2005].

La plupart des mesures sémantiques sont basées sur la relation hiérarchique « est - un », c'est-à-dire que la similarité entre les concepts d'une taxonomie ou d'une ontologie est basée sur les relations de spécialisation et de généralisation.

De nombreuses mesures de proximité sémantique ont été définies dans la littérature [Resnik, 1995 ; Hirst et St Onge, 1998 ; Rodriguez et Egenhofer, 2004]. Un état de l'art complet sur les mesures de similarité est présenté dans [Patwardham, 2003]. Nous présentons dans ce mémoire seulement la mesure de Wu-Palmer [Wu et Palmer, 1994].

La distance sémantique de Wu-Palmer est basée sur la notion de plus petit généralisant commun. Etant donnés deux concepts C_1 et C_2, la distance sémantique est définie de la manière suivante :

$$d_S = 1 - \frac{2 * prof(C)}{prof(C_1) + prof(C_2)} \quad (11)$$

où C est le plus petit ancêtre commun de C_1 et C_2, et la profondeur du concept C_i, notée prof(C_i), signifie la distance à la racine de C_i. Cette mesure est connue comme étant simple à utiliser et à implémenter. Un exemple d'utilisation de la mesure Wu-Palmer est illustré en Figure 33.

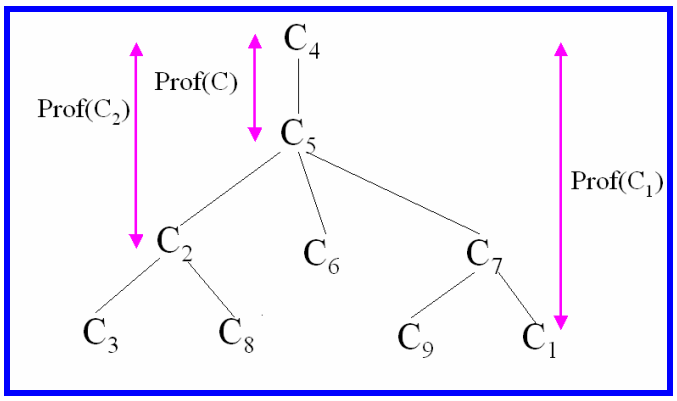


Figure 33. Exemple de calcul de profondeur des concepts dans le cadre de la mesure Wu-Palmer [Wu et Palmer, 1994]

A.4.2.4 Mesures utilisant les relations topologiques

Selon [Clementini *et al.*, 1993] les relations topologiques sont définies comme étant un sous-ensemble des relations spatiales caractérisées par la propriété d'être préservées après avoir subi des transformations telles que la translation, la rotation ou le changement d'échelle. Elles décrivent la position relative de deux objets géographiques au moyen des relations spatiales d'inclusion, d'intersection et d'adjacence [Egenhofer, 1989 ; Egenhofer et Herring, 1990 ; Clementini *et al.*, 1993 ; Zhan, 1998].

De nombreux travaux de recherche liés à la définition des relations topologiques entre les objets géographiques ont été menés. Un premier modèle appelé le modèle des 4 intersections a été défini par [Egenhofer, 1989 ; Egenhofer et Franzosa, 1991], et étendu ensuite à un modèle de neuf intersections [Egenhofer et Herring, 1990]. Le modèle des quatre intersections considère les intérieurs et les frontières de deux objets et analyse la manière dont ces quatre composantes s'intersectent à travers une matrice d'intersections de taille [2x2]. Le modèle des neuf intersections, quant à lui, utilise les notions d'intérieur, de frontière et d'extérieur. Par conséquent, la matrice d'intersection est de taille [3x3]. Le modèle des neuf intersections permet d'une part de formaliser les relations topologiques et d'autre part de déterminer une relation de proximité entre ces relations.

A titre d'exemple nous illustrons sur la Figure 34, les relations topologiques binaires issues du modèle des 4 intersections entre deux surfaces.

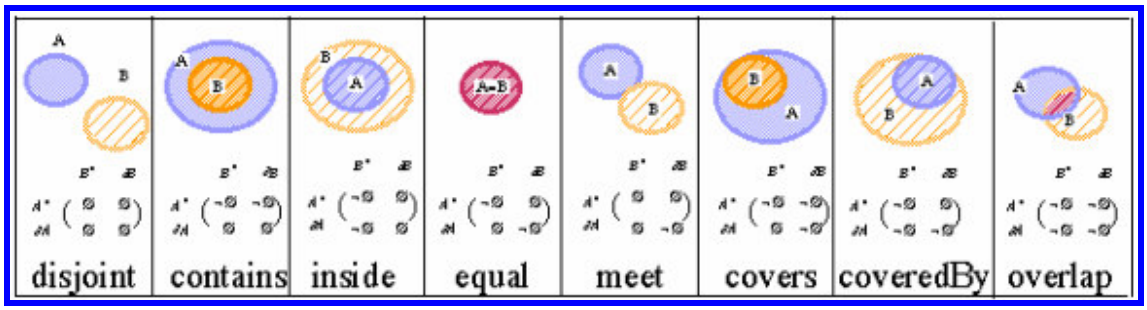


Figure 34. Relations topologiques binaires entre deux surfaces

Le modèle des 9 intersections illustrant les relations topologiques entre une ligne et une surface est présenté sur la Figure 35. Ce modèle, grâce à sa forme en réseau, permet de déterminer les relations de proximité entre les relations topologiques. Par exemple, la relation topologique « touche » est plus proche de la relation « disjoint » que de la relation « contient ».

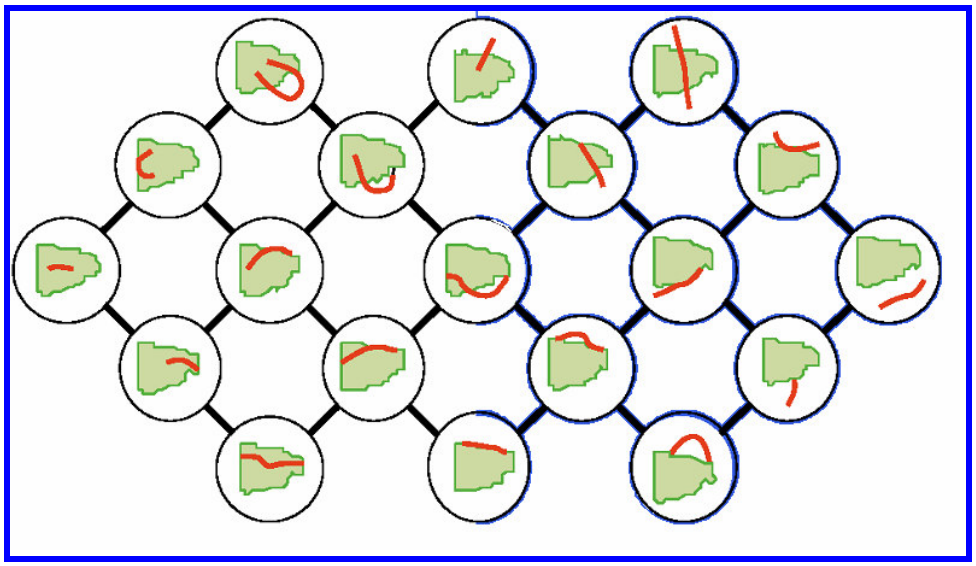


Figure 35. Relations topologiques binaires entre une ligne et une surface et leurs relations de proximité

A.4.3 Les étapes générales du processus d'appariement de données géographiques

Nous avons décrit, dans la partie précédente, les types de critères utilisés dans le processus d'appariement de données. Dans cette partie, nous montrons d'une part les idées de base sur lesquelles un processus d'appariement de données repose, et d'autre part la manière dont les critères d'appariement sont utilisés dans le processus.

Spécifications d'un module générique d'appariement de données géographiques

[Lemarié et Bucaille, 1998] définissent les spécifications d'un module générique d'appariement de données géographiques. Ce module permettrait de regrouper toutes les

méthodes d'appariement déjà existantes. Ainsi, en fonction des données à apparier et également en fonction des besoins, des mesures, des outils ainsi que des paramètres seront utilisés. Le module générique d'appariement est composé des quatre étapes suivantes :

- la sélection dans la première base, c'est-à-dire le choix d'un objet à apparier : un arc, un nœud important, etc.,
- la sélection dans la seconde base des candidats à l'appariement,
- l'étape d'appariement proprement dite, qui choisit l'algorithme d'appariement adapté. Ce dernier est basé sur des critères d'appariement qui exploitent les propriétés des objets géographiques et qui sont combinés afin de prendre une décision finale. Notons que dans cette étape, les résultats sont analysés et s'ils sont médiocres, un autre algorithme doit être proposé,
- le regroupement des objets correspondants, c'est-à-dire le regroupement de plusieurs liens en liens de cardinalité n : m.

Cette approche propose un processus basé sur l'enchaînement de plusieurs étapes. Cependant, la propagation des erreurs due à cet enchaînement est réduite grâce au fait que le processus est itératif au niveau de la sélection (si la sélection est insuffisante, alors il revient à la phase de sélection en changeant les paramètres), et au niveau de la validation des résultats (si la qualité est mauvaise, alors un autre algorithme est proposé). Ce dernier aspect nous semble très difficile à mettre en place. Des questions peuvent se poser telles que : « Quel type d'évaluation : *interactive ou automatique* ? » , « Comment évaluer automatiquement les liens : *sûrs, incertains, très incertains*, ou en termes de *précision et rappel* ? ».

Nous avons remarqué que la plupart des approches suivent à quelques différences près cette proposition, c'est-à-dire que le processus d'appariement est un enchaînement d'étapes. Cependant, nous pouvons noter quelques différences par rapport à ce module d'appariement de données proposé.

Etape de pré-traitement

Avant de passer à l'étape de sélection des candidats, il est possible d'effectuer des pré-traitements sur les deux jeux de données à apparier afin de les rendre plus comparables et donc de faciliter le processus d'appariement. Le recalage des deux jeux de données est un pré-traitement qui est souvent employé dans les approches d'appariement afin de corriger le biais systématique. Ainsi, [Walter et Fritsch, 1999] et [Blasby *et al.*, 2004] utilisent des points de contrôle choisis manuellement, pour recaler deux jeux de données linéaires ayant le même niveau de détail. [Zhang et Couligner, 2004 ; Haunert, 2005 ; Voltz, 2006] quant à eux, proposent un recalage à partir des points de contrôle issus d'un appariement des points. Le recalage de données peut également être vu comme un critère d'appariement, ce qui fait que l'approche met en avant la géométrie des données. D'autres pré-traitements ont été mis en place, tels que le découpage des arcs pour avoir des liens 1 : 1 le plus souvent possible [Voltz, 2006 ; Blasby *et al.*, 2004], l'élimination des nœuds qui n'ont que deux arcs connectés [Stigmar, 2005] ou la simplification du jeu de données plus détaillé afin que les deux jeux de données aient des niveaux de détail comparables [Zhang *et al.*, 2005].

Etape de sélection des objets à apparier

Concernant la sélection des objets à apparier, très peu d'auteurs se posent la question du choix de l'objet à apparier. Dans la plupart des cas, l'objet à apparier est choisi aléatoirement, l'ordre ne comptant pas.

Etape de sélection des candidats

Une fois que l'objet à apparier est choisi, la première étape consiste à sélectionner les candidats à l'appariement. Généralement, pour les objets ponctuels et linéaires, la sélection des candidats est basée sur la géométrie, en utilisant soit une distance maximale [Devogele, 1997 ; Beeri *et al.*, 2004 ; Voltz, 2006 ; Mustière et Devogele, 2008], soit un buffer [Sui *et al.*, 2004 ; Lüscher *et al.*, 2007], soit un buffer grandissant [Walter et Fritsch, 1999 ; Zhang *et al.*, 2005]. Le principe du buffer grandissant est le suivant : s'il n'y a pas de candidat, la valeur de la distance utilisée pour construire le buffer est augmentée (voir la Figure 36). Quelques auteurs proposent de sélectionner les candidats à l'appariement en utilisant à la fois la géométrie et la sémantique [Dunkars, 2003] ou seulement la sémantique, c'est-à-dire que les candidats à l'appariement sont uniquement ceux qui sont de la même nature [Anders et Bildirici, 2004 ; Voltz, 2005]. Pour les objets surfaciques, la sélection est basée sur l'intersection des objets, c'est-à-dire que tous les objets qui s'intersectent sont des candidats potentiels [Bel Hadj Ali, 2001 ; Sheeren, 2005 ; Anders et Bildirici, 2004].

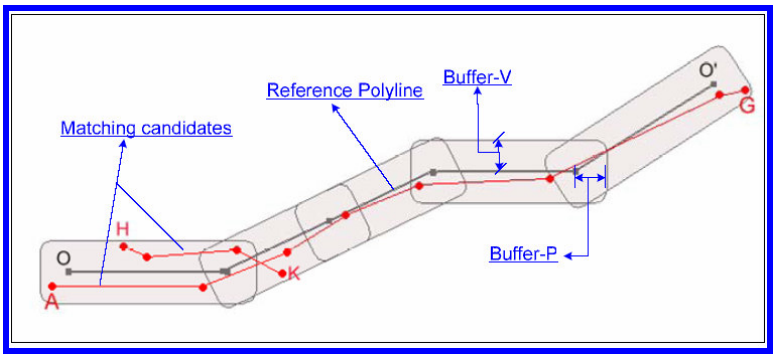


Figure 36. Sélection des candidats basée sur un buffer grandissant, [Zhang *et al.*, 2005]

L'étape de sélection permet de restreindre la recherche des homologues et d'avoir un gain de temps. La recherche des objets homologues est le cœur de l'appariement de données. A partir des différents critères, tous les candidats sélectionnés sont analysés afin de choisir celui qui est l'homologue de l'objet à apparier, en cours d'analyse.

Critères d'appariement

Comme nous l'avons vu dans la partie A.4.1, de nombreux critères d'appariement peuvent être définis à partir des propriétés des objets géographiques, à savoir la géométrie, l'information descriptive et les relations spatiales. Cependant, nous avons remarqué que la plupart des approches proposées dans la littérature utilisent seulement des critères basés sur la géométrie et la topologie.

Pour les objets ponctuels, le critère le plus employé est le critère basé sur l'écart de position entre les objets géographiques [Beeri *et al.*, 2004 ; Minami, 2000].

Pour les objets linéaires et notamment les réseaux, la géométrie et la topologie apparaissent dans la majorité des approches d'appariement existantes. En plus du critère basé sur l'écart de localisation, d'autres critères qui utilisent les informations implicites issues de la géométrie ont été définis. Il s'agit, par exemple, de l'orientation et de la longueur des arcs [Zhang *et al.*, 2005 ; Bouziani et Pouliot, 2008].

L'appariement des réseaux, que ce soit des réseaux ayant le même niveau de détail ou des réseaux ayant des niveaux de détail différents, consiste à apparier d'abord les nœuds et ensuite les arcs. Par conséquent, l'appariement de réseaux est un enchaînement d'étapes : l'appariement des nœuds est utilisé pour apparier les arcs [Devogele, 1997 ; Walter et Fritsch, 1999 ; Voltz, 2006 ; Safra *et al.*, 2006 ; Mustière et Devogele, 2008] ou bien l'appariement des arcs est employé pour apparier les nœuds [Bouziani et Pouliot, 2008].

Un réseau étant composé d'arcs et de nœuds, des relations topologiques existent entre eux : un nœud possède un certain nombre d'arcs entrants et d'arcs sortants, un arc a un nœud initial et un nœud final, etc. Par conséquent, les relations topologiques apportent une information très riche pour l'appariement de données. Les relations topologiques sont utilisées d'une part dans l'enchaînement des étapes : les nœuds sont appariés, puis les arcs des nœuds appariés sont analysés et éventuellement appariés [Stigmar, 2005; Mustière et Devogele, 2008], et d'autre part elles sont utilisées comme une mesure dans la définition du critère d'appariement : le nombre d'arcs incidents d'un nœud, l'angle des arcs incidents d'un nœud, etc. [Walter et Fritch, 1999 ; Zhang et Couloigner, 2004 ; Zhang *et al.*, 2005]. Des approches s'appuient sur les relations topologiques à la fois dans l'enchaînement des étapes et dans les mesures [Voltz, 2006 ; Lüscher *et al.*, 2007].

Pour les objets surfaciques, la géométrie est la seule information utilisée dans le processus d'appariement. Ainsi, l'écart de position est défini au moyen de la distance surfacique [Bel Hadj Ali, 2001 ; Bard, 2004] ou du degré d'intersection entre deux objets [Anders et Bildirici, 2004]. [Gomboši *et al.,* 2003], quant à eux, exploitent également la forme des objets pour pouvoir les apparier.

Combinaison des critères

Nous venons de voir qu'en fonction des données, de nombreux critères d'appariement peuvent être définis. Afin de prendre une décision finale, c'est-à-dire de définir les objets homologues, une étape de combinaison des critères est nécessaire. Nous pouvons distinguer deux types d'approches de combinaison, à savoir les approches qui utilisent les critères les uns après les autres afin de filtrer les candidats à travers des seuils [Devogele, 1997 ; Badard, 2000 ; Bouziani et Pouliot, 2008], et les approches qui utilisent les critères en parallèle afin de définir une valeur pondérée en fonction de leur importance et qui permettra la prise de décision [Dunkars, 2003].

[Zhang *et al.*, 2005] proposent deux méthodes de combinaison des critères. Une première méthode est basée sur un enchaînement des critères pour éliminer les candidats, c'est-à-dire que les candidats qui ne satisfont pas les critères (d'orientation, de longueur, de corde et de localisation) sont éliminés. Ensuite, s'il reste plus d'un candidat, les valeurs des seuils utilisés sont augmentées jusqu'à ce qu'un seul candidat reste. Une deuxième méthode consiste, après l'étape d'élimination des candidats par enchaînement des critères, à combiner les critères en parallèle, c'est-à-dire à calculer une somme pondérée des valeurs issues de chaque critère. Ensuite, le candidat choisi est celui pour lequel la somme pondérée est maximale.

Par rapport à l'approche précédente, [Lüscher *et al.*, 2007] suggèrent un enchaînement des critères pour éliminer les candidats aberrants. Les critères sont le nombre d'arcs incidents aux nœudx et la valeur moyenne des angles formés par les arcs incidents pour les nœuds, et la nature des objets pour les arcs. Ensuite, chaque nœud candidat restant est évalué au moyen de deux critères basés sur la distance euclidienne par rapport au nœud en cours d'analyse et la valeur moyenne de la somme des angles. Un candidat est choisi si les deux valeurs issues des deux critères sont significatives. Si aucun candidat ne se démarque, un expert intervient et réalise un appariement manuel. Ainsi, la manière de choisir le candidat qui a des valeurs significatives fait qu'aucune combinaison n'est effectuée réellement. L'appariement des arcs repose sur l'appariement des nœuds. Dans ce cas, l'appariement des nœuds peut être considéré comme un critère. Ainsi, si deux nœuds sont appariés, alors les arcs sont aussi appariés. Si plusieurs arcs sont connectés aux mêmes nœuds appariés, le choix est fait en utilisant l'algorithme du plus court chemin.

Dans le contexte de l'appariement des réseaux à la même échelle, [Voltz, 2006] propose d'apparier deux arcs en utilisant deux critères basés sur l'appariement des nœuds et sur une somme pondérée. Deux arcs sont appariés si les nœuds initiaux et finaux de ces arcs sont des nœuds appariés et si la somme pondérée est au-dessus d'un certain seuil. De ce fait, il propose d'enchaîner les critères pour apparier les arcs, en sachant qu'un des critères, c'est-à-dire celui basé sur la somme pondérée, est le résultat de la combinaison simultanée de différents critères d'appariement par pondération des mesures issues de chaque critère (de localisation, d'orientation et du nombre d'arcs incidents).

En se basant uniquement sur la localisation des objets ponctuels, [Beeri *et al.*, 2004] suggèrent de combiner deux mesures de probabilité pour apparier des objets ponctuels. Plus précisément, chaque candidat C_i à l'appariement possède une mesure issue de la combinaison de la probabilité que l'objet en cours d'analyse soit l'homologue du candidat C_i et la probabilité que le candidat C_i soit l'homologue de l'objet en cours d'analyse. Ainsi, pour calculer les deux probabilités, la recherche des candidats se fait dans les deux sens, c'est-à-dire que pour un objet appartenant à un jeu de données JD1, on cherche des candidats dans l'autre jeu de données JD2, ensuite pour chaque candidat de JD2 on cherche des candidats dans JD1. La mesure de probabilité est basée sur la distance euclidienne entre les objets ponctuels et sur le niveau de détails des deux jeux de données.

Etape d'évaluation des résultats

L'évaluation des résultats est une étape importante dans le processus d'appariement de données géographiques. Cependant, cette étape est une étape interactive dans la plupart des approches. Très peu d'approches proposent une évaluation automatique [Walter et Fritsch, 1999 ; Clodoveu *et al.*, 2007 ; Mustière et Devogele, 2008]. Une façon de présenter les résultats de l'évaluation du processus d'appariement est de s'appuyer sur des termes de confiance comme par exemple *incertain* et *sûr* [Clodoveu *et al.*, 2007 ; Mustière et Devogele, 2008]. Par exemple, parmi les liens d'appariement trouvés, il existe un certain pourcentage de liens sûrs, c'est-à-dire de liens d'appariement de confiance, et un certain pourcentage de liens d'appariement incertains, c'est-à-dire de liens d'appariement qui doivent être validés éventuellement par un opérateur. [Beeri *et al.*, 2004] et [Safra *et al.*, 2006], quant à eux, présentent la qualité de l'appariement en termes de précision et de rappel après avoir évalué les résultats d'appariement d'une manière interactive.

A.5 Appariement et imperfection dans les données géographiques

Les données en général, et plus particulièrement les données géographiques présentent des imperfections [Goodchild, 1995 ; Zhang et Goodchild, 2002 ; Hunter, 1998]. Malgré ce manque d'exactitude, nous devons les analyser et prendre des décisions.

La production d'une base de données géographiques nécessite plusieurs étapes telles que l'acquisition, l'abstraction, l'archivage, l'analyse, l'affichage de l'information. L'enchaînement des étapes de production, l'étape d'acquisition en particulier, ainsi que la complexité des données géographiques font que les données sont entachées d'erreurs.

Dans le domaine de l'information géographique, il existe de nombreuses taxonomies des imperfections réalisées à partir de la nature des objets géographiques ou des phénomènes qui engendrent l'imperfection. Ainsi, différents termes sont employés tels que : l'incertitude [Fisher, 2003 ; Fisher *et al.*, 2005], l'imprécision [Worboys, 1998 ; Virrantaus, 2003], le vague [Schneider, 1999 ; Cohn et Gotts, 1994], l'erreur [Wright, 2000], les « bona fide et fiat » [Smith et Varzi, 1997]. Ils sont employés avec des sens différents en fonction des domaines d'application, des communautés, des besoins et des points de vue, aucune définition standard n'existant.

Dans le domaine des bases de données géographiques on distingue deux concepts qui caractérisent les données géographiques et qui sont acceptés unanimement par les communautés : la précision et l'exactitude [Devogele *et al.*, 2002].

De nombreux auteurs ont employé le terme d'incertitude comme un concept général. Des travaux de recherche s'intéressent à la caractérisation de l'incertitude et à sa modélisation en proposant des modèles d'incertitude [Heuvelink, 1998 ; Hunter, 1998 ; Zhang et Goodchild, 2002]. Bien que l'incertitude touche à la fois à la localisation des objets géographiques et à leurs attributs, la plupart des modèles sont liés à la localisation des objets géographiques et peu nombreux sont ceux qui étudient l'incertitude de l'information attributaire [Brown, 1998 ; Wang *et al.*, 2005]. Les modèles dépendent d'une part de la nature de l'objet géographique (par exemple un bâtiment a une forme et une frontière bien définies, alors que pour une montagne la forme et la frontière ne sont pas bien définies), et d'autre part de sa représentation : point, ligne ou surface.

L'imprécision sur la localisation des points est en général mesurée et estimée en réalisant des mesures qui utilisent le modèle de l'erreur quadratique moyenne [Hunter et Goodchild, 1997].

De nombreux modèles d'incertitude liés à la localisation des objets géographiques ont été développés dans la littérature, tels que le modèle « circle normal » [Goodchild, 1991], le modèle « standard ellipse » pour estimer l'imprécision de la localisation d'un point [Mikhail et Ackerman, 1976], le modèle « epsilon-band » [Chrisman, 1982] et le modèle « error band » pour estimer l'imprécision de la localisation d'une ligne [Dutton, 1992]. Ces modèles spécifiques aux données ponctuelles et linéaires ont été également adaptés aux données surfaciques.

Lorsqu'il existe des éléments géographiques qui n'ont pas de définition claire ou de frontière bien définie, et donc pour lesquels les approches booléennes ne sont pas adaptées, il existe des modèles basés sur la théorie des ensembles flous [Schneider, 2001 ; Shi et al., 2002 ; Hansen, 2003 ; Hagen *et al.*, 2005 ; Fritz et See, 2004 ; Fonte et Lodwick, 2004 ; Dilo *et al.*, 2007], des modèles basés sur la théorie des ensembles grossiers [Ahlqvist *et al.*, 2003;

Wang *et al.*, 2002] ou des modèles vagues [Erwig et Schneider, 1997 ; Hazarika et Cohn, 2001 ; Tøssebro, 2002].

La taxonomie la plus utilisée et reconnue dans la communauté de l'information géographique est celle de [Fisher, 2003 ; Comber *et al.*, 2005b]. L'idée de base est que les données géographiques peuvent être accompagnées d'incertitude. Notons que Fisher utilise le concept d'incertitude comme un chapeau qui englobe tous les autres concepts, autrement dit, il s'agit de l'imperfection. Ainsi, toute analyse et tout processus qui manipulent des données géographiques doivent prendre en compte ces incertitudes afin de s'assurer que les données sont utilisées correctement et que les analyses et les décisions prises sont justes et de confiance. A partir de ces faits, [Fisher, 2003] s'intéresse à la nature de l'incertitude liée aux données, c'est-à-dire qu'il étudie quels sont les phénomènes et les éléments qui engendrent l'incertitude. Il considère que le principal facteur est le processus d'abstraction du monde réel, à travers par exemple la définition des classes et l'assignation d'un objet à une classe. A partir de la classification des objets du monde réel en objets bien définis et mal définis, Fisher distingue trois formes d'incertitude liées à la définition des classes d'objets observés et des objets individuels composant la classe : d'une part l'erreur si l'objet est bien défini (par exemple un bâtiment), et d'autre part le vague et l'ambiguïté (par exemple une vallée), qui peut être due à un désaccord ou à une non-spécificité, si l'objet est mal défini (voir Figure 37). Plus précisément :

- l'erreur est la différence entre la valeur d'un attribut d'un objet bien défini et la vraie valeur de la même propriété du même objet mesurée sans erreur,
- le caractère vague peut être dû aux spécifications floues ou encore à la nature floue d'un objet, par exemple une forêt. Quelle est la limite précise d'une forêt ?
- le caractère ambigu est dû d'une part à la non-spécificité d'une définition, et d'autre part aux désaccords entre les définitions des objets dans une base de données géographiques. Les causes d'un tel désaccord peuvent être les définitions d'un objet qui ne sont pas complètement spécifiques ou les différences de points de vue. Par conséquent, nous pouvons être confronté à la difficulté de classer une entité.

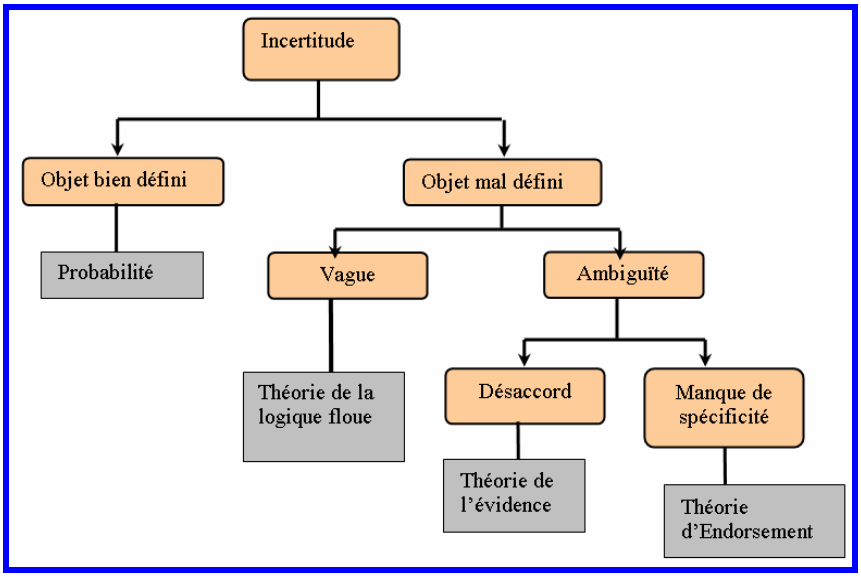


Figure 37. Modèle conceptuel de l'incertitude selon [Fisher, 2003]

[Fisher, 2003] souligne l'importance de la prise en compte de ces incertitudes dans tout processus de prise de décision. Afin de prendre en compte et de diminuer l'effet de l'incertitude, il propose de formaliser chaque concept présent dans la taxonomie à travers des théories mathématiques faisant partie de la famille des théories de l'incertain telles que la théorie des probabilités [Cowell, 1999], la théorie des fonctions de croyance [Shafer, 1976], la théorie des ensembles flous [Zadeh, 1965], la théorie d'Endorsement [Cohen, 1985].

Nous nous sommes inspirés dans notre approche des travaux de Fisher et des solutions qu'il propose. Notre objectif est de proposer une approche d'appariement générique, c'est-à-dire capable de traiter tout genre d'objets géographiques, que ce soit des objets à caractère précis, vague ou ambigu. Nous partons de l'idée que les trois aspects qui caractérisent les données géographiques, c'est-à-dire la localisation, l'information attributaire et les relations spatiales sont soumis à des imperfections.

Afin d'améliorer la robustesse et la qualité de la décision, notre objectif est de proposer un processus d'appariement guidé par des connaissances qui doivent définir des hypothèses. Cependant, la représentation explicite des connaissances n'est pas une tâche facile, celles-ci étant aussi imparfaites, c'est-à-dire imprécises, incertaines et incomplètes.

Pour conclure, notre travail a pour finalité le processus d'appariement de données géographiques, processus guidé par des connaissances qui sont explicitement représentées. Ces idées sont illustrées en Figure 38.

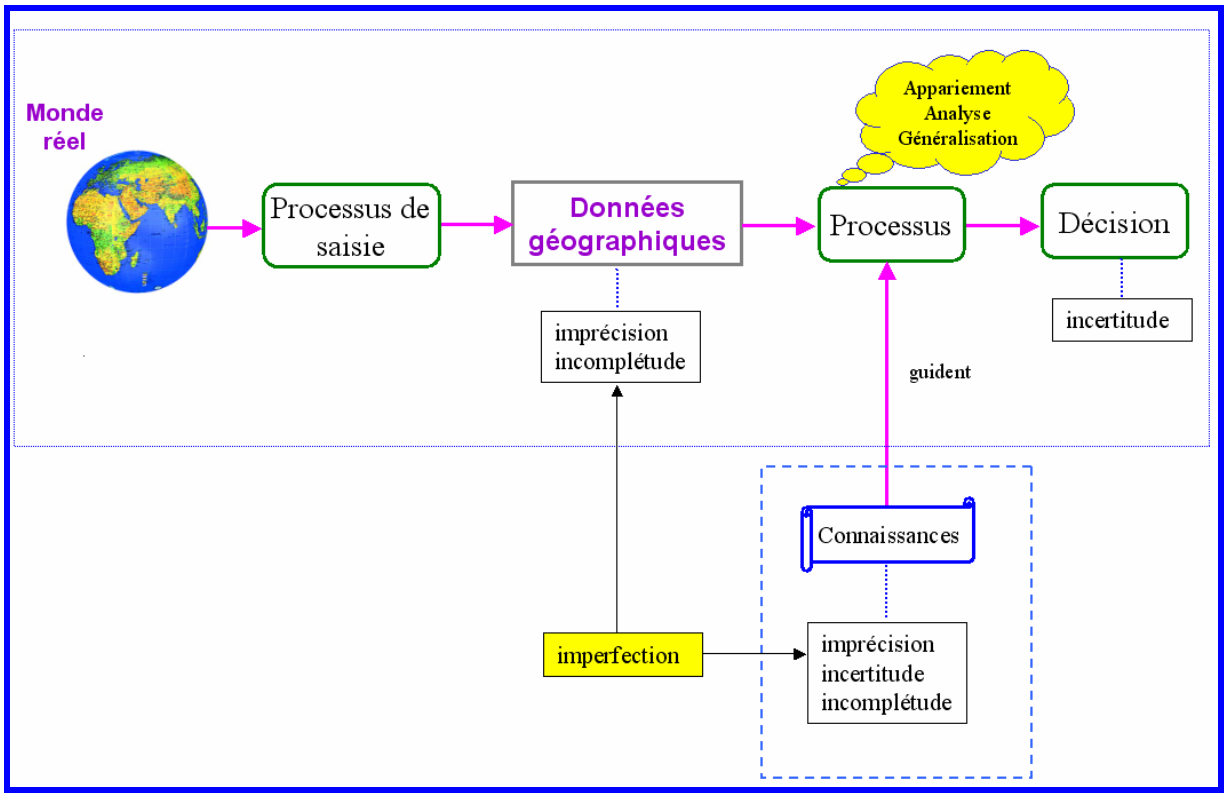


Figure 38. L'imperfection au cours des processus manipulant des données géographiques

Nous reviendrons plus en détail dans le chapitre B sur le concept d'imperfection et sur la manière dont celui-ci peut être explicitement représenté à l'intérieur du processus d'appariement.

A.6 Conclusion et définition du sujet

Dans la partie A.2, nous avons présenté quelques approches d'appariement de données géographiques classées d'une part en fonction des besoins, à savoir l'évaluation de la qualité des données géographiques, le recalage, la mise à jour et l'intégration des bases de données géographiques, et d'autre part en fonction des données géographiques, c'est-à-dire leur représentation (point, ligne ou surface) et leur niveau de détail.

Nous avons également présenté une analyse des approches d'appariement de données en nous intéressant aux critères d'appariement utilisés dans le processus d'appariement, à la manière dont ils sont combinés afin de prendre une décision, ainsi qu'aux étapes générales d'appariement de données (sélection des objets à apparier, sélection des candidats à l'appariement, définition des critères et leur combinaison, évaluation des résultats d'appariement).

Nous présentons en conclusion le bilan des approches d'appariement existantes et nous définissons plus précisément notre sujet de recherche.

Afin d'apparier des données géographiques, il est nécessaire d'évaluer l'écart entre une ou plusieurs propriétés de deux objets potentiellement homologues. L'évaluation des écarts repose sur de nombreux critères, appelés critères d'appariement. Ces derniers peuvent s'appuyer par exemple sur la géométrie des objets, sur l'information attributaire et sur les relations spatiales entre les objets géographiques.

Nous avons constaté que de nombreuses approches d'appariement de données présentées dans la partie A.2 sont basées sur la géométrie des objets géographiques. Ceci est dû au fait que l'information spatiale reste une information importante, pertinente et toujours présente dans une base de données géographiques. Cependant, la localisation des objets géographiques peut être imprécise, par exemple des objets remarquables du relief sont représentés par des points. Or, représenter une vallée par un point est évidemment très imprécis. Dans la même base de données il y a également des précisions différentes, par exemple la localisation d'un sommet est toujours plus précise que la localisation d'une plage.

En conséquence, un processus d'appariement basé uniquement sur la géométrie peut engendrer des résultats d'appariement erronés, parce qu'il ne faut pas toujours apparier à l'objet le plus proche. L'utilisation de la topologie peut améliorer le processus d'appariement.

La nature des objets géographiques, qui est un attribut qualitatif et que nous appelons par la suite *information sémantique*, peut être également utilisée pour apparier des données. Nous remarquons que l'information attributaire et l'information sémantique sont très peu utilisées dans les approches d'appariement de données géographiques, ce qui est justifié lorsque les bases de données à apparier se ressemblent fortement.

Des auteurs affirment que l'information attributaire est incomplète, parfois imprécise et entachée d'erreurs, et qu'elle rendrait le processus d'appariement dépendant des données [Badard, 2000 ; Beeri *et al.*, 2004]. Par exemple, pour le réseau routier, seules les routes principales possèdent un numéro. Il est vrai que, par exemple, une approche basée, en plus que sur de critères géométriques ou topologiques, sur un critère qui comparerait les numéros de route, dégraderait la qualité des résultats si elle ne savait pas gérer l'incomplétude dans les données, c'est-à-dire le cas où le numéro de route n'est pas rempli. De la même manière, la sémantique est hétérogène, différentes classifications existant en fonction de l'échelle de la base de données, des points de vue et des producteurs de bases de données géographiques. La

sémantique est alors uniquement utilisée dans quelques approches au niveau de l'étape de sélection des candidats à l'appariement.

Une autre justification donnée par des auteurs au fait qu'ils n'utilisent pas l'information attributaire et la sémantique est que de nombreuses approches se proposent comme objectif de définir un processus d'appariement de données générique. Il est évident qu'un processus d'appariement basé uniquement sur la géométrie et sur les relations topologiques entre les données pourrait être plus facilement générique.

Toutes ces raisons font que les approches d'appariement sont basées principalement sur la géométrie et la topologie. Cependant, nous avons vu que la géométrie et la topologie ne sont ni parfaites, ni suffisantes, surtout dans le cas des bases de données sensiblement différentes en contenu et en niveau de détail. Cet aspect nous amène à supposer que l'information attributaire et la sémantique peuvent être exploitables et utilisables dans le processus d'appariement, afin de remédier aux cas d'appariement où la géométrie et la topologie ne sont pas suffisantes pour prendre une décision.

Nous remarquons par ailleurs que les méthodes d'appariement, qu'elles soient appliquées sur des données ponctuelles, linéaires ou surfaciques et qu'elles soient utilisées pour apparier des jeux de données à la même échelle ou à des échelles différentes, sont basées soit sur un enchaînement de différents critères (voir Figure 39), soit sur la combinaison de critères (voir Figure 40).

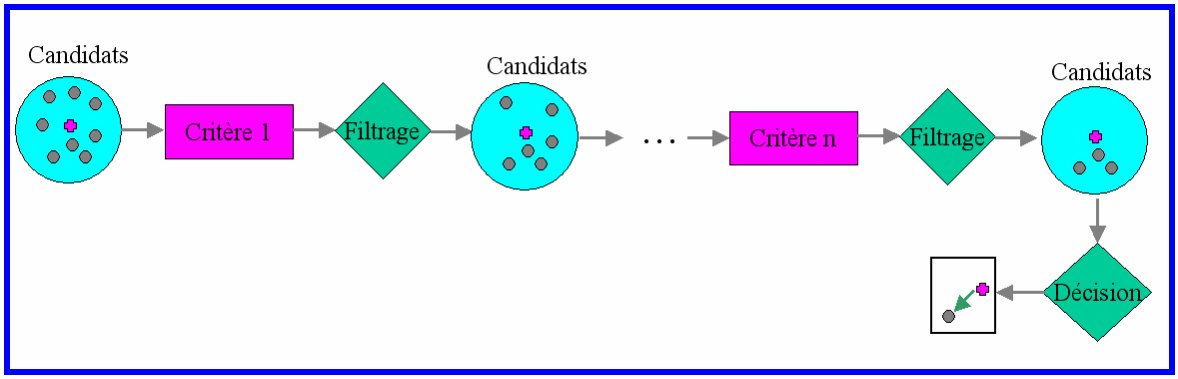


Figure 39. Appariement par enchaînement des critères

Si on enchaîne les critères, d'une manière générale, pour un objet donné, on sélectionne des candidats à l'appariement puis les candidats sont filtrés au fur et à mesure pour qu'à la fin les meilleurs candidats soient choisis. L'inconvénient de ce type d'approche est qu'il est dépendant d'une part de l'ordre des critères à apparier, et d'autre part des seuils fixés pour le filtrage des candidats. Si par exemple un candidat ne respecte pas le premier critère, il sera définitivement éliminé. L'avantage est qu'il est plus rapide que l'approche basée sur la combinaison des critères.

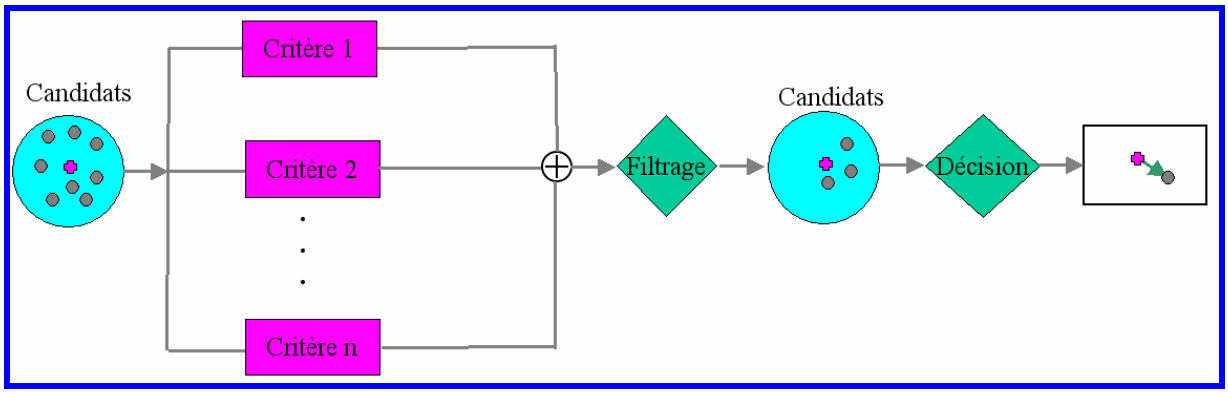


Figure 40. Appariement par combinaison des critères en parallèle

Contrairement à l'approche par enchaînement des critères, l'approche basée sur la combinaison des candidats ne nécessite pas la définition d'un ordre pour chaque critère d'appariement. La décision est prise après avoir fait une somme pondérée ou utilisé une autre technique de combinaison des valeurs issues de chaque critère. La difficulté consiste à définir les valeurs de pondération pour chaque critère. La somme est pondérée en fonction des critères indépendamment des valeurs des mesures issues de la comparaison des différentes propriétés.

Par ailleurs, d'une part la majorité des approches ne prennent pas en compte les imperfections dans les données d'une manière formalisée et d'autre part, pour la majorité des approches les connaissances sont à l'intérieur du processus.

Dans le cadre de ce travail de thèse plusieurs objectifs ont été fixés. Les objectifs visent d'une part à réaliser un processus d'appariement basé sur une approche générique capable de s'adapter à la fois au type de données (ponctuel, linéaire, surfacique) et aux spécificités des jeux de données (le même niveau de détail ou des niveaux de détail différents), et d'autre part à prendre en compte les imperfections en s'appuyant sur une combinaison des critères définis à partir des connaissances et des données elles-mêmes.

Un objectif est de proposer une approche d'appariement générique, c'est-à-dire capable de traiter tous types d'objets géographiques, qu'ils soient définis de manière précise, vague ou ambigue. Ainsi, nous partons de l'idée que les aspects qui caractérisent les données géographiques, c'est-à-dire la localisation, la forme, l'information attributaire et les relations spatiales, peuvent être imprécis.

Afin d'améliorer la robustesse et la qualité de la décision, notre objectif est de proposer un processus d'appariement guidé par des connaissances qui proviennent des données elles-mêmes, des spécifications ou des experts, et qui doivent générer des hypothèses.

Cependant, la représentation explicite des connaissances n'est pas une tâche facile, celles-ci étant aussi imprécises, incertaines ou incomplètes. Cette représentation fait l'objet de plusieurs travaux de recherche dans le domaine de l'intelligence artificielle [Bouchon-Meunier, 1995]. Une solution, pour améliorer la robustesse et la qualité de la décision, est de fusionner les sources d'information imparfaites, appelées dans notre cas des critères d'appariement. Afin d'atteindre nos objectifs, plusieurs besoins peuvent alors être exprimés à travers les questions suivantes :

- Comment prenons-nous en compte les imperfections ?
- Quelles sont les connaissances à utiliser dans le processus d'appariement ?
- Comment représenter d'une manière explicite les connaissances ?
- Comment renforcer la certitude d'une connaissance ou remédier à des connaissances imprécises, incertaines ou manquantes ?
- Quel outil utiliser pour prendre une décision ?
- Quelle est la certitude de la décision prise ?

Afin de répondre aux questions que nous venons d'énumérer et d'atteindre nos objectifs, nous avons ainsi choisi de nous appuyer sur la théorie des fonctions de croyance. Le choix de la théorie des fonctions de croyance a été adopté pour de nombreuses raisons :

- elle permet de prendre en compte et de modéliser à la fois l'imprécision, l'incertitude et l'incomplétude,
- elle permet de modéliser la connaissance parfaite et l'ignorance totale,
- en termes de fusion des connaissances, elle possède des outils qui permettent de combiner plusieurs avis,
- elle permet de mettre en évidence et de gérer le conflit, c'est-à-dire le désaccord entre les connaissances,
- elle possède des outils qui permettent la prise de la décision.

Nous décrirons plus en détail la notion d'imperfection dans les connaissances et la théorie des fonctions de croyance dans le chapitre B, puis nous détaillerons notre approche d'appariement basée sur cette théorie dans le chapitre C.

CHAPITRE B
Imperfection, représentation et fusion des connaissances

B Imperfection, représentation et fusion des connaissances

Nous présentons dans ce chapitre les types d'imperfection rencontrés dans les données et dans les connaissances. Après avoir présenté la raison pour laquelle nous avons choisi la théorie des fonctions de croyance pour réaliser notre approche d'appariement de données géographiques, nous présentons brièvement quelques applications. Enfin, nous décrivons le contexte général de la théorie des fonctions de croyance.

Introduction

D'une manière générale, un système d'information, que ce soit un système d'information géographique, un système de conception, un système embarqué, un système de diagnostic ou un système d'information de gestion, est un modèle informatique du monde réel et donc une abstraction de celui-ci. Le point le plus critique est porté sur la précision et la pertinence de la représentation. Malheureusement, nos connaissances et notre perception du monde réel sont le plus souvent subjectives.

Afin d'avoir un système d'information très fiable, deux solutions sont possibles : soit on restreint le système d'information à une portion du monde réel, c'est-à-dire qu'on utilise seulement la partie sur laquelle on possède des informations fiables, soit on définit un système d'information qui accepte des informations imparfaites. C'est cette dernière solution qui est employée dans la plupart des systèmes d'information.

Afin qu'il soit générique, un processus doit être ouvert pour permettre l'intégration des connaissances. Pour que les connaissances soient exploitables par le processus, elles doivent être formalisées et stockées d'une manière informatique, tout en prenant en compte leurs imperfections. Cet aspect est détaillé dans la suite de ce chapitre.

L'imperfection a fait l'objet de nombreux travaux dans le domaine de l'Intelligence Artificielle (IA) [Zadeh 1965 ; Niskanen, 1989 ; Bouchon-Meunier, 1995; Colot, 2000 ; Smets, 1997 ; Masson, 2005].

En général, en IA, la notion d'imperfection dans les connaissances fait appel à trois concepts : l'imprécision, l'incertitude et l'incomplétude, illustrés sur la Figure 41 [Bouchon-Meunier, 1995; Colot, 2000 ; Masson, 2005 ; Smets, 1997].

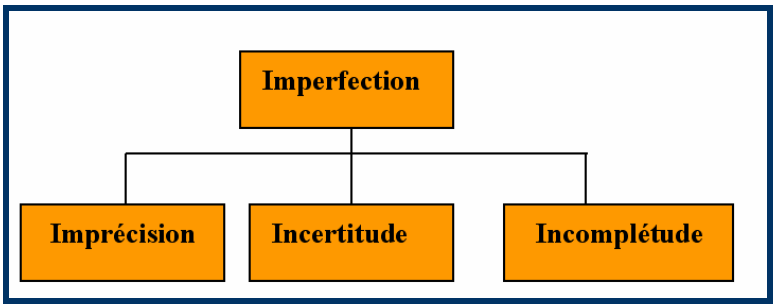


Figure 41. Grands types d'imperfection

Les trois concepts d'imperfection sont définis comme ci-dessous :

- Imprécision : concerne la difficulté d'exprimer clairement et précisément un état de la réalité par une proposition (par exemple « dans la salle il y a environ une centaine de personnes », ou « le poids de la table est d'environ 25 kg », ou « Jean est grand ». Modéliser l'imprécision consiste à formaliser les termes de « environ », « centaine » ou « grand »,
- Incertitude : concerne un doute sur la validité d'une connaissance. Elle est due à la fiabilité de l'observateur peu sûr de lui ou prudent qui ne peut pas déterminer la valeur de vérité de la connaissance. Ex : « Je crois que dans la salle il y a 100 personnes »,
- Incomplétude : il s'agit d'une absence de connaissance ou d'une connaissance partielle. Elle est due à une incomplétude dans les données, à l'absence d'une connaissance explicite ou à l'existence d'une connaissance générale. Par exemple, pour une instance d'une base de données la valeur de l'attribut *Nom* n'est pas remplie.

L'imprécision selon [Dubois et Prade, 1988] fait référence au contenu de l'information tandis que l'incertitude concerne plutôt la qualité de l'information, relativement à sa vérité.

Il existe d'autres taxonomies des types d'imperfection plus détaillées que celle présentée sur la Figure 41. Par exemple, [Niskanen, 1989] propose une taxonomie des types d'imperfection, illustrée sur la Figure 42, dans laquelle la notion de non-précision fait appel à quatre concepts : l'incertitude, l'imprécision, l'ambiguïté et la généralité. Pour lui, l'incertitude est un concept mesurable lié à la notion d'erreur, tandis que l'imprécision n'est pas mesurable. Il propose trois termes relatifs à la notion d'imprécision: l'imprécision ontologique, c'est-à-dire que les objets sont naturellement imprécis, l'imprécision linguistique, c'est-à-dire que les objets ne sont pas définis précisément, et l'imprécision épistémologique, c'est-à-dire que les objets ne peuvent pas être perçus d'une manière précise. D'après [Niskanen, 1989], la généralité peut être par exemple la représentation multiple de la réalité en fonction du niveau de détail. L'ambiguïté apparaît lorsque plusieurs points de vue existent sur le même phénomène du monde réel, par exemple la classification de l'occupation du sol.

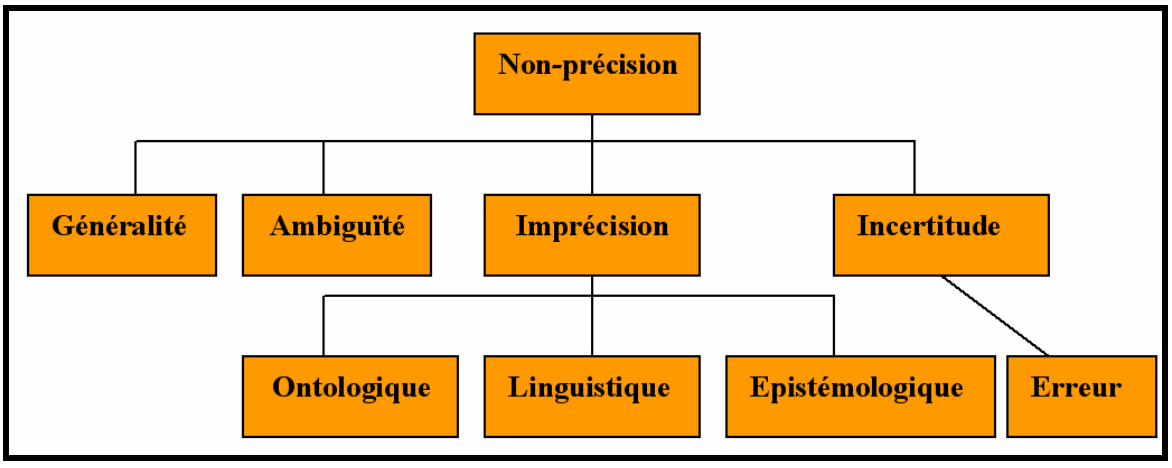


Figure 42. Taxonomie des imperfections selon [Niskanen, 1989] traduite de l'anglais

Dans le domaine de l'intelligence artificielle (IA), la notion d'imprécision peut être décrite à travers les termes « vague », « flou » ou « ambigu ». Le caractère vague ou flou d'une information concerne l'absence de contour bien délimité et défini. Le caractère ambigu est présent lorsqu'une affirmation renvoie à plusieurs hypothèses et il est lié au langage.

Comme nous l'avons vu, les connaissances sont imparfaites. Par conséquent, dans n'importe quel système expert ou processus, en présence de ces connaissances imparfaites, il est nécessaire, comme le souligne [Bouchon-Meunier, 1995], de « préserver des imperfections jusqu'à un certain point, ce qui permet de ne pas perdre une information intéressante, mais de parvenir à une représentation facilement manipulable de façon automatique ». De nombreux travaux de recherche liés à la formalisation des imperfections ont été réalisés. Ainsi, de nombreuses théories, telles que la théorie des probabilités, la théorie des ensembles flous [Zadeh, 1965 ; Bouchon-Meunier, 1995], la théorie des possibilités [Zadeh, 1965 ; Dubois et Prade, 1988], la théorie des fonctions de croyance [Dempster, 1967 ; Shafer, 1976], etc., possèdent des outils pour manipuler et modéliser les imperfections.

Nous y reviendrons plus en détail dans la suite de ce chapitre.

B.1 Analyse des imperfections des données géographiques

Dans cette partie, nous illustrons à travers un exemple les imperfections présentes dans les données géographiques.

Considérons deux jeux de données géographiques ponctuels représentant des points remarquables du relief ayant des niveaux de détail différents. Sur la Figure 43 et la Figure 44, les croix noires représentent les points remarquables du relief du jeu de données le plus détaillé JD2, et les points roses représentent les points remarquables du relief du jeu de données moins détaillé JD1.

Nous illustrons en Figure 43 les imperfections liées à la localisation des objets géographiques. Comme nous pouvons le remarquer, lorsque l'entité est de grande taille et qu'elle est représentée par un point, la localisation peut être imprécise. Par exemple, une vallée est représentée dans les jeux de données par un objet ponctuel signifiant en général le centre de la vallée. De plus, les objets géographiques d'un même jeu de données ont des précisions différentes en fonction de leur type. Par exemple, les vallées et les pics sont représentés de la même manière, c'est-à-dire par un objet ponctuel, alors que la localisation d'un pic est plus précise que la localisation d'une vallée.

En raison de la différence de niveau de détail entre les deux jeux de données, nous remarquons qu'il y a des points remarquables du relief qui sont représentés dans le jeu de données plus détaillé mais pas dans l'autre. Par conséquent, nous avons une incomplétude des données dans un jeu de données.

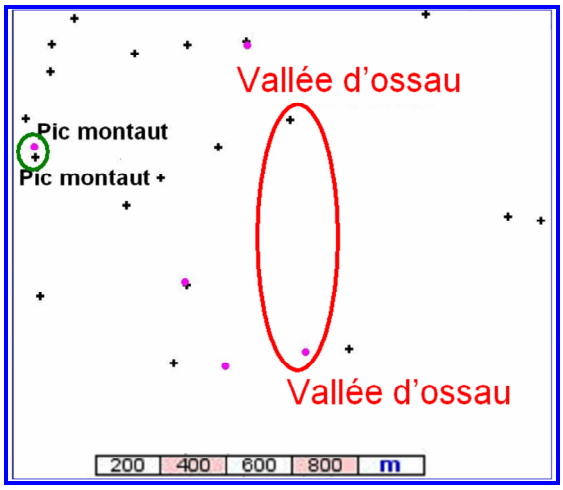


Figure 43. Imperfection de la localisation des données géographiques

Les attributs des données géographiques peuvent aussi présenter des imperfections, en particulier des imprécisions et des incomplétudes. Les imprécisions touchent surtout à la valeur de l'attribut, alors que les incomplétudes peuvent apparaître au niveau de l'attribut lui-même, lorsqu'un jeu de données possède un attribut qu'un autre jeu de données ne possède pas, ou au niveau de la valeur de l'attribut, lorsque l'attribut existe, mais pour certains objets la valeur n'est pas remplie.

La Figure 44 à gauche illustre quelques exemples d'imperfection qui peuvent apparaître pour l'attribut renseignant la nature des objets géographiques. Ainsi, sous l'hypothèse que les jeux de données possèdent l'attribut « Nature », les valeurs de cet attribut peuvent être imprécises ou incomplètes. Les imprécisions peuvent être dues aux regroupements des concepts, par exemple le point remarquable du relief de JD_1 appelé « La devèze » est de nature « sommet, crête, colline », alors que son homologue dans JD_2 est de nature « sommet ».

Un autre facteur qui peut générer de l'imperfection est la classification différente des objets géographiques. L'objet géographique nommé « Pic de l'Escarpu » est de nature « pic » dans JD_1 et de nature « sommet » dans JD_2. Dans ce cas, la raison pour laquelle les objets ont été classés différemment est due au caractère très proche d'un point de vue sémantique entre les concepts sommet et pic. Nous pouvons donc penser qu'il s'agit d'une interprétation humaine différente dans ce cas.

En ce qui concerne l'incomplétude, nous pouvons remarquer dans cet exemple que pour l'objet de JD_2 appelé « Munhoa » la valeur de l'attribut « Nature » n'est pas remplie, tandis que son homologue dans JD_1 est de nature « pic ».

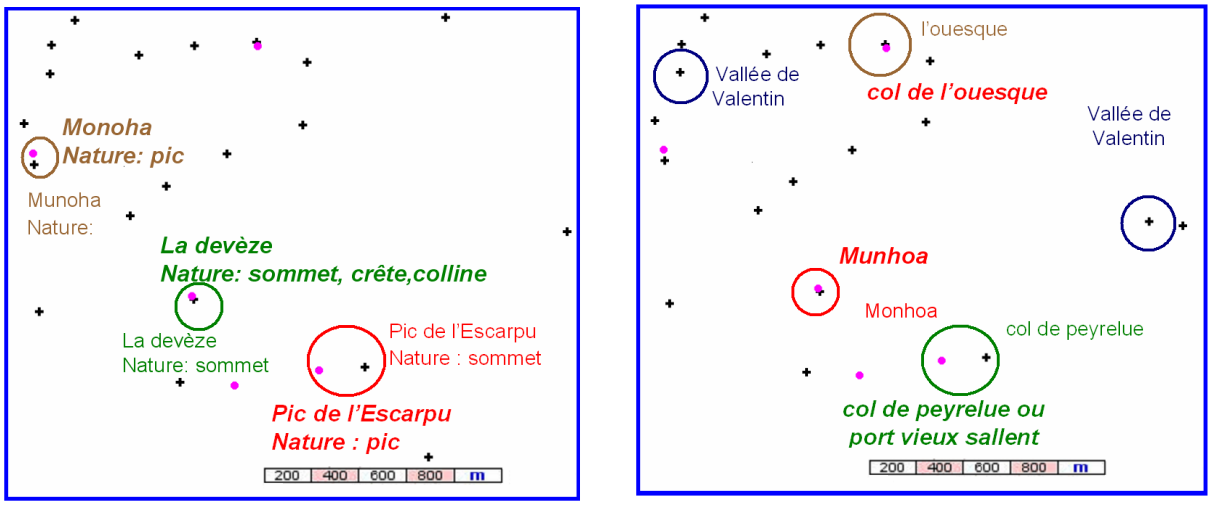


Figure 44. Imperfection dans les attributs des données géographiques (la nature de l'objet sur l'image à gauche et le toponyme sur l'image à droite)

Sur la Figure 44 à droite, nous illustrons quelques imperfections présentes dans l'attribut « Toponyme ». Toujours sous l'hypothèse que les jeux de données possèdent cet attribut, des objets géographiques homologues peuvent avoir des toponymes différents. Une des raisons peut être la différence de prononciation, par exemple « Munhoa » et « Monhoa ». L'imprécision dans cet exemple est illustrée par le mot « ou », le toponyme d'un objet de JD_1 étant « col de peyrelue ou port vieux sallent ». L'incomplétude peut aussi être présente, soit par le manque du toponyme, soit par le fait qu'un toponyme n'est pas complet : l'objet de JD_2 porte le toponyme « l'ouesque » alors que son homologue dans JD_1 s'appelle « col de l'ouesque ». Une autre spécificité des données géographiques est le fait qu'au sein d'un même jeu de données, deux objets distincts portent le même toponyme, par exemple « Vallée de Valentin ».

Nous pouvons remarquer à travers la partie B.1 qu'en général, l'incomplétude et l'imprécision peuvent avoir des formes à la fois objectives (des propriétés appartenant aux données) et subjectives (des interprétations d'une connaissance) tandis que l'incertitude est subjective et elle apparaît lorsque l'observation d'un expert doit être prise en compte. Ainsi, on pourrait dire que l'incertitude qualifie la relation entre les données elles-mêmes et les connaissances du monde réel.

B.2 Représentation de l'imperfection en utilisant la théorie des fonctions de croyance

Dans cette partie nous exposons d'abord notre motivation du choix de la théorie des fonctions de croyance en vue de définir une nouvelle approche d'appariement de données, ensuite nous présentons brièvement quelques applications basées sur la théorie des fonctions de croyance. Enfin, nous présentons le cadre général de la théorie des fonctions de croyance.

B.2.1 Motivation du choix de la théorie des fonctions de croyance

Nous avons vu dans le chapitre A que l'appariement de données géographiques est basé principalement sur des critères utilisant la géométrie et les relations spatiales, mais qu'il existe des cas où la géométrie utilisée seule n'est plus suffisante, d'autres critères moins discriminants étant nécessaires pour prendre une décision finale.

Comme nous l'avons dit précédemment, notre objectif est de mettre en œuvre un processus d'appariement de données géographiques basé sur une approche générique et qui soit guidé par des connaissances, c'est-à-dire qui les représente explicitement.

L'enjeu est de représenter explicitement les connaissances que nous avons sur des faits réels et de raisonner sur ces connaissances afin de les utiliser pour prendre une décision finale. D'une manière générale, certaines connaissances se présentent sous la forme de règles, tandis que d'autres sont parfois difficiles à modéliser explicitement.

Dans le processus d'appariement, afin de pouvoir décider si deux objets sont homologues ou pas, nous comparons différents attributs des deux objets et nous exprimons les résultats des comparaisons par des mesures de distance. Ainsi, si par exemple nous comparons leur localisation, nous utilisons une distance euclidienne, si nous comparons leur nature, nous utilisons une distance sémantique qui nous indique si les natures des deux objets sont proches conceptuellement parlant, ou encore si nous comparons leurs toponymes, nous employons une distance toponymique entre les deux chaînes de caractères.

Les résultats de ces comparaisons n'ont pas la même fiabilité puisqu'ils ne représentent pas les mêmes connaissances. De plus, ils peuvent être imprécis car ils dépendent de la manière dont nous avons décidé de calculer les distances. Un autre aspect concerne l'absence d'une information dans les données. Par exemple l'écart de position est discriminant mais parfois imprécis, la sémantique est en générale précise mais moins discriminante et peu comporter des erreurs. Pour les autres attributs, la discrimination dépend du type de l'attribut, mais ils sont parfois incomplets. Que se passe-t-il, par exemple, lorsque quelques objets ne possèdent pas de nom ? Comment gérer cette incomplétude dans les données ?

D'autres connaissances utiles dans le processus d'appariement sont les seuils. En effet, que se soit pour sélectionner les candidats à l'appariement ou pour éliminer les candidats qui ne sont pas fiables, les seuils jouent un rôle important dans le processus. Une question à laquelle nous devons répondre est : Comment fixons-nous ces seuils pour qu'ils soient fiables et comment rendre le processus peu sensible aux seuils?

Nous cherchons également à traiter les connaissances fournies par des experts, donc riches, subjectives et souvent en langage naturel. Ces connaissances permettent par exemple de définir la croyance qu'un expert a dans une proposition ou d'exprimer des conditions nécessaires mais pas toujours suffisantes. Par exemple, pour que deux objets soient homologues, ils doivent être de la même nature, mais tous les objets de la même nature ne sont pas appariés à un objet donné.

En se basant sur des faits ou sur d'autres connaissances telles que les seuils ou les distances, les experts raisonnent afin de prendre la bonne décision. Mais il peut s'avérer que dans certains cas spécifiques l'expert ne possède pas certaines connaissances, soit à cause de l'imperfection dans les données et dans les connaissances, soit à cause d'une situation ambiguë. Par exemple, un candidat n'est ni très proche de l'objet à apparier, ni très loin. Quelle décision pouvons-nous prendre dans ce cas ? Dans cette situation l'expert a un doute, il ne peut se prononcer avec certitude ni sur le fait que les deux objets sont homologues ni sur

le fait qu'ils ne le sont pas. Afin de ne pas prendre une mauvaise décision, il préfère dire que dans ce cas nous ne pouvons pas trancher. Par conséquent, le besoin d'exprimer le doute ou l'ignorance nous semble nécessaire dans le processus d'appariement.

La question qui se pose est : comment représenter explicitement le doute ? Par quel formalisme ?

Nous avons vu précédemment que les connaissances ne sont pas parfaites, pouvant être imprécises, incertaines ou incomplètes. De plus, les connaissances proviennent de sources différentes ayant des niveaux de fiabilité différents ; elles expriment des conditions nécessaires mais pas suffisantes ; ou encore elles sont exprimées dans des formats hétérogènes.

Les connaissances qui peuvent être utilisées dans le processus d'appariement sont nombreuses et hétérogènes. Nous nous sommes bornés dans notre processus d'appariement à trois types de connaissances. Elles proviennent des données elles-mêmes (les distances), des spécifications (les seuils) et des experts (les croyances en des propositions).

En conséquence des questions se posent telles que : Quelles sont les connaissances qui pourront guider le processus d'appariement de données géographiques ? Comment raisonner sur ces connaissances ? Comment représenter explicitement des connaissances différentes dans le même formalisme ? Comment représenter explicitement dans le même formalisme les imprécisions, les incertitudes et les incomplétudes ? Quelle confiance accordons-nous à ces connaissances ?

Pour renforcer la certitude que nous avons sur la décision ou encore pour remédier aux connaissances manquantes, nous avons besoin de fusionner d'une manière pertinente ces connaissances.

La représentation explicite des connaissances pour l'appariement est illustrée à titre illustratif à travers en exemple en fin de cette partie.

Afin de répondre à notre objectif, nous avons choisi la théorie des fonctions de croyance pour de nombreuses raisons :

- elle permet de prendre en compte et de modéliser à la fois l'imprécision, l'incertitude et l'incomplétude.
- elle permet de modéliser la connaissance parfaite, partielle et l'ignorance totale,
- elle permet de représenter plusieurs types de connaissance, ce qui offre un cadre riche et flexible,
- en termes de fusion des connaissances, elle possède des outils qui permettent de combiner plusieurs avis, c'est-à-dire de combiner plusieurs masses de croyance pour avoir, à la fin de l'étape de combinaison, une nouvelle distribution des masses qui permet ensuite de prendre une décision,
- elle permet de mettre en évidence et de gérer le conflit entre les connaissances,
- elle possède des outils qui permettent la prise de la décision.

Les avantages que nous venons d'énumérer semblent justifier l'utilisation de la théorie des fonctions de croyance vis-à-vis de notre besoin et ils sont illustrés à travers l'exemple suivant.

Exemple

Soit un objet O_1 appartenant au jeu de données JD_1 dont nous cherchons l'objet homologue dans le jeu de données JD_2. Considérons qu'après une sélection des candidats, nous avons gardé trois objets qui sont des homologues potentiels, C_1, C_2 et C_3, comme illustré sur la Figure 45. Supposons que le problème consiste à modéliser la connaissance concernant le choix parmi les trois candidats à l'appariement, exprimée par un expert. Ainsi, la solution à notre problème se trouve dans le cadre de discernement $\Theta=\{C_1, C_2, C_3\}$ ou éventuellement il peut s'avérer qu'aucun des trois objets ne soit l'homologue de O_1.

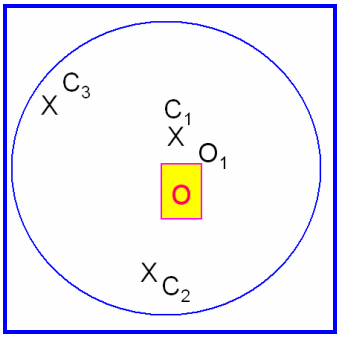


Figure 45. Sélection des candidats à l'appariement

Supposons que l'expert donne son avis par rapport à différentes mesures qu'il peut définir entre l'objet O_1 et chacun des candidats. Considérons dans cet exemple, que l'expert compare l'objet à apparier et ses candidats en utilisant leur toponyme et plus précisément la distance toponymique entre les toponymes.

Si l'objet O_1 ne possède pas de toponyme, l'expert ne peut pas se prononcer sur les trois candidats. Dans ce cas, l'expert attribue toute la masse de croyance au cadre de discernement $m(\Theta) = 1$, c'est-à-dire aucune masse de croyance (non-nulle) n'intervient, ni en faveur d'une hypothèse ni en sa défaveur. Ainsi, l'incomplétude est représentée formellement par l'ignorance. Nous remarquons à travers cette représentation la différence entre la théorie des probabilités et la théorie des fonctions de croyance. En effet, dans le cadre de la théorie des probabilités, cette connaissance est représentée en attribuant une probabilité égale aux trois hypothèses $P(C_1) = P(C_2) = P(C_3) = 0{,}33$. Le concept d'équi-probabilité n'est pas identique au concept d'ignorance totale. La même connaissance est représentée dans le cadre de la théorie des possibilités par $\Pi(C_1) = \Pi(C_2) = \Pi(C_3)=1$, c'est-à-dire que la possibilité que le candidat C_1 soit l'homologue de l'objet O_1 est totale (elle vaut 1) et elle est égale aux possibilités de chacun des autres candidats.

Supposons maintenant que les candidats C_1 et C_2 possèdent chacun un toponyme très ressemblant au toponyme de l'objet O_1. Par conséquent, la distance toponymique d_T (toponyme O_1, toponyme C_1) est très proche de la distance toponymique (toponyme O_1, toponyme C_2). Dans ce cas, il y a une imprécision. L'expert n'ayant pas plus d'information attribue la masse de croyance à l'union des deux candidats, $m(\{C_1, C_2\}) = 1$. Cela signifie que l'objet O_1 peut être apparié soit avec le candidat C_1 soit avec le candidat C_2.

Si le candidat C_2 possède le même toponyme que l'objet O_1, ce qui implique que la distance toponymique est égale à 0, l'expert peut représenter explicitement sa connaissance en attribuant la totalité de la masse de croyance à l'hypothèse C_2. Ainsi, il exprime sa connaissance parfaite, $m(C_2)=1$.

Supposons maintenant que les trois candidats possèdent des toponymes différents. L'expert sait que dans un jeu de données les toponymes représentent les noms d'usage et que dans l'autre jeu de données ils représentent les noms officiels. Il sait également que le toponyme de l'objet O_1 ressemble plus au toponyme du candidat C_1 qu'aux toponymes de C_2 et C_3. Ainsi, nous sommes dans une situation incertaine. Dans le cadre de la théorie des fonctions de croyance il est possible de représenter explicitement cette incertitude par les masses de croyance partielle dans les hypothèses. Afin de formaliser l'incertitude, l'expert peut intégrer d'une part sa propre connaissance et d'autre part la connaissance issue des distances toponymiques. Ainsi, il attribue à chacune des hypothèses une masse de croyance de la manière suivante :

$$\begin{cases} m_1(C_1) = 0{,}6 \\ m_1(C_2) = 0{,}2 \\ m_1(C_3) = 0{,}2 \end{cases} \quad (12)$$

Une décision prise en prenant en compte seulement des connaissances issues de la comparaison des toponymes n'est pas fiable. En effet, nous avons besoin d'intégrer d'autres connaissances afin de remédier au manque de connaissance, à l'imprécision et à l'incertitude. Par conséquent, nous pouvons faire appel à la fusion des connaissances. Dans ce contexte, continuons notre exemple.

Supposons que l'expert utilise une nouvelle source d'information basée sur la distance euclidienne entre l'objet O_1 et chacun de ses candidats à l'appariement. Il croit également que plus un candidat est proche de l'objet O_1, plus il y a de chances que les deux objets soient homologues et plus un candidat est loin, plus cette hypothèse est impossible. Afin de formaliser les connaissances de cette source d'information, il prend en compte à la fois les distances euclidiennes calculées et ses propres connaissances.

Si le candidat C_1 est très proche, le candidat C_2 n'est ni très loin ni très proche et le candidat C_3 est très loin. L'expert définit donc le jeu de masses de la manière suivante :

$$\begin{cases} m_2(C_1) = 0{,}9 \\ m_2(C_2) = 0{,}1 \\ m_2(C_3) = 0 \end{cases} \quad (13)$$

Les deux sources d'information peuvent être ensuite fusionnées en utilisant la combinaison conjonctive. Le jeu de masses résultant est le suivant :

$$\begin{cases} m_{12}(C_1) = m_1(C_1) * m_2(C_1) = 0{,}54 \\ m_{12}(C_2) = m_1(C_2) * m_2(C_2) = 0{,}02 \\ m_{12}(C_3) = 0 \\ m_{12}(\phi) = m_1(C_1)[m_2(C_2) + m_3(C_3)] + m_1(C_2)[m_2(C_1) + m_3(C_3)] + m_1(C_3)[m_2(C_2) + m_3(C_3)] \\ \qquad = 0{,}44 \end{cases} \quad (14)$$

Sous l'hypothèse du monde fermé (voir la partie B.4), $m_{12}(\phi)$ représente le conflit entre les sources d'information et est utilisée pour normaliser le jeu de masses résultant de la fusion des deux sources d'information.

Par conséquent, nous obtenons le jeu de masses suivant :

$$\begin{cases} m_{12}(C_1) = 0{,}96 \\ m_{12}(C_2) = m_1(C_2) * m_2(C_2) = 0{,}04 \\ m_{12}(C_3) = 0 \\ m_{12}(\phi) = 0 \end{cases} \quad (15)$$

Dans cet exemple, le jeu de masses est composé uniquement d'hypothèses singleton, ce qui fait que les fonctions de crédibilité, de plausibilité et de probabilité pignistique sont égales.

$$\begin{cases} Cr(C_1) = Pl(C_1) = P(C_1) = 0{,}96 \\ Cr(C_2) = Pl(C_2) = P(C_2) = 0{,}04 \\ Cr(C_3) = Pl(C_3) = P(C_3) = 0 \end{cases} \quad (16)$$

En se basant sur l'équation (21) et en maximisant la crédibilité accordée à chaque candidat, l'expert décide que l'objet homologue de l'objet O_1 est le candidat C_1 (voir Figure 46).

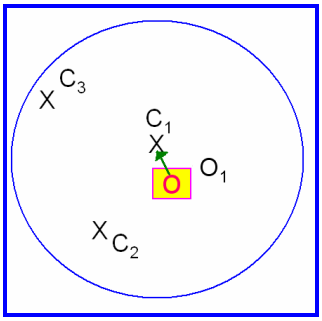


Figure 46. Résultat d'appariement basé sur la théorie des fonctions de croyance

B.2.2 Aperçu sur les théories de l'incertain

Les théories mathématiques et les techniques capables de modéliser, de gérer et de contrôler des données imparfaites, font l'objet de plusieurs travaux dans la littérature. Les théories employées sont nombreuses : citons la théorie des probabilités, la théorie des possibilités [Zadeh, 1978 ; Dubois et Prade, 1985], la théorie des sous-ensembles flous [Zadeh, 1965 ; Bouchon-Meunier, 1995], la théorie des fonctions de croyance [Dempster, 1967], la théorie des ensembles grossiers [Pawlak, 1982 ; Pawlak, 1991 ; Duckham *et al.*, 2001]. Une description complète des théories de l'incertain se trouve dans [Cohen *et al.*, 1987; Dubois et Prade, 1985 ; Bouchon-Meunier, 1995].

Nous présentons dans l'Annexe 1 un aperçu des autres théories faisant partie de la même famille des théories de l'incertain : la théorie des probabilités, la théorie des ensembles flous, la théorie des ensembles grossiers et la théorie des possibilités.

Plusieurs travaux consacrés à comparer les théories mathématiques ont été réalisés dans la littérature. Il semblerait qu'elles ne soient pas contradictoires, et le choix de l'utilisation dépend de l'application visée.

La théorie des sous-ensembles flous [Zadeh, 1965] permet de représenter des connaissances exprimées en langage naturel (« grand », « petit », « environ », etc.) présentant des imperfections et de réaliser une interface entre les connaissances symboliques et les descriptions numériques. Cependant, elle ne permet pas de modéliser les incertitudes. Or,

comme d'ailleurs Bouchon-Meunier l'affirme, l'imprécision et l'incertitude sont fortement liées [Bouchon-Meunier, 1995]. Par conséquent, Zadeh a introduit la théorie des possibilités [Zadeh, 1978].

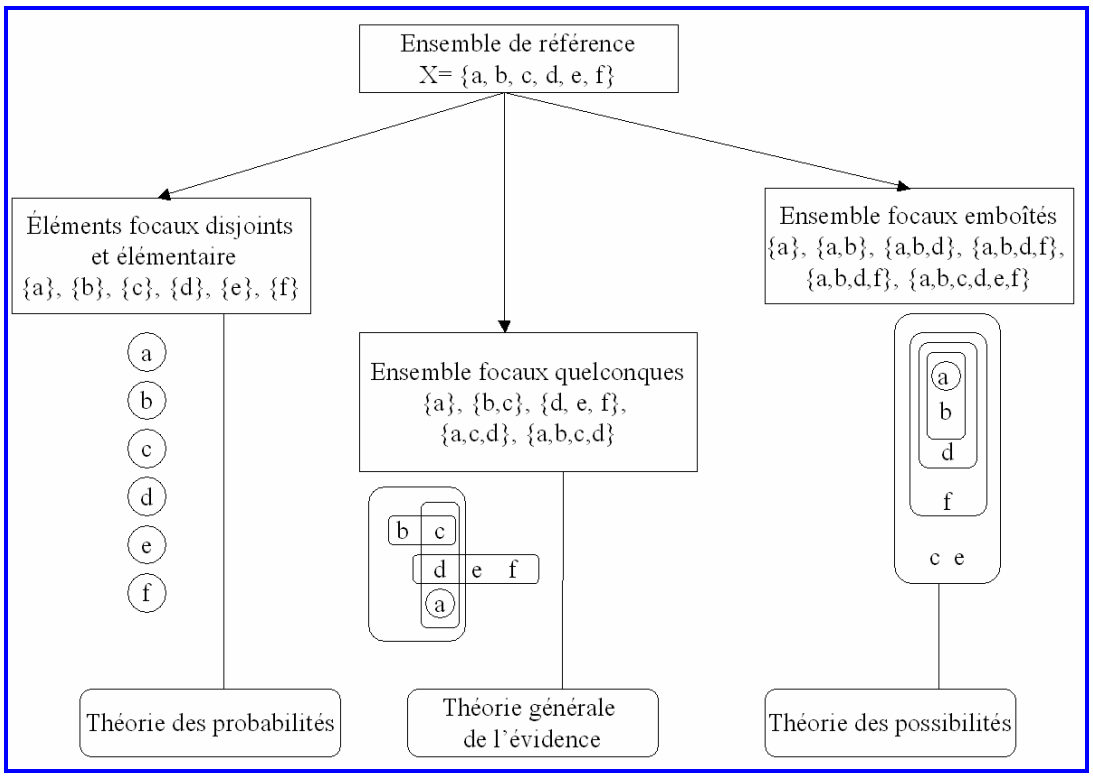


Figure 47. Position de la théorie des possibilités et de la théorie des probabilités dans le cadre de la théorie de l'évidence, d'après [Bouchon-Meunier, 1995], page 89.

Nous pouvons dire que la théorie des possibilités est utilisée lorsque les connaissances concernant la réalisation des événements sont insuffisantes. Elle permet de modéliser l'incertitude, l'imprécision et l'incomplétude. Cependant la représentation explicite de l'incomplétude est moins fine que dans le cadre de la théorie des fonctions de croyance. En effet, la nécessité est nulle et la possibilité est égale à 1.

Bouchon-Meunier montre que la théorie des fonctions de croyance n'est pas incompatible avec les autres théories [Bouchon-Meunier, 1995]. La théorie des probabilités et la théorie des possibilités sont des cas particuliers de la théorie des fonctions de croyance, lorsque les éléments focaux sont respectivement disjoints et élémentaires (théorie des probabilités) ou emboîtés (théorie des possibilités) (voir Figure 47).

B.3 Quelques applications de la théorie des fonctions de croyance

Grâce à ses nombreux avantages, la théorie des fonctions de croyance a montré son utilité dans de nombreux domaines d'application tels que l'analyse de données, le traitement d'image, le diagnostic ou l'aide à la décision ainsi que dans le domaine de l'information géographique.

Nous pouvons remarquer qu'il y a une interaction entre les domaines que nous venons d'énumérer. Des problématiques liées à la classification, à la reconnaissance des formes ou à

la fusion d'informations sont rencontrées dans plusieurs domaines. En conséquence, une classification rigoureuse des applications selon leur domaine d'utilisation nous semble difficile à réaliser.

Dans cette partie, nous donnons un aperçu des travaux de recherche qui utilisent la théorie des fonctions de croyance.

B.3.1 Analyse de données

L'analyse de données concerne par exemple la reconnaissance des formes ou la classification. D'une manière générale, la problématique consiste à assigner un vecteur x à une classe C_i en utilisant un ensemble d'apprentissage contenant N exemples. Le modèle le plus utilisé actuellement pour résoudre ce problème avec la théorie des fonctions de croyance est celui proposé par [Denœux, 1995], connu sous le nom d'*algorithme des k plus proches voisins*. Il est basé sur les fonctions de croyance définies à partir de mesures de distance effectuées entre un vecteur x à classifier et ses k plus proches voisins. La décision est prise en combinant les k fonctions de croyance au moyen de règles de combinaison proposées dans le cadre de la théorie des fonctions de croyance. Une extension de l'*algorithme des k plus proches voisins* a été proposée par [Wang et Bell, 2004] dans le but d'optimiser le processus de classification. L'extension consiste à considérer tous les voisins du vecteur x à classifier comme faisant partie d'une seule source d'information. Ainsi, le classifieur crédibiliste n'utilise pas de règle de combinaison pour agréger les fonctions de masse de croyance.

Au sujet des données relationnelles, [Masson, 2005 ; Denœux et Masson, 2003] s'intéressent à leur classification automatique en se basant sur des mesures de similarité entre les objets. Les données sont classifiées en groupes ou en classes homogènes dans le but d'obtenir un résumé des données. L'incertitude sur l'appartenance d'un objet obj_K aux différentes classes est modélisée par les fonctions de masses partielles définies sur le cadre de discernement.

Le diagnostic consiste à identifier un problème, une maladie ou même une nouvelle espèce par rapport à une taxonomie ou une ontologie.

Dans le cadre de l'aide à la conduite, [Lauriette-Rougegrez, 2006] définit un système qui prend en compte et formalise les incertitudes grâce aux fonctions de masse de croyance et qui permet de détecter la réalisation des manœuvres, en l'occurrence le dépassement.

La théorie des fonctions de croyance est employée par [Diaz *et al.*, 2007] pour enrichir les QCM (Questionnaires à choix multiples) en définissant des QCM « évidentiels ». Par rapport aux QCM classiques où l'apprenant (un étudiant par exemple) doit prendre absolument une décision parmi les choix donnés même s'il a un doute, les QCM évidentiels permettent aux étudiants de répondre aux questions en qualifiant l'état de leurs connaissances, ainsi que celui de leur ignorance. Ainsi, grâce à des distributions de masses de croyance issues des réponses aux QCM évidentiels, un diagnostic de l'état des connaissances des étudiants plus proche de la réalité peut être réalisé.

Les fonctions de croyance sont employées également dans le domaine du traitement de signal pour la classification parole/musique [Mauclair et Pinquier, 2004]. A partir des passages enregistrés, la classification parole/musique consiste à détecter les endroits où il n'y a que de la parole, que de la musique ou de la parole et de la musique en même temps.

Dans le domaine de l'aide à la décision, l'objectif consiste à choisir la meilleure solution ou la solution optimale parmi un ensemble de solutions. L'alternative de type booléenne *oui-*

non n'étant plus suffisante, de nombreux auteurs ont fait appel aux fonctions de croyance pour gérer l'imperfection afin d'avoir la décision la plus fiable possible.

La théorie des fonctions de croyance a été utilisée également pour résoudre des problèmes environnementaux. [Vannorenberghe *et al.*, 2000] ont modélisé et fusionné des informations imparfaites issues de mesures de pollution afin de prédire une situation anormale, c'est-à-dire l'apparition d'un pic de pollution, à l'aide des fonctions de croyance. Ces dernières sont employées pour mettre en place des modèles climatiques permettant de prédire la formation des ouragans [Poroseva *et al.*, 2006], la couverture des sols [Corgne, 2004] ou les risques sismiques [Rohmer, 2007].

Un outil d'aide à la décision dans le domaine de la gestion environnementale a été proposé par [Omrani *et al.*, 2007] afin d'évaluer les impacts environnementaux liés à la mobilité urbaine. L'outil permet de formaliser les informations prévenant des experts à l'aide de la logique floue et de les fusionner à l'aide de la théorie des fonctions de croyance.

B.3.2 Traitement d'image

La théorie des fonctions de croyance a fait l'objet de nombreux travaux dans le domaine du traitement d'image. Les problématiques sont liées à la segmentation des images ou à la classification des objets dans l'image. Généralement en traitement d'image, les sources d'information sont les plans de couleur. Par exemple, pour les images couleur, les plans rouge, vert, bleu sont utilisés alors que pour les images satellitaires ou aériennes, les plans rouge, vert, bleu et infra-rouge définissent les sources d'information.

La classification des images consiste soit à associer à chaque vecteur de l'ensemble des objets x_i une classe parmi un ensemble d'apprentissage de N classes (approche supervisée), soit à classer chaque vecteur en lui attribuant des étiquettes en fonction de différentes règles, puis à répéter les mêmes opérations jusqu'à ce que les étiquettes des vecteurs ne changent plus (approche non-supervisée).

Le processus de segmentation consiste à diviser l'image en zones homogènes afin d'extraire des informations utiles pour la reconnaissance des objets présents dans l'image. La redondance et la complémentarité des informations sont utilisées à travers la fusion de données afin de remédier aux imperfections.

Afin d'améliorer la fiabilité de la détection des défauts sur les clichés radiologiques de soudures, [Dupois, 2000] définit un cadre de discernement exhaustif contenant tous les défauts et ensuite une fusion de deux techniques de détection représentées à l'aide de fonctions de croyance est mise en place.

[Lemeret *et al.*, 2004] ont réalisé un simulateur utilisant un Laser à balayage et un Radar pour la détection et le suivi d'obstacles dans une scène. Le simulateur prend en compte pour chaque point détecté à chaque pas de balayage différentes mesures telles que la distance et l'angle, qui sont fusionnées grâce à un algorithme de fusion basé sur la théorie des fonctions de croyance. D'autres travaux liés aux problèmes du suivi des cibles faisant appel aux fonctions de croyance ont été réalisés [Appriou, 1999 ; Rombaut, 1998 ; Ramasso *et al.*, 2007].

De nombreux travaux liés à la segmentation et à la classification des images font appel aux fonctions de croyance, que se soit sur des photographies [Vannoorenberghe *et al.*, 2003 ; Vannoorenberghe et Flouzat, 2006 ; Mathevet *et al.*, 1999 ; Faux et Luthon, 2007], sur des images médicales [Colot, 2000 ; Capelle, 2003 ; Bloch, 1996 ; Dupuis, 2000], des images

sonars [Martin, 2005 ; Maussang *et al.*, 2008] ou bien sur des images satellitaires ou aériennes [Rottensteiner *et al.*, 2005 ; Saber Naceur *et al.*, 2000 ; Chitroub, 2004].

B.3.3 Géomatique

Dans le domaine de l'information géographique, la plupart des travaux qui sont basés sur la théorie des fonctions de croyance concernent la classification des images satellitaires ou aériennes et la classification de l'occupation des sols. A notre connaissance, il existe très peu de travaux qui utilisent la théorie des fonctions de croyance et qui traitent des données vecteur dans la littérature [Royère, 2002 ; El Najar, 2003].

Une application de la théorie des fonctions de croyance qui utilise des données géographiques vecteur concerne l'aide à la conduite automobile. Ainsi, afin de localiser un véhicule sur une route à partir d'une position estimée, [Royère 2002 ; El Najar, 2003] utilisent un capteur de perception et une base de données géographiques vecteur. L'approche s'appuie sur la notion des sources spécialisées, c'est-à-dire que chaque source se spécialise dans une hypothèse, dans leur cas un tronçon de route candidat à la localisation du véhicule. Les fonctions de croyance sont définies à partir des mesures issues de deux critères géométriques, un critère de proximité et un critère de colinéarité au cap du véhicule. Afin de prendre une décision, deux approches de fusion ont été mises en place : une approche qui combine les sources d'information par candidat, et une approche qui combine d'abord les sources par candidat, puis les candidats entre eux. La modélisation réalisée permet de prendre en compte les imperfections des capteurs de perception ainsi que les imperfections dans les données géographiques intégrées dans un GPS. Cette approche a été pour nous une source d'inspiration dans le processus d'appariement de données géographiques.

La classification des images satellitaires ou aériennes rentre dans le domaine du traitement d'image dont nous avons exposé les problématiques ci-dessus.

Les approches basées sur la connaissance experte sont très appropriées à la classification de l'occupation du sol. Dans ce cas, le problème est lié à la difficulté d'assigner un objet à une classe justement à cause de l'ambiguïté. La théorie des fonctions de croyance permet par exemple de fusionner plusieurs informations provenant des images satellitaires pour mieux extraire les types d'occupation du sol [Saber Naceur *et al.*, 2000 ; Le Hégarat-Mascle *et al.*, 2003] ou de modéliser les connaissances qu'un expert peut avoir sur la relation entre les classes de deux jeux de données différents [Comber *et al.*, 2005b; Comber, 2007].

B.4 Cadre général de la théorie des fonctions de croyance

Dans cette partie nous présentons le cadre général de la théorie des fonctions de croyance, connue également sous le nom de théorie de l'évidence ou de théorie de Dempster-Shafer. Elle a été introduite par Shafer en 1976 à la suite des travaux de Dempster sur les probabilités inférieure et supérieure [Dempster, 1967]. La théorie des fonctions de croyance [Shafer, 1982 ; Shafer, 1987] permet de modéliser les connaissances en utilisant deux mesures floues, une mesure de crédibilité et une mesure de plausibilité. Ces mesures sont appelées fonctions de croyance et elles se placent dans un cadre plus général que la théorie des possibilités et la théorie des probabilités.

B.4.1 Représentation explicite des connaissances

La théorie des fonctions de croyance est illustrée tout au long de cette partie par un exemple d'appariement de données géographiques.

Dans le cadre de la théorie des fonctions de croyance on considère un univers de référence appelé le cadre de discernement Θ, composé d'un ensemble de N hypothèses supposées répondre à un problème donné :

$$\Theta = \{ H_1, H_2, H_3 ... H_N \} \quad (17)$$

Exemple

Considérons un objet géographique ponctuel A appartenant à un jeu de données JD_1, pour lequel nous cherchons un objet homologue dans le jeu de données JD_2. Supposons qu'après une étape de sélection, nous avons trois objets candidats : B, C et D (voir Figure 48).

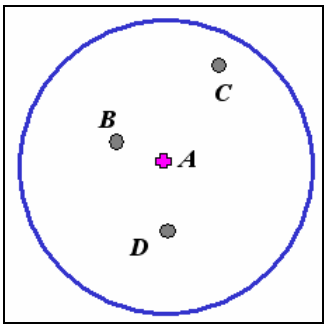


Figure 48. Sélection des objets candidats à l'appariement

Chaque candidat étant un potentiel objet homologue de l'objet A, le cadre de discernement peut être composé de trois hypothèses :AappB, « l'objet A est apparié avec le candidat B, AappC, « l'objet A est apparié avec le candidat C » et AappD, « l'objet A est apparié avec le candidat D ».

$$\Theta = \{ AappB, AappC, AappD \}$$

A partir du cadre de discernement, un référentiel de définition, noté 2^{Θ}, contenant tous les sous-ensembles de Θ, est défini :

$$2^{\Theta} = \{ \{ H_1 \}, \{ H_2 \}, \{ H_1, H_2 \} ... \{ H_1 ... H_{N-1} \}, \Theta \} \quad (18)$$

où le sous-ensemble $\{H_i, H_j\}$ représente la proposition P, que la solution au problème est une des deux hypothèses, H_i ou H_j.

Exemple

Dans notre exemple le référentiel de définition est le suivant :

$$2^{\Theta} = \{ AappB, AappC, AappD, \{AappB, AappC\}, \{AappB, AappD\}, \{AappD, AappC\}, \Theta \}$$

Par exemple, la proposition {AappC, AappD} signifie que l'objet A est apparié soit avec le candidat C soit avec le candidat D.

Un point clé de la théorie des fonctions de croyance réside dans la possibilité de représenter de manière explicite la croyance que nous avons en une hypothèse à travers les fonctions de croyance.

Une fonction de croyance peut être définie par une fonction de masse m_j, $m_j : 2^\Theta \rightarrow [0, 1]$ qui respecte les conditions suivantes :

$$\sum_{P \in 2^\Theta} m_j(P) = 1 \qquad (19)$$

Dans l'équation précédente, m_j (P) s'appelle masse de croyance et représente le degré de croyance avec lequel une source $S_{j,\ j=1\ldots M}$ croit en la proposition P. L'indice *j* représente donc la source d'information qui a défini la masse de croyance associée à la proposition P. Le degré de croyance représente le degré de précision d'une proposition par rapport à l'ensemble des connaissances d'un expert, à l'opposé du degré de vérité qui fait référence au degré de précision d'une proposition vis-à-vis la réalité.

Exemple

Dans notre exemple une source d'information peut être un critère d'appariement. Par exemple j=1 peut représenter un critère basé sur l'écart de position et j=2 signifie un critère basé sur la comparaison de la sémantique.

Le critère d'écart de position « croit », en analysant l'écart de position entre un objet et un candidat, que l'objet A est apparié avec :

-le candidat B, en attribuant à cette hypothèse une masse de croyance de 0,5 : m_1(AappB)=0,5 ;

-le candidat C, en attribuant à cette hypothèse une masse de croyance de 0,1 m_1(AappC)=0,1 ;

-le candidat D ou B, en attribuant à cette hypothèse une masse de croyance de 0,4 : m_1({AappD, AappB})=0,4.

De la même manière, le critère sémantique attribue à chaque hypothèse un degré de croyance en fonction de l'écart sémantique entre les objets.

La masse de croyance se différencie de la notion de probabilité par le fait que la totalité de la masse de croyance est répartie non seulement sur les hypothèses singletons, H_1, H_2, H_3, mais aussi sur les hypothèses combinées, par exemple $\{H_1, H_2\}$, $\{H_1, H_3\}$, $\{H_1, H_2, H_3\}$. Par conséquent, la masse de croyance présente une grande analogie avec la notion de distribution de probabilité, à la différence près que l'on répartit une masse unité parmi les éléments de 2^Θ. L'avantage est que l'on attribue à une hypothèse combinée une croyance en la réalisation de chacune d'entre elles sans prendre parti pour l'une d'elles précisément. Cette manière de définir une masse de croyance sur une union d'hypothèses, c'est-à-dire une proposition, permet une modélisation de l'ignorance partielle de l'information. Cela signifie que nous pouvons qualifier aussi une réponse imprécise. Ainsi, plus les hypothèses non singleton sont importantes, plus la réponse est imprécise.

Tous les sous-ensembles de la proposition P pour lesquels la masse de croyance est non nulle sont appelés *éléments focaux* et ils constituent le noyau N_{mj} de la structure de croyance, défini par :

$$N_{mj} = \left\{ P \in 2^\Theta / m_j(P) > 0 \right\} \qquad (20)$$

Lorsque les éléments focaux se réduisent aux seuls singletons H_i, la notion de masse élémentaire est assimilable à celle de probabilité. La structure de croyance devient une distribution de probabilité, si elle ne contient que des singletons.

Sous l'hypothèse du *monde fermé*, on considère que le cadre de discernement est exhaustif, c'est-à-dire que la solution au problème donné se trouve parmi les hypothèses définies dans le cadre de discernement, et que les hypothèses sont exclusives [Shafer, 1976]. Dans ce cas, l'ensemble vide ϕ désigne l'événement impossible. En revanche, sous l'hypothèse du *monde ouvert* défini par [Smets et Kennes, 1994] dans le modèle de croyance transférable, ϕ signifie que la solution au problème posé ne se trouve pas dans le cadre de discernement et il correspond à toutes les hypothèses omises dans le cadre de discernement, constituant une classe de rejet ou de conflit. Ainsi dans le cas du *monde ouvert*, les hypothèses sont considérées toujours exclusives, et la condition du *monde fermé* $m_j(\phi)=0$ imposée par [Shafer, 1976], est relâchée.

D'autres auteurs [Rombaut, 1998 ; Royère, 2002] ont introduit la notion de *monde ouvert étendu*. Dans le cadre du *monde ouvert étendu* une nouvelle hypothèse singleton est introduite qui regroupe toutes les hypothèses non définies dans le cadre de discernement. En conséquence, le cadre de discernement devient exhaustif, donc s'approche de l'hypothèse du *monde fermé*. Par contre lorsque la condition $m(\phi)=0$ est relâchée, ϕ est interprété comme une classe de conflit dû uniquement à la non-fiabilité des sources.

L'hypothèse du *monde fermé* est très courante dans les applications dont on connaît l'ensemble des hypothèses possibles, par exemple en diagnostic, alors que l'hypothèse du *monde ouvert* est plutôt adaptée dans des applications de reconnaissance de forme où il est parfois impossible d'envisager toutes les formes rencontrées.

Les fonctions de croyance permettent de modéliser et de formaliser la crédibilité attribuée à des propositions. A partir de la masse de croyance, on définit deux fonctions de croyance :

1. La fonction de crédibilité (voir la Figure 49)

La fonction de croyance, $Cr_j(P)$, mesure la force avec laquelle on croit en la véracité de la proposition P et elle est définie par [Bouchon-Meunier, 1995] comme étant la somme de toutes les masses de croyance des éléments focaux $P' \in 2^{\Theta}$ inclus dans P :

$$Cr_j(P) = \sum_{P' \subseteq P} m_j(P') \qquad (21)$$

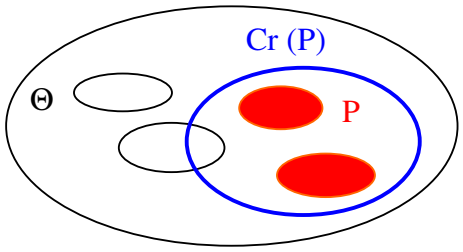


Figure 49. La fonction de crédibilité définie dans le cadre de discernement

2. La fonction de plausibilité (voir la Figure 50).

La fonction de plausibilité mesure l'intensité avec laquelle on ne doute pas de la proposition P, c'est-à-dire qu'elle est la somme des masses de croyance des éléments focaux

dont l'intersection avec P n'est pas nulle [Bouchon-Meunier, 1995]. Elle est définie par l'équation suivante :

$$Pl_j(P) = \sum_{P \cap P' \neq \phi} m_j(P') \qquad (22)$$

La fonction de plausibilité, $Pl_j(P)$, est une fonction duale de la fonction de crédibilité pouvant être également définie à partir de la fonction de crédibilité de l'événement contraire à P :

$$Pl_j(P) = 1 - Cr(\overline{P}) \qquad (23)$$

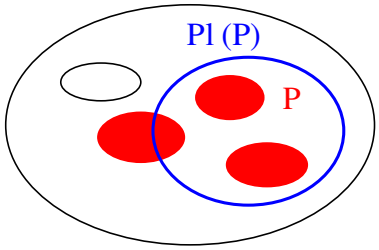


Figure 50. La fonction de plausibilité définie dans le cadre de discernement

La fonction de croyance est une mesure pessimiste, alors que la fonction de plausibilité est une mesure optimiste.

Dans le cadre de la théorie des fonctions de croyance, la croyance en une hypothèse peut être interprétée comme l'incertitude minimum sur l'hypothèse et la plausibilité comme l'incertitude maximum sur l'hypothèse. Ainsi, l'incertitude est représentée par l'ensemble des valeurs de l'intervalle [Cr, Pl] appelé intervalle de croyance et sa longueur représente l'imprécision.

Exemple

A partir des masses de croyance définies ci-dessus par le critère d'écart de position, nous pouvons déterminer les fonctions de crédibilité et de plausibilité pour chaque hypothèse.

$Cr_1(AappB) = m_1(AappB) + m_1(\{AappD, AappB\}) = 0,9$;

$Pl_1(AappB) = m_1(AappB) + m_1(\{AappD, AappB\}) = 0,9$;

$Cr_1(AappC) = m_1(AappC) = 0,1$;

$Pl_1(AappC) = m_1(AappC) = 0,1$;

$Cr_1(\{AappD ; AappB\}) = m_1(\{AappD, AappB\}) = 0,4$;

$Pl_1(\{AappD ; AappB\}) = m_1(AappB) + m_1(\{AappD, AappB\}) = 0,9$

B.4.2 Initialisation des masses de croyance

La principale difficulté dans la théorie des fonctions de croyance est l'initialisation des masses de croyance. Cette étape consiste à déterminer les masses de croyance qui serviront après pour prendre la décision finale. L'étape d'initialisation est primordiale puisque les résultats en dépendent. La mauvaise initialisation des masses de croyance peut être aussi une

source de conflit. Plusieurs méthodes d'initialisation des masses de croyance en fonction du problème à traiter existent dans la littérature.

Les jeux de masses peuvent être initialisés en utilisant les fonctions de vraisemblance construites à partir d'un ensemble de données d'apprentissage [Appriou, 1991, Smets, 1993]. [Colot, 2000] modélise les connaissances par des gaussiennes où la moyenne et l'écart-type sont également calculés sur un ensemble d'apprentissage.

Une approche des modèles très employée, basée sur des mesures de distances, est celle proposée par [Denœux, 1995], connue sous le nom d'*algorithme des k plus proches voisins*. Cette approche consiste à initialiser les jeux de masses pour un vecteur inconnu à partir des distances entre le vecteur inconnu et ses k plus proches voisins. La décision est prise sur un jeu de masses final issu de la combinaison des k jeux de masses définis pour les k voisins. Elle nécessite également un ensemble de données d'apprentissage.

Les matrices de confusion calculées sur un ensemble d'apprentissage permettent aussi d'initialiser un jeu de masses [Mercier, 2006 ; Comber *et al.*, 2005a]. En imagerie, les masses de croyance peuvent être initialisées à partir des histogrammes des niveaux de gris [Rombaut et Zhu, 2002].

Un autre modèle d'initialisation des masses de croyance repose sur des arbres de décision. Après l'apprentissage de l'arbre de décision, une fonction de croyance est construite pour chaque nœud de l'arbre [Denœux et Bjanger, 2000 ; Vannoorenberghe et Denœux, 2002].

Des grandeurs physiques telles que la distance, l'angle, et des représentations floues peuvent être utilisées pour l'initialisation des fonctions de croyance [Rombaut, 1998 ; Royère 2002 ; El Najjar, 2003].

B.4.3 Combinaison des sources d'information

Un autre avantage de la théorie des fonctions de croyance réside dans la possibilité de combiner les informations issues de sources d'informations distinctes. Ainsi, les informations sont fusionnées à l'aide de l'opérateur de Dempster, aussi appelé opérateur conjonctif. Celui-ci est le premier opérateur de combinaison défini dans le cadre de la théorie des fonctions de croyance [Dempster, 1967]. Son utilisation suppose que les sources à fusionner soient indépendantes. La masse fusionnée est notée de la manière suivante :

$$\forall P \in 2^{\Theta}, m_{1..M}(P) = m_1(P) \oplus m_2(P) \oplus ... \oplus m_M(P) \qquad (24)$$

où M représente le nombre de sources d'information.

L'opérateur de Dempster [Dempster, 1967] vérifie les propriétés suivantes :

- commutativité : $m_{12}(P) = m_1(P) \oplus m_2(P) = m_2(P) \oplus m_1(P)$;
- associativité : $(m_1(P) \oplus m_2(P)) \oplus m_3(P) = m_1(P) \oplus (m_2(P) \oplus m_3(P))$. Cette propriété est très importante lorsque nous avons plus de deux sources à fusionner ;
- la somme est non idempotente : $m_1(P) \oplus m_1(P) \neq m_1(P)$, c'est-à-dire qu'un jeu de masses issu d'une source d'information combiné avec lui-même fournit un autre jeu de masses. La somme orthogonale $m_1(P) \oplus m_1(P)$ favorise les mêmes sous-hypothèses que $m_1(P)$, sauf que la croyance a un poids double.

La masse de croyance totale m_{12} attribuée à l'hypothèse P pour deux sources S_1 et S_2 est donnée par l'équation suivante :

$$\forall P \in 2^{\Theta},\ m_{12}(P) = \frac{1}{1 - m_{12}(\phi)} \sum_{P' \cap P'' = P} m_1(P') m_2(P''),\ P' \ et\ P'' \in 2^{\Theta} \qquad (25)$$

où $\frac{1}{1 - m_{12}(\phi)}$ est un coefficient de normalisation de l'opérateur de Dempster et $m(\phi)$ représente le conflit entre les deux sources, c'est-à-dire qu'il exprime le fait que les hypothèses ne sont pas compatibles (l'intersection entre les sous-hypothèses de l'hypothèse P est vide) :

$$m_{12}(\phi) = \sum_{P' \cap P'' = \phi} m_1(P') m_2(P'') \qquad (26)$$

Si la masse de l'élément vide $m(\phi)$ est égale à 1, les deux sources sont en conflit total et en conséquence la fusion ne peut pas avoir lieu. A l'inverse, lorsque la valeur de $m(\phi)$ est 0, les sources sont en accord total. De nombreux travaux ont contesté l'opérateur de Dempster parce qu'il ne gère pas le conflit, la masse attribuée à l'ensemble vide étant utilisée uniquement pour normaliser le jeu de masses. [Zadeh, 1979] a démontré que ce coefficient de normalisation est très sensible, au voisinage de $m(\phi)=1$, aux petites variations des deux jeux de masses. Malgré ses limites, l'opérateur de Dempster est utilisé grâce à ses propriétés que nous venons d'énumérer auparavant.

Exemple

Considérons que les critères d'écart de position et sémantique fournissent les jeux de masses suivants :

$$\begin{cases} m_1(AappB) = 0{,}5, \\ m_1(AappC) = 0{,}1, \\ m_1(\{AappD, AappB\}) = 0{,}4, \end{cases} \qquad \begin{cases} m_2(AappB) = 0{,}7, \\ m_2(\{AappC, AappD\}) = 0{,}3, \end{cases}$$

Ajoutons que les autres hypothèses du référentiel de définition ont des masses de croyance nulles.

Après la combinaison des masses de croyance issues des deux critères, nous obtenons le jeu de masses suivant :

$m_{12}(AappB) = m_1(AappB) * m_2(AappB) + m_2(AappB) * m_1(\{AappD, AappB\}) = 0{,}63$

$m_{12}(AappC) = m_1(AappC) * m_2(\{AappC, AappD\}) = 0{,}03$

$m_{12}(AappD) = m_1(\{AappD, AappB\}) * m_2(\{AappC, AappD\}) = 0{,}12$

$m_{12}(\phi) = m_1(AappB) * m_2(\{AappC, AappD\}) + m_1(AappC) * m_2(AappB) = 0{,}22$

B.4.4 Analyse et redistribution du conflit

Lors de la combinaison des sources d'information, un conflit entre les sources peut apparaître. Les causes du conflit peuvent être multiples : contradiction entre les informations soutenues par les sources (par exemple deux appariements différents), mauvaise définition du cadre de discernement (par exemple une mauvaise sélection des candidats à l'appariement), mauvaise initialisation des masses de croyance (par exemple une mauvaise pondération des critères d'appariement), etc. L'utilisation du conflit pour normaliser les jeux de masses révèle en fait l'ignorance du conflit, c'est-à-dire qu'aucune analyse du conflit n'est faite.

En conséquence, d'autres opérateurs de combinaison ont été proposés [Smets, 1990 ; Dubois et Prade, 1988; Yager, 1987 ; Lefevre *et al.*, 2000b] de façon à analyser et redistribuer le conflit, c'est-à-dire que la masse attribuée au conflit est redistribuée aux hypothèses du référentiel de définition.

[Smets 1990] considère, sous l'hypothèse de sources totalement fiables, que le conflit entre les sources, lorsqu'il existe, est dû à la non-exhaustivité du cadre de discernement, c'est-à-dire que la solution ne fait pas partie du cadre de discernement. Dans ce cas la masse attribuée à l'ensemble vide est vue comme une classe de rejet représentant les hypothèses non représentées dans le cadre de discernement. Celle-ci n'intervient pas dans le calcul de la masse fusionnée. En conséquence, l'opérateur de fusion est défini de la manière suivante :

$$\forall P \in 2^{\Theta}, m_{12}(P) = \sum_{P' \cap P'' = P} m_1(P') m_2(P'') \qquad (27)$$

Dans le cas de sources non-fiables, plusieurs opérateurs ont été proposés, permettant d'interpréter le conflit d'une manière différente. Ainsi, le conflit peut être réparti globalement [Dubois et Prade, 1988], ou localement [Lefevre *et al.*, 2002]. A titre d'illustration supposons que la masse associée au conflit soit $m_{12}(\phi) = m_1(H_1) * m_2(H_2)$ avec $H_1 \neq H_2$. Ainsi, selon [Dubois et Prade, 1988] la masse associée au conflit est attribuée à l'hypothèse $\{ H_1, H_2 \}$:

$m(\{ H_1, H_2 \}) = m_1(H_1) * m_2(H_2)$,

Selon la proposition de [Lefevre *et al.*, 2002], la masse attribuée au conflit est redistribuée d'une manière pondérée aux masses associées à chaque hypothèse impliquées dans le conflit :

$m(H_1) = w_1 * (m_1(H_1) * m_2(H_2))$,

$m(H_2) = w_2 * (m_1(H_1) * m_2(H_2))$,

où w_1 et w_2 représentent les poids associés à chaque hypothèse et calculés à l'aide des masses de chacune des hypothèses impliquées dans le conflit.

Lorsque le conflit apparaît, [Yager, 1987] propose de redistribuer le conflit sur le cadre de discernement, tandis que [Dubois et Prade, 1988] proposent deux alternatives de combinaison.

- la première alternative est une stratégie conjonctive-disjonctive, c'est-à-dire que si les sources sont en accord, la combinaison est faite en utilisant un opérateur conjonctif (la masse de croyance combinée des deux hypothèses est attribuée à leur intersection), et que si les sources sont en désaccord, la combinaison est faite en employant un opérateur disjonctif (la masse de croyance combinée est attribuée à l'union des hypothèses en conflit, dont l'intersection est égale à l'ensemble vide). Cette stratégie conjonctive-disjonctive fait que la masse de croyance associée au conflit, $m(\phi)$, est toujours égale à 0,
- la deuxième alternative est une stratégie basée sur une combinaison conjonctive suivie d'une redistribution du conflit. Ainsi, après la combinaison conjonctive, le conflit est

redistribué à l'union des propositions qui ont engendré le conflit. Cette stratégie de combinaison n'est pas associative.

[Lefevre *et al.*, 2002] proposent deux solutions de fusion basées sur une redistribution du conflit à l'aide de poids accordés à chaque proposition. Une première proposition consiste à distribuer la masse de conflit sur toutes les disjonctions de sous-hypothèses à partir des hypothèses composées ayant produit du conflit, à la différence de la deuxième proposition qui consiste à redistribuer la masse de croyance conflictuelle sur les hypothèses singletons ayant généré le conflit. L'inconvénient des deux opérateurs de fusion est qu'ils ne sont pas associatifs, nécessitant ainsi une stratégie de fusion définie a priori.

La redistribution du conflit a été traitée également par [Royère, 2002], qui propose deux méthodes. La première méthode est une méthode inspirée des propositions de [Dubois et Prade, 1988] mais qui a l'avantage par rapport à celle-ci d'être associative. Elle consiste à redistribuer les portions de masse de croyance composant la masse conflictuelle sur l'union des propositions. La deuxième méthode proposée consiste à redistribuer les portions de masses composant la masse conflictuelle sur l'union des hypothèses qui ont engendré le conflit. Le deuxième opérateur n'est associatif que dans certains cas.

Par exemple, supposons que nous avons une portion de masse conflictuelle :

$m(\phi)=m_1(H_1) * m_2 * (H_2) * m_3(\Theta)$

Le premier opérateur redistribue la masse conflictuelle sur l'union des hypothèses, c'est-à-dire Θ, alors que la masse conflictuelle est redistribuée par le deuxième opérateur sur l'union des hypothèses qui sont la cause du conflit, c'est-à-dire $\{H_1, H_2\}$.

[Denœux, 2006] définit une nouvelle règle de combinaison appelée *règle de combinaison prudente,* qui s'avère être commutative, associative et idempotente. Elle s'applique lors de la combinaison de sources fiables, étant basée sur le principe d'engagement minimal, c'est-à-dire que parmi les fonctions de masses d'un ensemble compatible avec un ensemble de contraintes, la fonction de masses la plus appropriée est celle qui est la moins informative. Grâce à la propriété d'idempotence, la condition sur l'indépendance des sources peut être relâchée.

B.4.5 Affaiblissement des sources

Une façon d'exprimer la fiabilité des sources d'information, et donc de diminuer le conflit, est l'affaiblissement des sources qui consiste à associer à chaque source d'information un coefficient d'affaiblissement. Cette solution s'applique uniquement si on connaît la source qui n'est pas fiable. Nous avons vu précédemment que parmi les causes qui engendrent le conflit il y a les sources d'information peu fiables. La notion de fiabilité peut être vue comme la capacité d'une source à rendre compte de la réalité qu'elle observe avec fidélité.

L'affaiblissement des sources consiste à accorder un coefficient de confiance $\alpha_j \in [0,1]$ à chaque source d'information $S_{j,(j=1..M)}$. Ainsi, le jeu de masses est pondéré par le coefficient de confiance α_j de la manière suivante :

$$\begin{cases} \forall P \in 2^{\Theta}, m_{\alpha,j}(P) = \alpha_j m_j(P) \\ m_{\alpha,j}(\Theta) = (1-\alpha_j) + \alpha_j m(\Theta) \end{cases} \quad (28)$$

où P est une proposition, c'est-à-dire une union d'hypothèses simples. Si α_j est égal à 0, la source j est considérée comme non fiable, alors que si α_j est égal à 1 la source est considérée comme fiable.

Concernant la détermination de ces coefficients d'affaiblissement, plusieurs travaux existent dans la littérature [Appriou, 1991 ; Lefevre *et al.*, 2000a].

B.4.6 Décision

Une fois les sources combinées, la théorie des fonctions de croyance offre de nombreux outils pour la prise de décision, tels que :

- le maximum de plausibilité :

  décision H_j : si $Pl(H_j) = \max(Pl(H_k), K_k=2^{\Theta})$ (29)

- le maximum de crédibilité :

  décision H_j : si $Cr(H_j) = \max(Cr(H_k), K_k=2^{\Theta})$ (30)

- le maximum de crédibilité sans recouvrement des intervalles de confiance [Bloch, 2005] :

  décision H_j : si $Cr(H_j) \geq \max(Pl(H_k), K_k=2^{\Theta}, k\neq j)$ (31)

- le maximum de probabilité pignistique :

  décision H_j : si $P(H_j) = \max(P(H_k), K_k=2^{\Theta})$ (32)

La fonction de probabilité pignistique est une mesure qui se trouve entre la mesure de crédibilité et la mesure de plausibilité. Elle est une fonction de probabilité, mais [Smets et Kennes, 1994 ; Smets, 1998] l'appelle fonction de probabilité pignistique parce qu'elle intervient au niveau de la prise de décision. Elle est définie de la manière suivante :

$$\forall A \in 2^{\Theta},\ P(A) = \sum_{A \subseteq B} m(B) \frac{|A \cap B|}{|B|} \frac{1}{1-m(\phi)} \tag{33}$$

où |B| représente le nombre d'hypothèses simples contenues dans la proposition B. Le terme $\frac{1}{1-m(\phi)}$ est utilisé pour la normalisation des probabilités pignistiques.

La probabilité pignistique a été définie dans le cadre du modèle de croyances transférables (TBM) [Smets et Kennes, 1994 ; Smets et Kruse, 1997]. Le TBM est un modèle utilisé pour quantifier les croyances à l'aide de fonctions de croyance. Selon [Smets et Kennes, 1994], il représente l'interprétation du modèle Dempster-Shafer. Les croyances peuvent être utilisées à deux niveaux, d'une part au niveau dit « crédal » lorsqu'elles sont représentées par des fonctions de croyance, et d'autre part au niveau dit « pignistique » lorsqu'elles sont utilisées pour prendre des décisions, ce dernier niveau étant quantifié par des fonctions de probabilité. [Smets et Kennes, 1994] montrent le lien entre les fonctions de croyance et les fonctions de probabilité lorsqu'une décision doit être prise. Ainsi, à ce niveau, les croyances du niveau crédal induisent une mesure de probabilité au niveau pignistique.

Le TBM est proche du modèle Dempster-Shafer, à quelques exceptions près telles que les notions de monde fermé et de monde ouvert, l'introduction de fonctions de croyance non-

normalisées, le modèle à deux niveaux et la transformation pignistique, le concept transférable de « part de croyance » et la dissociation complète d'utilisation de probabilité afin de calculer les fonctions de croyance.

B.5 Approches des sources spécialisées

[Appriou, 1991] propose deux modèles d'initialisation des masses de croyance, connus dans la littérature sous le nom de *sources spécialisées*. D'une manière générale, chaque source « se spécialise » sur une hypothèse du cadre de discernement, c'est-à-dire que la source d'information analyse l'hypothèse donnée et se prononce en sa faveur, en sa défaveur ou elle ne se prononce pas parce qu'elle n'a pas assez de connaissances sur cette hypothèse. Dans cette partie nous présentons brièvement ces deux modèles. Définis de manière axiomatique, ils ont été initialement définis pour la modélisation des données d'apprentissage statistique incertain.

Modèle 1 : un jeu de masses qualifie trois éléments focaux (H, ¬H, Θ). Une source d'information se spécialise et se prononce sur une hypothèse H en attribuant une masse de croyance à ces éléments focaux, chacun représentant une affirmation. Ainsi, à partir d'une mesure, l'hypothèse est évaluée à travers les trois propositions suivantes :

- l'hypothèse est vraie, H,
- l'hypothèse n'est pas vraie, ¬H,
- la source ne sait pas, Θ.

Pour chacune des hypothèses H_i et pour chacune des sources S_j, [Appriou, 1991] construit le jeu de masses noté m_{ij} de la manière suivante :

$$\begin{cases} m_{ij}(H_i) = d_{ij} * R_j * p(m_j / H_i) / [1 + R_j * p(m_j / H_i)] \\ m_{ij}(\neg H_i) = d_{ij} / [1 + R_j * p(m_j / H_i)] \\ m_{ij}(\Theta) = 1 - d_{ij} \end{cases} \quad (34)$$

Dans les équations précédentes, d_{ij} représente le coefficient qui caractérise la fiabilité des sources. Si ce coefficient est égal à 1, les jeux de masses ne sont pas affaiblis. A l'inverse s'il est égal à 0, nous sommes dans une situation de méconnaissance totale, c'est-à-dire que les densités de probabilités sont inconnues, et toute la masse de croyance est attribuée au cadre de discernement $m_{ij}(\Theta)=1$. Dans ce cas le jeu de masses $m_{ij}(\Theta)$ est ignoré puisqu'il devient élément neutre de la règle de combinaison de Dempster.

Le terme $p(m_j/H_i)$ représente la densité de probabilité de la mesure m_j issue de la source S_j sous l'hypothèse H_i. Les distributions de probabilité *a priori* sont issues d'un apprentissage dans l'approche d'Appriou.

Le facteur R_j est un facteur de normalisation contraint par :

$$R_j \in [0, \frac{1}{\max_{i\in[1,N]}(p(m_j / H_i))}] \quad (35)$$

Modèle 2 : un jeu de masses possède deux éléments focaux (¬H, Θ). Une source d'information se spécialise et se prononce sur une hypothèse H en attribuant une masse de croyance à ces éléments focaux, chacun représentant une affirmation. Ainsi, à partir d'une mesure, l'hypothèse est évaluée à travers les deux propositions suivantes :

- l'hypothèse n'est pas vraie, ¬H
- la source ne sait pas, Θ.

Le jeu de masses est construit comme ci-après :

$$\begin{cases} m_{ij}(H_i) = 0 \\ m_{ij}(\neg H_i) = d_{ij} * [1 - R_j * p(m_j / H_i)] \\ m_{ij}(\Theta) = 1 - d_{ij} + d_{ij} * R_j * p(m_j / H_i) \end{cases} \quad (36)$$

Les significations des notations utilisées pour définir le modèle 2 sont similaires à celles que nous avons expliquées pour le modèle 1.

D'autres travaux liés à l'initialisation des fonctions de masses de croyance et proches des modèles d'Appriou ont été menés. En fonction du nombre d'éléments focaux, les travaux de recherche peuvent être classés en quatre groupes :

- ceux qui utilisent les éléments focaux : (¬H, Θ) [Smets, 1993]. Ce modèle suit le même principe que le modèle 2 proposé par [Appriou, 1991]. Les deux modèles sont identiques lorsque le facteur de normalisation R_j est égal à 0. Ce modèle est utilisé lorsque l'information que nous avons sur une hypothèse est incomplète : on sait qu'il a y a des chances qu'elle ne soit pas vraie, mais pour ne pas la rejeter complètement, on attribue une fraction de masse de croyance à la masse de croyance attribuée au cadre de discernement.
- ceux qui utilisent les éléments focaux : (H, ¬H) [Barnett, 1981]. Ce modèle composé d'une hypothèse simple et de son contraire a été défini afin de diminuer la complexité algorithmique. Nous savons que plus il y a d'hypothèses dans le cadre de discernement, plus la complexité algorithmique augmente à cause de la combinaison de toutes les hypothèses. Remarquons que si l'hypothèse H et son contraire sont des hypothèses singleton, nous sommes dans le cas de la théorie des probabilités où on attribue une probabilité à une hypothèse et à son contraire. Dans ce type de modélisation, le conflit peut prendre des valeurs très importantes. En effet, il suffit que deux sources ne partagent pas les mêmes avis pour engendrer un conflit.
- ceux qui utilisent les éléments focaux : (H, Θ) [Zouhal et Denoeux, 1997]. Ce modèle a été défini pour la première fois dans le cadre de l'algorithme des k plus proches voisins utilisé pour la reconnaissance de formes. Lors de l'utilisation de ce jeu de masses, la source d'information se focalise uniquement sur l'hypothèse H. Cette modélisation permet de traduire la réalisation d'une hypothèse indépendamment des autres. A travers ce modèle, si deux objets sont proches, alors on peut conclure qu'ils sont homologues, par contre s'ils sont très éloignés, on ne peut rien conclure. Etant donné que m(Θ) est l'élément neutre de la règle de combinaison de Dempster, une observation très éloignée n'aura qu'une influence négligeable sur le jeu de masses combinées. Ce type de modélisation ne permet pas de mettre en évidence une hypothèse qui n'a pas été définie et modélisée dans le cadre de discernement.

B.6 Conclusion

Dans ce chapitre nous avons abordé les problèmes liés à l'imperfection dans les connaissances et nous avons décrit les outils disponibles pour traiter des connaissances imparfaites.

Nous avons vu que l'imperfection dans les connaissances fait appel à trois concepts : l'imprécision, l'incertitude et l'incomplétude. Afin de gérer et de modéliser les imperfections, de nombreuses théories mathématiques existent dans la littérature. Les plus employées sont celles faisant partie de la famille des théories de l'incertain telles que : la théorie des probabilités, la théorie des sous-ensembles flous, la théorie des fonctions de croyance, la théorie des possibilités, la théorie des ensembles grossiers. Plusieurs travaux consacrés à comparer les théories mathématiques ont été réalisés dans la littérature.

La théorie des probabilités permet de représenter explicitement des connaissances statistiques, tandis que la théorie des sous-ensembles flous, la théorie des fonctions de croyance et la théorie des possibilités permettent une représentation plus fine des connaissances qu'un individu a sur un phénomène observé. Au-delà des différences subtiles, parfois des différences d'interprétation, aucune conclusion n'a été donnée concernant la supériorité de l'une sur l'autre. Ainsi, nous pouvons affirmer que le choix de l'utilisation dépend de l'application visée.

Nous avons vu que lors de la fusion des sources d'information dans le cadre de la théorie des fonctions de croyance un conflit peut apparaître. Plusieurs interprétations ont été définies au sujet de la nature du conflit.

Ainsi, si le cadre de discernement est considéré exhaustif, c'est-à-dire sous l'hypothèse du *monde fermé*, alors le conflit provient de la non-fiabilité des sources d'information. Si les sources d'information qui ne sont pas fiables sont connues, la solution proposée est de les affaiblir, ce qui revient à accorder plus de poids à l'ignorance. Au contraire, si les sources non-fiables ne sont pas identifiées avec précision, alors de nombreux auteurs proposent des opérateurs de redistribution du conflit. Malheureusement, la plupart de ces opérateurs ne sont pas associatifs, ce qui implique qu'une stratégie de séquencement des fusions doit être définie au préalable.

Par contre, si le cadre de discernement n'est pas exhaustif, c'est-à-dire l'hypothèse du *monde ouvert*, alors le conflit est interprété comme le fait qu'il existe des hypothèses omises dans le cadre de discernement.

Pour notre application, l'appariement de données géographiques, nous avons considéré que la théorie des fonctions de croyance répond le mieux à nos besoins et sera à la base de notre approche.

Dans le chapitre suivant nous présentons notre processus d'appariement de données géographiques basé sur la théorie des fonctions de croyance.

CHAPITRE C
Processus d'appariement de données géographiques basé sur la théorie des fonctions de croyance

C Processus d'appariement de données géographiques basé sur la théorie des fonctions de croyance

Nous présentons dans ce chapitre notre approche d'appariement de données géographiques. Les expérimentations que nous avons mises en œuvre afin d'étudier la faisabilité de notre approche seront présentées dans le chapitre D, et nous exposons les points forts et les points faibles de celle-ci dans le chapitre E.

Introduction

Dans notre approche, l'initialisation des masses de croyance est réalisée en utilisant l'un des deux modèles proposés par [Appriou, 1991], que nous avons présentés dans le chapitre B. D'une manière générale, chaque source se spécialise dans une hypothèse du cadre de discernement, c'est-à-dire que la source d'information analyse l'hypothèse donnée et se prononce en sa faveur, en sa défaveur ou elle ne se prononce pas parce qu'elle n'a pas assez de connaissances sur cette hypothèse.

Nous avons utilisé un jeu de masses composé de trois éléments focaux (H, ¬H, Θ)**.** La raison de notre choix est que ce jeu de masses permet de modéliser la connaissance complète, m(H), la connaissance incomplète, m(¬H) et l'ignorance m(Θ). Par conséquent, nous modélisons les masses de croyance en une hypothèse simple {H}, en un ensemble d'hypothèses {¬H} et en toutes les hypothèses du cadre de discernement (Θ).

Afin d'initialiser les jeux de masses, nous avons utilisé des mesures de distance issues des comparaisons des différentes propriétés des objets géographiques. Ce choix a été privilégié par rapport à l'utilisation d'un ensemble d'apprentissage, parce que nous considérons que dans le cas de l'appariement de données géographiques, il est difficile de trouver une zone d'apprentissage représentative contenant tous les cas d'appariement possibles et parce que cette méthode est plus coûteuse. De plus nous disposons des connaissances sur les critères et de leur validité. Le recours à des méthodes d'apprentissage pourrait bien sûr être envisagé.

C.1 Le processus d'appariement de données géographiques

Dans cette partie, nous présentons notre processus d'appariement, qui est composé de cinq étapes principales, illustrées en Figure 51.

Nous considérons deux jeux de données JD_1 et JD_2. Chaque objet *obj1* de *JD1* est traité tour à tour, et l'objectif est de trouver son homologue dans JD_2. La première étape du processus d'appariement consiste à sélectionner les candidats, c'est-à-dire à chercher pour chaque objet obj_1 appartenant à JD_1 des objets homologues potentiels dans le jeu de données JD_2, appelés candidats à l'appariement et notés $C_{i,(i=1..N)}$. Ensuite, chaque candidat à l'appariement est analysé afin de déterminer le lien d'appariement.

L'initialisation des masses de croyance (deuxième étape du processus) consiste pour chaque critère d'appariement à se prononcer sur chaque candidat en attribuant une croyance dans les hypothèses définies à l'égard de chaque candidat. Les critères d'appariement sont basés sur les distances entre obj_1 et chaque candidat sélectionné et sur les connaissances issues des spécifications, des données elles-mêmes et des experts. Pour que les connaissances soient exploitables dans le processus d'appariement, elles doivent être représentées d'une

manière explicite. La représentation explicite des connaissances peut être vue également comme l'initialisation de masses de croyance, elle fait l'objet de la partie C.2.

Ensuite, après une fusion des critères d'appariement pour chaque candidat (troisième étape du processus), nous effectuons une fusion des candidats afin d'avoir une vue d'ensemble sur les croyances attribuées à tous les candidats (quatrième étape du processus). Enfin, nous avons l'étape de décision, où le meilleur candidat est choisi.

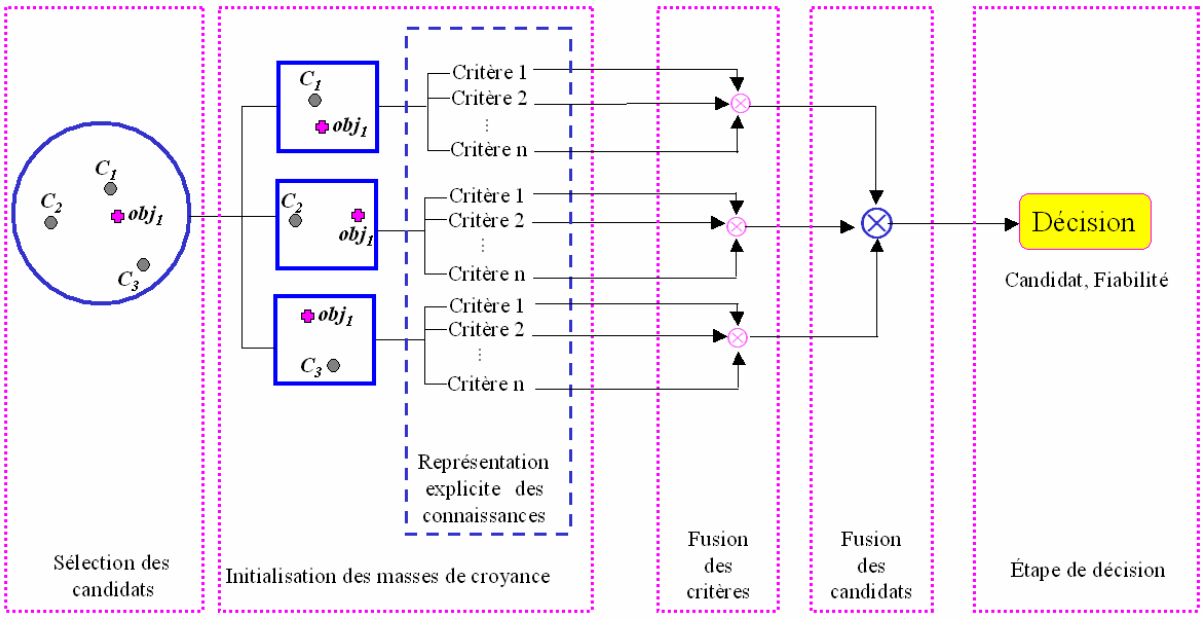


Figure 51. Processus d'appariement de données géographiques détaillé

C.1.1 Sélection des candidats

La première étape de notre processus est la sélection des candidats à l'appariement (voir Figure 52).

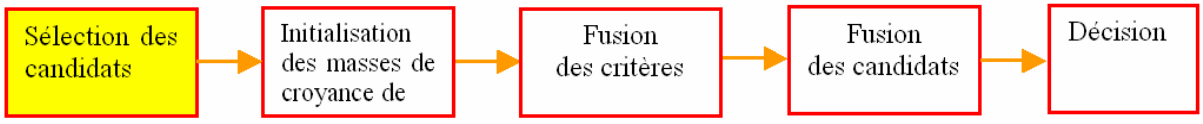


Figure 52. Etape de sélection des candidats à l'appariement

L'étape de sélection des candidats consiste à définir le cadre de discernement associé à obj_1. Pour cela nous sélectionnons les objets les plus proches (obj_{2i}), i=1..N dans le jeu de données JD2 selon un critère de distance géométrique (voir Figure 53). A partir de ce moment nous notons chaque objet candidat (obj_{2i}), i=1..N par C_i, i=1..N.

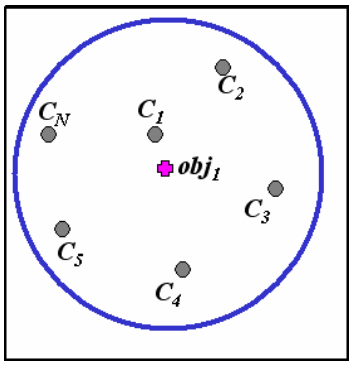


Figure 53. Candidats à l'appariement pour l'objet obj_1 sélectionnés selon un critère de distance géométrique

Le choix du seuil de distance dépend en général de la connaissance de l'erreur moyenne sur la position des éléments présents dans les jeux de données et de la nature des objets. Ainsi, un seuil de sélection élevé permet de prendre en compte l'imprécision sur la localisation des données géographiques. Cependant, un seuil très élevé augmente la complexité algorithmique, parce que plus le seuil est élevé, plus il y a de candidats et plus le nombre de combinaisons possibles dans 2^{Θ} augmente. C'est pour cela que nous considérons que le seuil de sélection doit prendre en compte la nature de l'objet obj_1. Par exemple, le seuil de sélection pour un objet de nature *vallée* sera plus élevé que le seuil de sélection pour un objet de nature *pic,* mieux localisé en général.

Cette sélection de candidats permet la définition du cadre de discernement. Ainsi, afin de construire le cadre de discernement, nous considérons qu'apparier l'objet obj_1 à chaque candidat sélectionné constitue une hypothèse, donc une solution possible.

Par conséquent, le cadre de discernement est un ensemble d'hypothèses $appC_{i,(i=1..N)}$ exprimant chacune « le candidat C_i est l'homologue de l'objet obj_1 ». De plus, nous avons constaté que des objets peuvent ne pas avoir d'homologues dans l'autre base et donc ne pas être appariés. Cela nous amène à définir une nouvelle hypothèse NA signifiant : « l'objet obj_1 n'est pas apparié ». Ainsi, l'ajout de l'hypothèse NA rend le cadre de discernement exhaustif, c'est-à-dire que la solution se trouve parmi les hypothèses définies. Grâce à l'hypothèse NA, le conflit après la fusion des critères est moins important parce que la solution se trouve dans le cadre de discernement.

Le cadre de discernement pour un objet obj_1 ayant N candidats à l'appariement est défini par :

$$\Theta_{obj_1} = \{appC_1, appC_2, ..., appC_i ... appC_N, NA\} \quad (37)$$

A partir du cadre de discernement, nous définissons le référentiel de définition 2^{Θ}. Le référentiel de définition contient donc toutes les combinaisons possibles des hypothèses définies dans le cadre de discernement (conformément au chapitre B). Ainsi, pour un cadre de discernement composé de N hypothèses issues de la sélection de N candidats, auquel nous ajoutons l'hypothèse NA, nous avons 2^N combinaisons.

$$2^{\Theta_{obj1}} = \{\ appC_1, ..., appC_N, \{\ appC_1, appC_2\ \}, \{\ appC_1, appC_3\ \}, ... \{\ appC_1, ... appC_N, NA\}, NA\}$$

(38)

Exemple

Nous illustrons un exemple typique d'appariement de données. Cet exemple est repris à la fin de chaque étape de notre processus.

Considérons que nous cherchons à apparier l'objet obj_1. L'étape de sélection nous permet donc de définir les candidats à l'appariement. Supposons qu'il y a trois candidats à l'appariement C_1, C_2, C_3 (voir Figure 54), qu'un seul candidat peut être choisi, et que le bon candidat est le candidat C_1. Le cadre de discernement est donc défini comme suit :

$\Theta = \{appC_1, appC_2, appC_3, \text{NA}\}$.

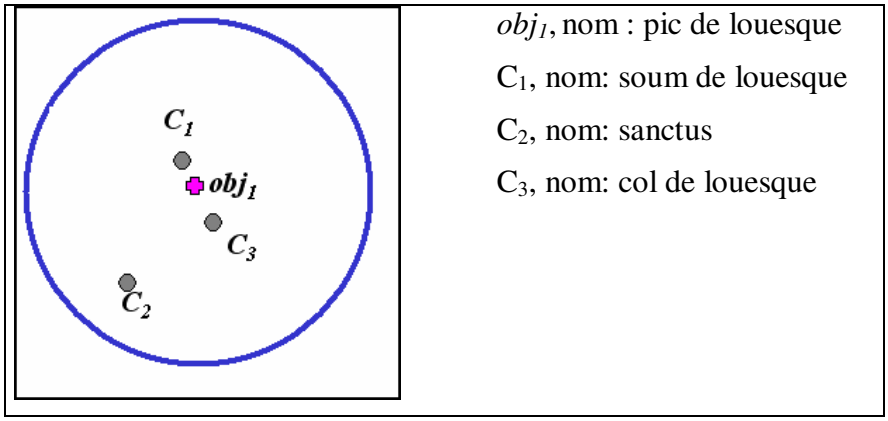


Figure 54. Exemple typique d'appariement de données géographiques

C.1.2 Initialisation des masses de croyance

La deuxième étape consiste à initialiser les masses de croyance pour chaque candidat indépendamment des autres et pour chaque source d'information. L'initialisation de masses de croyance est une tâche primordiale, dont dépendent les résultats.

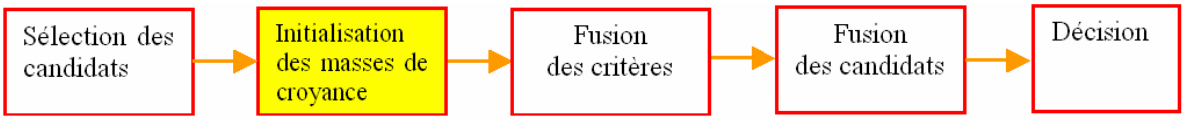


Figure 55. Deuxième étape du processus : initialisation des masses de croyance

Dans notre approche, l'initialisation des masses de croyance repose sur des fonctions floues définies à partir des connaissances. Les connaissances que nous utilisons sont : les seuils issus des spécifications, la taille usuelle des entités, les mesures de distance déterminées à partir des comparaisons de différents attributs des données géographiques utilisées et les règles d'appariement implicites définies par les experts. La mise en oeuvre est réalisée en utilisant l'un des modèles proposés par [Appriou, 1991] que nous avons présenté au chapitre B. Ce modèle est connu dans la littérature sous le nom de sources spécialisées ou focalisées. Ainsi, à partir d'un sous-ensemble du référentiel de définition composé de trois éléments focaux, chaque source se spécialise et se prononce sur un seul élément focal.

Dans notre cas, une source d'information est un critère guidant l'appariement, comme la proximité spatiale, la ressemblance des toponymes, la sémantique. Dans la suite de ce

mémoire de thèse nous utilisons le terme de critère à la place de source, terme plus utilisé en Géomatique et plus particulièrement pour qualifier les processus d'appariement.

L'initialisation explicite des masses de croyance pour différents critères d'appariement est détaillée dans la partie C.2.

Conformément au modèle d'Appriou, considérons un sous-ensemble S_i de 2^Θ, défini comme ci-après:

$$S_i = \{appC_i, \neg appC_i, \Theta\} \quad (39)$$

où :

- $appC_i$ représente l'hypothèse que l'objet obj_1 en cours d'analyse est apparié avec le candidat C_i, signifiant la connaissance complète,
- $\neg appC_i = \{appC_1, appC_2 \ldots appC_{i-1}, appC_{i+1} \ldots appC_N, \text{NA}\}$ représente l'hypothèse que l'objet obj_1 en cours d'analyse est apparié avec un autre candidat que C_i ou pas apparié du tout, signifiant la connaissance incomplète),
- $\Theta = \{ appC_1, appC_2 \ldots appC_i \ldots appC_N, \text{NA}\}$ représente l'hypothèse que nous ne pouvons pas nous prononcer sur ce candidat, signifiant l'ignorance,

Signalons que la somme des masses de croyance attribuées aux trois hypothèses doit être égale à 1, $m(appC_i) + m(\neg appC_i) + m(\Theta) = 1$.

La raison pour laquelle nous avons choisi le modèle d'Appriou est que ce jeu de masses permet de modéliser la connaissance complète, $m(appC_i)$, la connaissance incomplète, $m(\neg appC_i)$ et l'ignorance $m(\Theta)$. De plus, l'utilisation d'un sous-ensemble du référentiel de définition 2^Θ, conformément à l'équation (39), réduit le nombre de combinaisons possibles et donc la complexité algorithmique est réduite. L'utilisation d'un sous-ensemble seulement du référentiel de définition a aussi l'avantage que nous devons initialiser seulement les masses de croyance attribuées à ces trois hypothèses, et non pas celles attribuées à toutes les hypothèses du référentiel de définition.

Signalons que l'hypothèse NA n'est pas initialisée, elle apparaît après la fusion des candidats.

Exemple (voir *Figure 54*)

Supposons qu'afin d'apparier l'objet obj_1, nous utilisons deux critères : le critère d'écart de position, basé sur la proximité entre l'objet obj_1 et chacun des candidats, et le critère toponymique, basé sur la comparaison des toponymes. Les deux critères se spécialisent sur chacun des candidats et fournissent les jeux de masses suivants :

$$\begin{cases} m_1(appC_1)=0,4 \\ m_1(\neg appC_1)=0 \\ m_1(\Theta)=0,6 \\ \\ m_2(appC_1)=0,3 \\ m_2(\neg appC_1)=0 \\ m_2(\Theta)=0,7 \end{cases} \quad \begin{cases} m_1(appC_2)=0,1 \\ m_1(\neg appC_2)=0,9 \\ m_1(\Theta)=0 \\ \\ m_2(appC_2)=0 \\ m_2(\neg appC_2)=1,0 \\ m_2(\Theta)=0 \end{cases} \quad \begin{cases} m_1(appC_3)=0,35 \\ m_1(\neg appC_3)=0.0 \\ m_1(\Theta)=0,65 \\ \\ m_2(appC_3)=0,3 \\ m_2(\neg appC_3)=0 \\ m_2(\Theta)=0,7 \end{cases}$$

Dans le jeu de masses défini ci-dessus, $m_1(X)$ et $m_2(X)$ représentent respectivement les masses de croyance attribuées par le critère 1 (le critère d'écart de position) et par le critère 2 (le critère toponymique), à une hypothèse X appartenant au référentiel de définition. Nous remarquons que les deux critères attribuent des masses de croyance relativement proches aux candidats C_1 et C_3. Ceci est dû au fait que les deux candidats se trouvent à peu près à la même distance de l'objet obj_1 et que les toponymes se ressemblent.

C.1.3 Fusion des critères

Une fois que les masses de croyance sont initialisées, pour chaque candidat C_i, nous fusionnons les jeux de masses issus de chaque critère d'appariement, avec l'opérateur conjonctif, c'est-à-dire que nous appliquons l'opérateur de Dempster sans normaliser les masses de croyance (voir Figure 56).

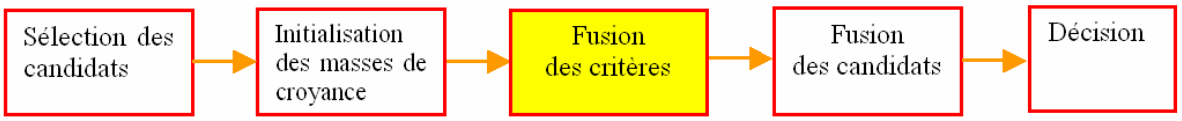


Figure 56. Troisième étape du processus : fusion des critères

Au final, nous obtenons pour chaque candidat C_i un jeu de masses composé des hypothèses C_i, $\neg C_i$, Θ, synthétisant l'avis de tous les critères qui se sont prononcés individuellement.

Par exemple, soit deux critères d'appariement Critère 1 et Critère 2 qui se prononcent chacun sur le candidat C_i, comme illustré en Figure 57. On note $\{m_1(appC_i), m_1(\neg appC_i), m_1(\Theta)\}$ le jeu de masses défini par le critère Critère 1 et $\{m_2(appC_i), m_2(\neg appC_i), m_2(\Theta)\}$ le jeu de masses défini par le critère Critère 2. Afin de prendre en compte les avis issus des deux critères, les deux jeux de masses sont fusionnés en utilisant l'opérateur conjonctif. Le jeu de masses obtenu après la fusion des critères est $\{m_{12}(appC_i), m_{12}(\neg appC_i), m_{12}(\Theta), m_{12}(\phi)\}$. La masse de croyance $m_{12}(\phi)$ représente le conflit engendré par le désaccord entre les connaissances issues de chaque critère. Par exemple, selon les connaissances représentées dans le critère Critère 1, nous croyons que c'est le candidat C_i le vrai homologue alors que selon les connaissances modélisées par le critère Critère 2 nous croyons le contraire, c'est-à-dire que ce n'est pas le candidat C_i le vrai homologue.

Nous rappelons ci-dessous la définition de l'opérateur conjonctif :

$$\forall P \in 2^{\Theta},\ m_{12}(P) = \frac{1}{1 - m_{12}(\phi)} \sum_{P' \cap P'' = P} m_1(P') m_2(P'') \qquad (40)$$

Dans notre exemple, la fusion des deux critères Critère 1 et Critère 2 consiste donc, pour chacune des hypothèses $appC_i$, $\neg appC_i$ et Θ, à additionner les produits des masses de croyance issues des deux critères pour lesquelles l'intersection des hypothèses est égale respectivement à $appC_i$, $\neg appC_i$, Θ.

Le jeu de masses après la fusion est défini sur la Figure 57 :

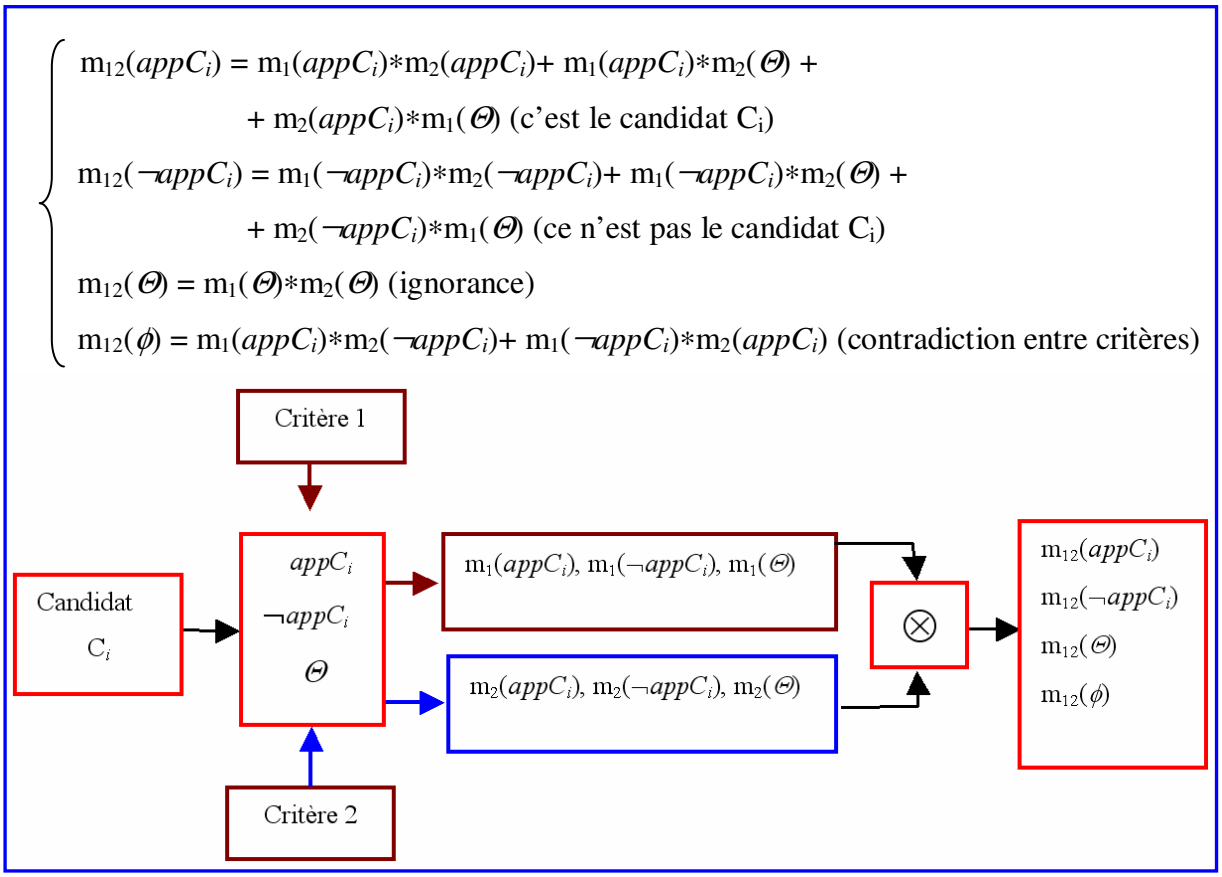


Figure 57. Fusion des critères pour le candidat C_i

Exemple

Nous continuons notre exemple avec l'étape de fusion des critères pour chaque candidat.

La Figure 58 illustre les jeux de masses issus de la fusion des critères pour chaque candidat. En analysant la Figure 58, nous pouvons exprimer avec certitude que le candidat C_2 n'est pas l'objet homologue de l'objet obj_1. Par contre, nous constatons que les candidats C_1 et C_3 ont des masses de croyance très proches, donc aucune décision ne peut être prise avec certitude à cause de l'ambiguïté existante.

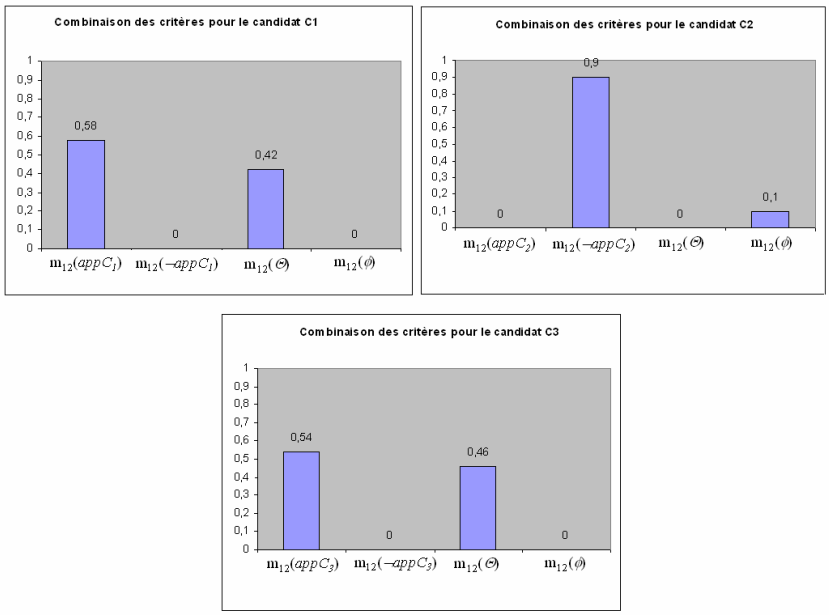


Figure 58. Jeux de masses après la combinaison des critères pour chaque candidat

C.1.4 Fusion des candidats

La quatrième étape de notre processus d'appariement consiste à fusionner les candidats entre eux (voir Figure 59).

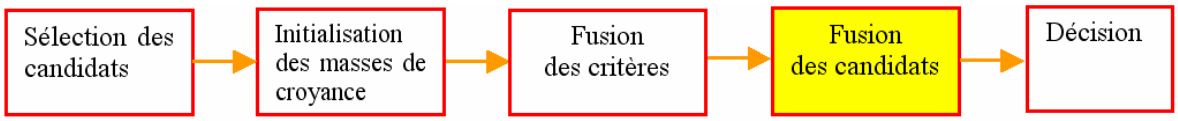


Figure 59. Quatrième étape du processus : fusion des candidats

A la fin de la troisième étape de notre approche, nous avons un jeu de masses fusionnées pour chaque candidat à l'appariement. A ce stade, une décision pourrait être directement prise, c'est-à-dire que nous pourrions choisir le candidat pour lequel la masse de croyance attribuée à l'hypothèse $appC_i$ a la valeur maximale. Cependant, d'une part il s'avère qu'il y a des cas d'ambiguïté pour lesquels il n'existe pas un seul candidat qui se distingue particulièrement, mais plusieurs. D'autre part, cette unique fusion ne permet pas de mettre en évidence l'hypothèse NA, c'est-à-dire que l'hypothèse « l'objet obj_1 n'est pas apparié » n'est jamais choisie, elle fait seulement partie de l'hypothèse $\neg appC_i$. Or, nous savons que l'absence d'homologue n'est pas un cas exceptionnel pour l'appariement de données géographiques. Ainsi, afin de lever l'ambiguïté, et de faire apparaître l'hypothèse NA, nous avons réalisé une quatrième étape, de fusion des candidats illustrée sur la Figure 60. Une telle stratégie de fusion a été proposée par [Royère, 2002] dans le cadre d'une application de localisation d'un véhicule sur une carte.

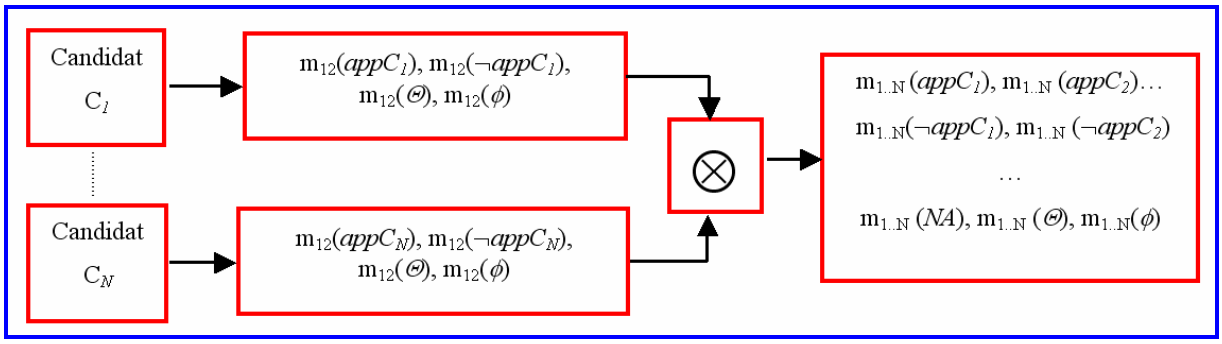


Figure 60. Combinaison des masses de croyance pour chacun des candidats

Pour chaque candidat à l'appariement nous avons un jeu de masses issu de la fusion des critères. En Figure 60 nous avons considéré deux critères et N candidats à l'appariement. Ainsi, après la fusion des critères, nous avons pour chaque candidat un jeu de masses noté m_{12}. Par exemple, pour le candidat C_1 nous avons le jeu de masses $m_{12}(appC_1)$, $m_{12}(\neg appC_1)$, $m_{12}(\Theta)$ et $m_{12}(\phi)$.

La fusion des candidats est réalisée en fusionnant avec l'opérateur conjonctif, le jeu de masses pour le candidat C_1 avec le jeu de masses pour le candidat C_2, puis le résultat avec le jeu de masses pour le candidat C_3, et ainsi suite jusqu'au candidat N. Rappelons que, l'opérateur conjonctif étant associatif, l'ordre de fusion des candidats n'est pas important. La masse de croyance après la fusion des N candidats est notée $m_{1..N}$.

La masse de croyance attribuée à l'hypothèse NA résulte de la fusion de toutes les masses de croyance attribuées aux hypothèses $\neg appC_i$, pour i = 1..N. Elle est définie de la manière suivante :

$$m_{1..N}(NA) = m_{1..N}(\neg appC_1) * m_{1..N}(\neg appC_2) * ... m_{1..N}(\neg appC_i) * .... * m_{1..N}(\neg appC_N) \qquad (41)$$

L'étape de fusion des candidats peut augmenter la masse de croyance associée au conflit quand différents critères soutiennent plusieurs candidats en même temps. Ceci est dû justement à l'approche que nous avons adoptée pour initialiser les masses de croyance, en les initialisant pour chaque candidat indépendamment des autres. Cependant, nous avons supposé qu'une modélisation prudente, c'est-à-dire que nous attribuons une masse de croyance importante à l'ignorance dans le cas où nous ne sommes pas sûrs de la vérité d'une hypothèse, pourrait prévenir l'apparition d'un fort conflit entre les critères.

Exemple

Si nous effectuons une fusion des candidats, toujours en utilisant l'opérateur conjonctif, nous obtenons les jeux de masses illustrés sur la Figure 61. Nous remarquons qu'après la fusion des candidats, les valeurs des masses de croyance attribuées aux hypothèses $appC_1$ et $appC_3$ sont relativement proches et la valeur de la masse de croyance associée au conflit 0,39, est assez importante, par rapport aux autres valeurs obtenues. Ce conflit est dû au fait que les critères d'appariement soutiennent les candidats C_1 et C_3 d'une manière sensiblement identique.

Nous donnons à titre d'illustration la formule mathématique après la fusion des critères, pour la masse de croyance attribuée au candidat C_1 :

$$m_{1..3}(appC_1) = m_{12}(appC_1) * [m_{12}(\neg appC_2) + m_{12}(\Theta)] * [m_{12}(\neg appC_3) + m_{12}(\Theta)] \quad (42)$$

où l'élément appartenant au premier crochet est issu de la fusion avec les jeux de masses du candidat C_2, tandis que le deuxième crochet est issu de la fusion avec les jeux de masses du candidat C_3.

En analysant le jeu de masses illustré sur la Figure 61, nous pouvons dire avec certitude que l'objet obj_1 ne peut pas être apparié avec le candidat C_2, puisque l'hypothèse $appC_2$ n'apparaît pas dans le jeu de masses final. La masse de croyance attribuée à l'hypothèse $\neg appC_2$ signifie que l'objet obj_1 peut soit être apparié avec le candidat C_1, soit être apparié avec C_2, soit ne pas être apparié.

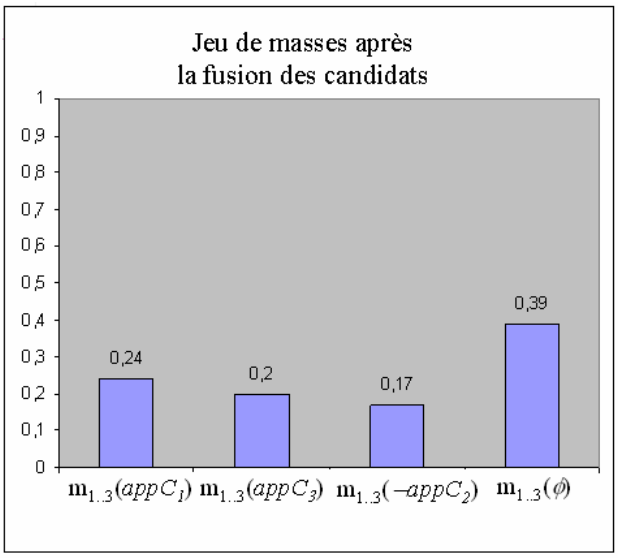


Figure 61. Fusion des candidats

C.1.5 Décision

La cinquième étape de notre processus est la prise de la décision (voir Figure 62).

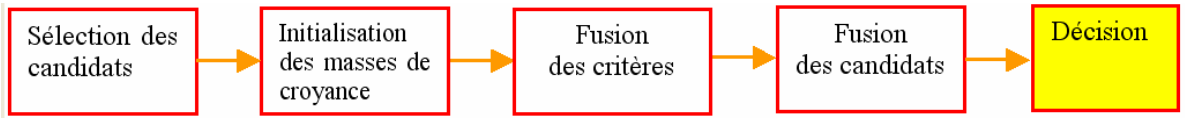


Figure 62. Cinquième étape du processus : décision

La décision est prise après l'étape de fusion des candidats et après avoir normalisé les masses résultantes en utilisant l'opérateur de Dempster. Rappelons qu'à ce niveau le conflit vaut 0, étant utilisé pour la normalisation des masses de croyance.

Notre objectif ici est d'apparier un objet à un seul et unique candidat. Ainsi, notre approche permettrait d'apparier d'une part les objets qui ont un seul homologue dans l'autre jeu de données, mais aussi les objets qui ont plusieurs homologues. Cette dernière tâche peut être réalisée en regroupant les appariements obtenus. Nous y reviendrons dans le chapitre E.

Compte tenu de notre objectif, le choix d'utiliser la probabilité pignistique lors de l'étape de décision nous semble approprié, puisque la probabilité pignistique privilégie les hypothèses simples.

Afin de prendre une décision, nous avons utilisé le maximum de probabilité pignistique, c'est-à-dire l'hypothèse simple avec la masse de croyance la plus élevée. Par exemple la Figure 63, montre qu'à partir du jeu de masses calculé après la fusion des candidats, nous calculons les probabilités pignistiques pour toutes les hypothèses simples, ensuite nous choisissons celle pour laquelle la probabilité pignistique a la valeur maximale. Dans notre exemple, il s'agit de l'hypothèse $appC_1$ signifiant que le candidat homologue de l'objet *obj1* est le candidat C_1.

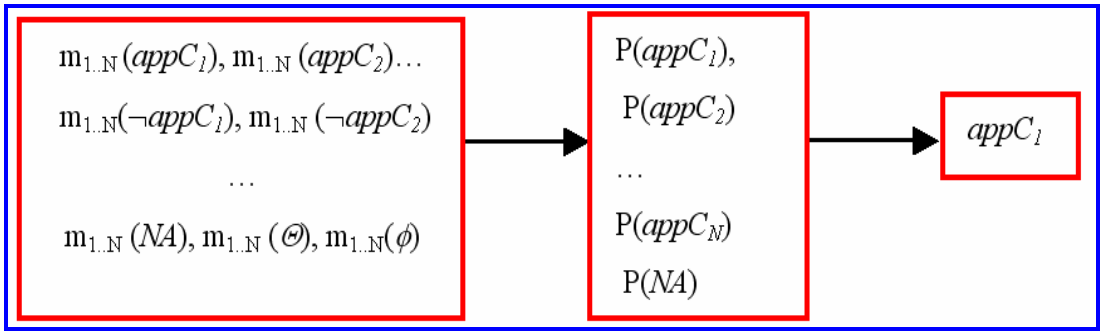


Figure 63. Décision selon le maximum de probabilité pignistique

Nous rappelons ci-dessous la formule de calcul de la probabilité pignistique :

$$\forall A \in 2^{\Theta},\ P(A) = \sum_{A \subseteq B} m(B) \frac{|A \cap B|}{|B|} \qquad (43)$$

où | X | signifie le nombre d'hypothèses contenues dans la proposition X.

Nous avons vu dans le chapitre B que plusieurs outils existent pour prendre une décision. L'utilisation de la fonction de crédibilité ou de la fonction de plausibilité fournit comme décision soit une hypothèse simple, soit une proposition, c'est-à-dire une union d'hypothèses, tandis que l'utilisation de la probabilité pignistique privilégie les hypothèses simples.

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. La décision est portée sur une seule hypothèse simple $appC_i$, par conséquent sur un seul candidat C_i, et donc il n'y a pas d'incertitude à ce niveau. Toutefois, précisons que dans le calcul de la probabilité pignistique associée à une hypothèse $appC_i$, notée $P(appC_i)$, toutes les propositions contenant l'hypothèse $appC_i$ sont prises en compte, dans le but de choisir la « meilleure » hypothèse. Par exemple, dans le calcul final de la probabilité pignistique attribuée à l'hypothèse $appC_1$ s'additionnent toutes les fractions de masse de croyance de chaque hypothèse contenant $appC_i$, c'est-à-dire $\neg appC_2$, $\neg appC_3$, …, $\neg appC_N$ ou Θ.

L'hypothèse *NA* peut être également choisie en utilisant la probabilité pignistique.

Notons que l'hypothèse choisie est celle pour laquelle la probabilité pignistique a la valeur maximale. Une incertitude au niveau de la décision peut néanmoins toujours exister lorsque cette dernière est prise en choisissant l'hypothèse qui correspond à la valeur maximale. Ainsi, il est important d'étudier la fiabilité des résultats d'appariement obtenus.

Notre processus d'appariement permet de vérifier automatiquement la fiabilité des résultats.

Dans le but d'analyser la confiance que l'on a dans l'hypothèse choisie, [Appriou, 1999] par exemple propose de vérifier si l'hypothèse choisie est celle qui correspond à une consistance K_i minimale définie de la manière suivante :

K_i = 1-P1 (hypothèse choisie) (44)

Où Pl représente la fonction de plausibilité attribuée à l'hypothèse choisie.

Etant donné que la fonction de plausibilité est une mesure optimiste, plus le facteur K_i est petit, plus nous sommes sûrs de l'hypothèse choisie.

Afin d'étudier la fiabilité des résultats obtenus, nous proposons d'étudier l'écart entre le premier et le deuxième maximum de probabilité pignistique correspondant aux hypothèses simples. Si l'écart est important, par exemple s'il est supérieur à un seuil donné, nous considérons que le résultat est fiable et le lien d'appariement est qualifié de sûr. Au contraire, si l'écart entre le premier et le deuxième maximum est faible, c'est-à-dire si le deuxième maximum est proche du premier, nous considérons que le résultat est peu fiable. Par conséquent, le lien d'appariement est qualifié d'incertain. Les cas incertains sont mis en valeur, puis vérifiés d'une manière interactive.

Exemple

Après la normalisation des masses selon l'opérateur de Dempster, c'est-à-dire que toutes les masses de croyance sont normalisées par la masse de croyance attribuée au conflit, nous remarquons sur la Figure 64 que l'hypothèse associée au candidat C_1 se détache un peu plus de celle attribuée au candidat C_3, ayant une masse de croyance égale à 0,4.

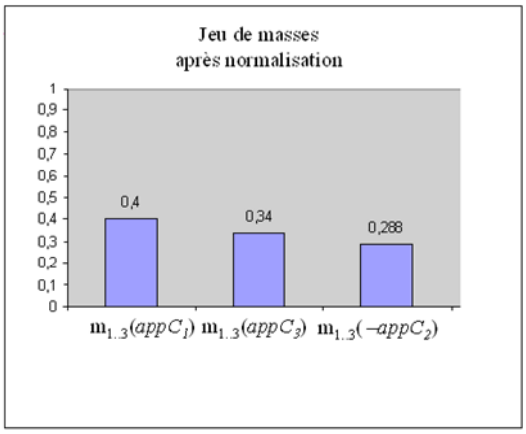


Figure 64. Jeu de masses de croyance après normalisation

Après le calcul de la probabilité pignistique, l'hypothèse $appC_1$ est choisie, puisque la probabilité pignistique attribuée a la valeur maximale, P($appC_1$)=0,5. En conséquence, l'objet obj_1 est apparié avec le candidat C_1.

A titre d'illustration, nous donnons ci-dessous la formule de calcul de la probabilité pignistique pour l'hypothèse $appC_1$:

$$P(appC_1) = m_{1..3}(appC_1) + \frac{|appC1 \cap \neg appC_2|}{|\neg appC_2|} m_{1..3}(\neg appC_2) = 0{,}4 + \frac{1}{3}(0{,}288) = 0{,}5 \quad (45)$$

C.1.6 Discussion sur la stabilité du processus

Nous rappelons que la fusion des critères et celle des candidats a été réalisée en utilisant l'opérateur conjonctif de Dempster. Cet opérateur a été contesté pour le fait que le conflit n'est pas redistribué mais utilisé pour normaliser le jeu de masses final. La faiblesse de cet opérateur est due surtout à la sensibilité du facteur de normalisation $[1\text{-}m(\phi)]^{-1}$ lorsque la valeur du conflit est élevée [Zadeh, 1986]. A partir de ce constat, [Colot, 2000] a étudié la sensibilité de l'opérateur de Dempster au conflit. D'après la courbe de variation du coefficient de normalisation en fonction du conflit présentée dans [Colot, 2000], nous remarquons que le facteur de normalisation devient très sensible lorsque la valeur du conflit est supérieure à 0,95, c'est-à-dire qu'une faible variation du conflit entraîne une forte variation du facteur de normalisation (voir Figure 65).

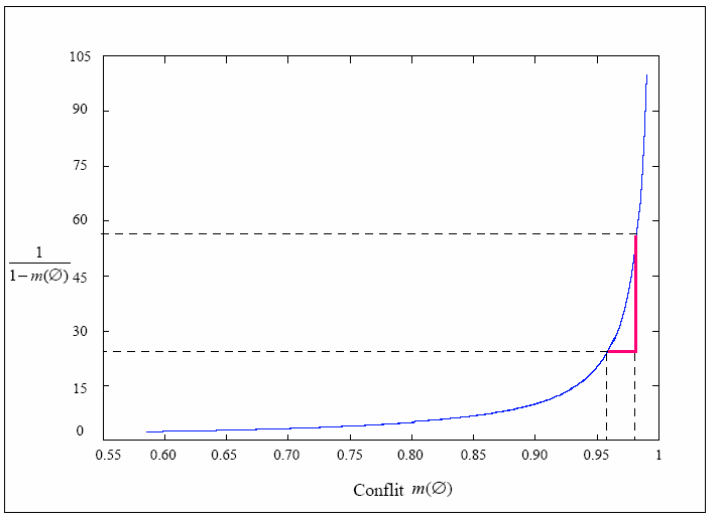


Figure 65. Variation du coefficient de normalisation en fonction du conflit, [Colot, 2000]

Nous avons néanmoins choisi d'utiliser dans nos expérimentations l'opérateur de Dempster premièrement pour ses propriétés (commutativité, associativité), pour la simplicité de la méthode, et du fait que nous n'avons pas de connaissances sur le meilleur ordre de prise en compte des critères. De nombreux travaux ont proposé d'autres opérateurs de fusion qui gèrent le conflit d'une manière efficace. Mais ces opérateurs sont en général non-associatifs. Notons que dans notre cas la propriété d'associativité est importante, parce que nous fusionnons plus de deux critères et plus de deux candidats à l'appariement. Deuxièmement, nous avons constaté dans nos expérimentations que le conflit est faible. Cela a été possible grâce à notre modélisation prudente des connaissances (voir partie C.2). Cependant, afin de rester prudent, aucune fusion n'est faite, en conséquence aucune décision n'est prise, si le conflit est supérieur à 0,9. Dans ce cas, nous émettons un avertissement, et un appariement interactif est réalisé.

Concernant la redistribution du conflit, nous avons testé plusieurs propositions, telles que la proposition de [Colot, 2000] ou de [Royère, 2002]. A titre d'exemple nous montrons les résultats que nous avons obtenus pour un cas d'appariement de données.

Exemple. Considérons un cas d'appariement où seul un candidat, parmi les quatre sélectionnés, doit être apparié à un objet obj_1. Supposons qu'après les étapes de fusion des critères pour chaque candidat et de fusion des candidats nous obtenons le jeu de masses illustré sur la Figure 66.

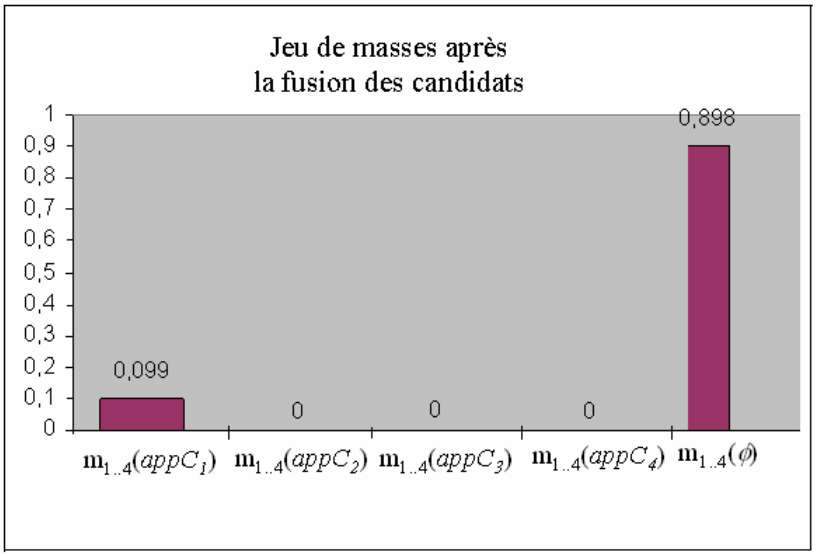


Figure 66. Exemple de jeu de masses après la fusion des quatre candidats

Afin de redistribuer le conflit, nous avons testé l'approche de Dempster, [Dempster, 1967 ; Shafer, 1976] qui utilise le conflit pour normaliser le jeu de masses, les deux approches de [Colot, 2000] et les deux approches de [Royère, 2002] (Figure 67). Nous pouvons constater que l'approche de Dempster [Dempster, 1967 ; Shafer, 1976] ne redistribue le conflit que sur l'hypothèse non nulle du jeu de masses, $appC_1$, alors que les autres approches redistribuent le conflit à l'union des hypothèses ou bien aux hypothèses simples qui ont engendré le conflit.

Nous remarquons que dans ce cas de conflit très fort les opérateurs de Royère [Royère, 2002] nous semblent plus pertinents, parce que la masse de croyance est attribuée à l'hypothèse app{C_1, C_3, C_4}, c'est-à-dire que l'objet est apparié soit avec le candidat C_1 soit avec C_3 soit avec C_4. Cependant, compte tenu de notre application, dans un tel cas de fort conflit, nous considérons que la solution la plus raisonnable est de ne pas conclure, c'est-à-dire de ne pas redistribuer le conflit.

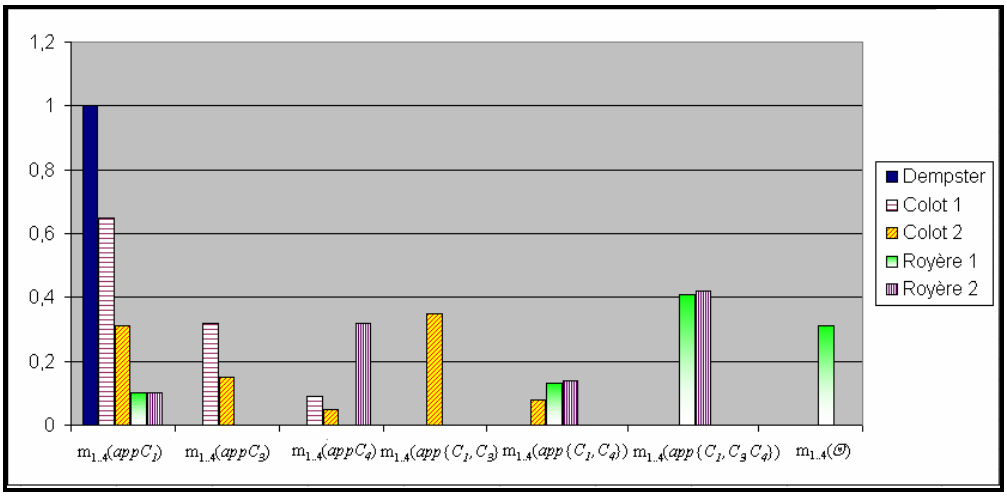


Figure 67. Redistribution du conflit selon plusieurs opérateurs

Nous avons mentionné dans le chapitre B qu'il y a plusieurs possibilités pour prendre une décision telles que : le maximum de crédibilité, le maximum de plausibilité et le maximum de probabilité pignistique. Selon la redistribution du conflit à travers les approches énoncées ci-dessus, nous obtenons les résultats suivants :

	Max (Crédibilité)	Max (Plausibilité)	Max (Probabilité pignistique)
Dempster	$appC_1$	$appC_1$	$appC_1$
Colot 1	$appC_1$	$appC_1$	$appC_1$
Colot 2	$appC_1$	$app(\{C_1, C_3\})$	$appC_1$
Royère 1	$appC_1$	$app(\{C_1, C_3, C_4\})$	$appC_1$
Royère 2	$appC_1$	$app(\{C_1, C_3, C_4, NA\})$	$appC_1$

Tableau 1. Décision finale en fonction de la redistribution du conflit

A travers divers exemples extrêmes que nous avons définis manuellement, c'est-à-dire des situations où le conflit entre les critères est élevé, nous avons constaté que la décision dépend de la manière dont le conflit est redistribué, surtout si nous utilisons le maximum de plausibilité comme outil de décision. Par contre, la décision est toujours la même si elle est prise en regardant le maximum de probabilité pignistique ou le maximum de crédibilité. Même si la décision est identique, les différentes techniques de redistribution du conflit mettent en valeur la certitude que l'on a dans la décision prise. Nous pouvons remarquer que la décision prise avec l'opérateur de Dempster est considérée sûre à 100%, alors que pour les autres opérateurs elle ne l'est pas, d'autres hypothèses simples ayant des probabilités pignistiques non nulles. Ainsi, on peut dire que l'opérateur de Dempster est parfois trop optimiste.

C.2 Modélisation des critères d'appariement

L'objectif de la fusion des informations est d'améliorer la robustesse et la qualité de la décision. Lorsque les informations sont imparfaites, plus on ajoute des connaissances, plus la décision sera sûre. Cependant, les connaissances sont de nature différente, elles sont parfois

hétérogènes ou basées sur des données incomplètes. De ce fait, le défi réside dans la représentation explicite des connaissances au sein d'un même formalisme.

Comme nous l'avons vu précédemment, afin de prendre une décision, nous fusionnons les critères d'appariement puis les candidats. Chaque critère d'appariement repose d'une part sur des mesures de distance provenant des propriétés des données géographiques, telles que la géométrie, la nature, le nom ou les relations spatiales entre les objets, et d'autre part sur des connaissances d'experts et sur des connaissances qui proviennent des données elles-mêmes et des spécifications.

Dans cette partie, nous présentons quelques critères d'appariement typiques basés sur des connaissances exprimées à l'aide de seuils, de distances ou de règles, ainsi que la représentation explicite de ces connaissances dans le cadre de la théorie des fonctions de croyance, c'est-à-dire l'initialisation des jeux de masses de croyance pour chacun des critères. Notons que les courbes illustrées ci-dessous ainsi que les seuils choisis sont des exemples typiques. Tout le processus de modélisation des critères d'appariement dépend des données utilisées.

La définition des règles est l'étape proprement dite d'initialisation des masses de croyance. Définir une règle consiste à attribuer une masse de croyance à chaque hypothèse définie dans le sous-ensemble du référentiel de définition, noté S_j, en analysant les autres connaissances, c'est-à-dire les seuils et les distances.

Nous rappelons le principe général de représentation explicite des connaissances. Etant donné un objet obj_1 appartenant au jeu de données JD_1, nous cherchons un objet homologue parmi les N candidats C_1, C_2… C_N dans le jeu de données JD_2. Ensuite chaque candidat C_i, i=1...N est analysé, c'est-à-dire qu'en fonction des distances et des écarts d'orientation mesurés entre l'objet obj_1 et le candidat C_i, nous exprimons nos croyances en ce candidat à travers trois hypothèses : $appC_i$ signifiant « le candidat C_i est l'homologue de l'objet obj1 », $\neg appC_i$ signifiant « le candidat C_i n'est pas l'homologue de l'objet obj_1 » et Θ signifiant « on ne sait pas », c'est-à-dire l'ignorance. Rappelons aussi que la somme des trois masses de croyance est toujours égale à 1.

C.2.1 Connaissances sur la géométrie

Les données géographiques ont une composante spatiale, la géométrie, et une composante attributaire, qui décrit l'objet géographique à travers des attributs tels que le nom, la nature, le nombre de voies, etc. Chaque objet géographique possède une géométrie qui peut être représentée généralement par une des trois primitives : point, ligne ou surface. La géométrie décrit donc la localisation, la forme dans le cas de la ligne et de la surface, ainsi que les relations spatiales implicites entre les objets géographiques.

La géométrie des objets géographiques est une information toujours présente dans les bases de données géographiques, et donc toujours exploitable. Comme nous avons pu le constater dans le chapitre A, la plupart des travaux d'appariement de données s'appuient sur la géométrie car elle est la plus discriminante.

Il existe de nombreuses mesures qui permettent d'exploiter et de comparer la géométrie des objets géographiques. Ainsi, pour comparer les localisations, on peut utiliser une mesure de distance telle que la distance euclidienne, la distance de Hausdorff ou la distance surfacique. La forme des objets peut être comparée à travers des mesures qui évaluent l'écart

de forme, comme par exemple la fonction à distance radiale, la fonction angulaire ou la sinuosité. Lorsque nous avons à comparer des lignes, nous pouvons analyser leur orientation.

C.2.1.1 Initialisation des masses de croyance pour le critère d'écart de position

Dans notre approche, le critère d'écart de position a fait ses preuves avec la distance euclidienne appliquée aux données isolées représentant les points remarquables du relief [Olteanu, 2007], et avec la distance de Hausdorff appliquée aux réseaux routiers [Olteanu-Raimond et Mustière, 2008]. Nous croyons que le critère reste exploitable pour d'autres mesures de distance.

La représentation explicite des connaissances pour le critère d'écart de position est illustrée dans le Tableau 2.

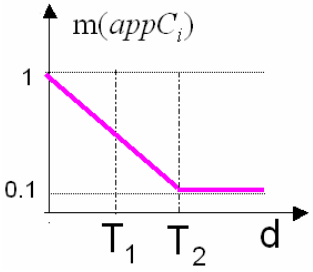
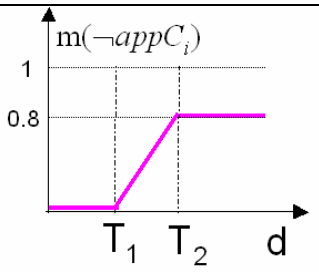
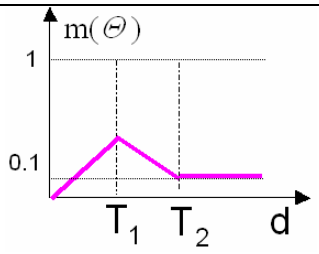
Hypothèse	Critère d'écart de position
$appC_i$	
$\neg appC_i$	
Θ	

Tableau 2. Représentation des connaissances pour le critère d'écart de position

La courbe figurant dans la première ligne du Tableau 2 représente la masse de croyance attribuée à l'hypothèse $appC_i$, c'est-à-dire « le candidat C_i est l'homologue de l'objet obj_1 ». Elle signifie le fait que plus le candidat à l'appariement C_i est proche de l'objet obj_1, plus il y a de chances que celui-ci soit l'objet homologue. Elle signifie aussi qu'à partir d'un seuil donné T_2, nous considérons que cette hypothèse devient presque impossible, quelle que soit la distance. Afin d'éviter l'élimination définitive d'un candidat qui est relativement éloigné, la masse de croyance attribuée à cette hypothèse, m($appC_i$) n'est pas nulle mais égale à 0,1 dans notre exemple. Précisons qu'au delà d'une distance c'est effectivement improbable puisque nous ne sommes plus dans le cadre de discernement.

La courbe de la deuxième ligne représente la masse de croyance attribuée à l'hypothèse $\neg appC_i$, signifiant « le candidat C_i n'est pas l'homologue de l'objet obj_1 ». Afin de modéliser l'imprécision sur la localisation des objets géographiques, nous avons introduit un deuxième seuil T_1 qui représente la résolution des données. Si la distance est inférieure au seuil T_1, nous considérons cette proposition très improbable, par contre si la distance est comprise entre T_1 et T_2, la proposition devient de plus en plus crédible et enfin lorsque la distance est supérieure au seuil T_2, la proposition devient très crédible. Plus la distance est élevée, plus nous croyons que la proposition est vraie.

Enfin, la courbe figurant dans la dernière ligne représente la masse de croyance attribuée à l'ignorance, $m(\Theta)$. Pour ce critère d'écart de position, l'ignorance est importante lorsque la distance est dans le voisinage de T_1, c'est-à-dire lorsque le candidat n'est ni suffisamment éloigné pour conclure d'une manière sûre que ce n'est pas lui, ni suffisamment proche pour conclure que c'est lui le vrai homologue.

C.2.1.2 Initialisation des masses de croyance pour le critère écart d'orientation

Le critère d'orientation est représenté dans le Tableau 3. Il peut être utilisé pour les objets géographiques linéaires et il consiste dans cet exemple à évaluer le degré de co-linéarité local entre l'objet obj_1 et le candidat C_i. Nous avons expliqué dans le chapitre A la manière d'étudier l'écart d'orientation entre les lignes. Nous rappelons brièvement le principe.

Le critère d'orientation mesure l'écart entre les orientations des tangentes à obj_1 et au candidat C_i respectivement au point de obj_1 le plus proche du candidat C_i, et au point du candidat C_i le plus proche de obj_1. Si la valeur de l'angle entre les deux objets est proche de 0, alors les objets sont relativement parallèles et ils ont la même direction. Si la valeur de l'angle est proche de π, alors les objets sont parallèles et dans la direction opposée. Par contre, si la valeur de l'angle est proche de $\pi/2$, alors les objets sont perpendiculaires.

En raison notamment de la complexité des données géographiques linéaires, par exemple les réseaux routiers, il est possible que plusieurs candidats à l'appariement soient localement parallèles à l'objet *obj1*. Ainsi, le fait que deux objets linéaires aient la même orientation ne signifie pas qu'ils soient homologues.

D'autre part, en raison de niveaux de détail ou de segmentation différents, deux objets homologues peuvent ne pas avoir la même orientation. En conséquence, si l'écart d'orientation est dans le voisinage de $\pi/2$, nous croyons que les deux objets ne sont pas homologues mais nous n'avons aucune certitude.

De plus, cette mesure ne présente pas une grande fiabilité parce elle est calculée localement, c'est-à-dire que l'orientation entre deux arcs est déterminée au point le plus proche, ce qui est une forte approximation de la colinéarité. Cela veut dire que l'écart d'orientation entre deux objets linéaires peut être nul, sans toutefois que les deux objets soient vraiment semblables.

Les raisons que nous venons d'énumérer font que le critère d'orientation est peu fiable, et que l'ignorance a un rôle important. Ainsi, l'ignorance permet de gérer l'imprécision due à la mesure utilisée et aux caractéristiques des données.

Plus précisément, si les objets sont relativement parallèles, c'est-à-dire si l'angle θ est soit proche de 0 soit proche de π, nous croyons que l'objet obj_1 et le candidat C_i peuvent être homologues et nous attribuons une masse de croyance de 0,5 à l'hypothèse « le candidat C_i

est l'homologue de l'objet obj_1 », mais aussi à l'ignorance, afin d'exprimer le fait que le critère seul n'est pas suffisamment significatif. La masse de croyance attribuée à l'hypothèse « le candidat C_i n'est pas l'homologue de l'objet obj_1 » est le complément de $(m(appC_i)+m(\Theta))$.

Plus l'écart d'orientation entre deux objets linéaires tend vers $\pi/2$, plus la masse de croyance attribuée à l'hypothèse $appC_i$ diminue et proportionnellement avec cette diminution, la masse de croyance attribuée à l'hypothèse $\neg appC_i$ augmente.

Lorsque les objets sont perpendiculaires, c'est-à-dire que l'angle θ est égal à $\pi/2$, nous attribuons une masse de croyance importante à l'hypothèse « le candidat C_i n'est pas l'homologue de l'objet obj_1 » et à l'ignorance, afin d'exprimer notre doute relatif à la mesure utilisée et aux caractéristiques des données.

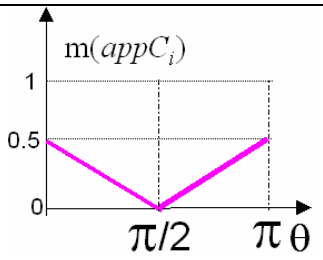
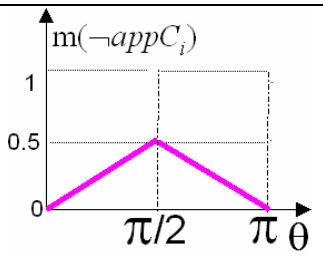
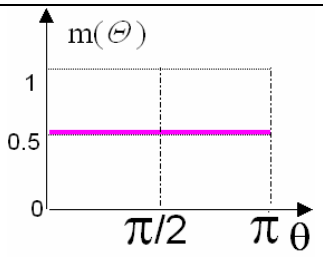
Hypothèse	Critère Orientation
$appC_i$	
$\neg appC_i$	
Θ	

Tableau 3. Représentation des connaissances pour le critère d'orientation

C.2.2 Connaissances sur la sémantique

Dans une base de données géographiques, les données géographiques sont regroupées dans des thèmes tels que le réseau routier, l'hydrographie, les points remarquables du relief, l'administratif, l'occupation du sol. Chaque thème peut contenir plusieurs classes d'objets. Cependant, il s'avère qu'au sein de la même classe d'objets il peut y avoir des objets géographiques qui n'ont pas exactement la même nature, en fonction du niveau de détail de la classification. Ainsi, les objets géographiques possèdent un attribut appelé par exemple *nature* ou *type* qui précise ce que nous appelons l'information sémantique.

L'information sémantique a fait ses preuves dans le processus d'appariement de schémas et elle est très souvent employée. Malgré cela, la sémantique est très peu utilisée dans le processus d'appariement de données géographiques, plusieurs auteurs affirmant que la sémantique pourrait améliorer un processus d'appariement, mais son utilisation reste très limitée à cause de la difficulté de gérer son hétérogénéité parmi les jeux de données et ses imperfections.

C.2.2.1 Problématique

Pour simplifier, nous supposons désormais que l'attribut qui désigne la sémantique des objets géographiques est l'attribut *nature*. De nombreux concepts géographiques existent et en fonction de leur interprétation et de leur but final, deux concepts différents peuvent désigner la même chose ou encore, un même concept peut désigner deux choses sensiblement différentes dans des applications différentes. Ce problème est souvent appelé l'hétérogénéité sémantique. Par exemple, une entité du monde réel représentant un barrage est vue par un opérateur de saisie comme un ouvrage physique, tandis que la même entité du monde réel est vue par un autre opérateur comme une zone d'eau.

L'attribut *nature* ne présente donc pas le même niveau de détail pour toutes les données. Par exemple, dans un jeu de données il peut y avoir un regroupement de concepts dans une même valeur de l'attribut : « sommet, crête, colline », tandis que dans l'autre jeu de données les concepts sont distingués.

De plus, les données géographiques, telles que les montagnes, les sommets, les pics, les vallées, les cols, etc. sont imprécises d'une part par définition, par exemple 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 de données peuvent être floues, comme par exemple entre sommet et pic. Les concepts de sommet et de pic sont proches d'un point de vue sémantique, et dans la pratique ils sont très souvent confondus.

Cependant, l'analyse des données géographiques montre qu'il existe des objets géographiques qui ont le même toponyme, qui sont proches les uns des autres, mais qui ne sont pas de la même nature et par conséquent ne peuvent pas être mis en correspondance, comme par exemple un sommet avec un col.

Toutes ces difficultés font que la comparaison de l'attribut *nature* n'est pas immédiate, et qu'une simple comparaison des valeurs d'attribut sera inexploitable. C'est pour cela que nous devons évaluer plus finement le degré de ressemblance sémantique entre les concepts.

Nous considérons que la prise en compte de la nature des objets géographiques peut être utile dans le processus d'appariement. Cependant, un critère d'appariement basé sur la sémantique est moins discriminant qu'un critère d'appariement basé par exemple sur l'écart de position. Il est évident que la condition nécessaire pour qu'un objet géographique obj_1 soit apparié avec un candidat est que les deux objets appartiennent à la même classe ou aient une nature proche, mais tous les candidats de la même classe ne sont pas appariés avec l'objet obj_1.

Si pour certaines propriétés, telles que la géométrie, la topologie, les attributs quantitatifs ou même qualitatifs, des mesures de distance existent déjà, qui peuvent s'appliquer directement aux propriétés pour mesurer le degré de ressemblance sémantique entre deux concepts, la seule comparaison des chaînes des caractères désignant la nature des objets ne suffit pas. La question qui se pose est : Comment définir une mesure qui permette d'évaluer

cette distance sémantique ? Une solution est de demander à des experts qu'ils évaluent l'écart sémantique entre les données, puis d'exploiter leur évaluation pour pouvoir calculer une distance sémantique. Une autre solution est de s'appuyer sur d'autres sources telles qu'un dictionnaire, une taxonomie ou une ontologie de domaine afin de déterminer les distances sémantiques.

Utiliser une ontologie ou une taxonomie de domaine a l'avantage entre autres, de faciliter la mise en oeuvre et l'automatisation de l'étape de détermination de la distance sémantique. Dans notre approche, afin d'évaluer la pertinence de l'évaluation de la ressemblance sémantique à partir d'une taxonomie, nous avons comparé deux méthodes. D'une part nous avons réalisé une enquête auprès d'experts. Elle consiste à leur demander d'attribuer des notes entre 0 et 1 pour évaluer l'écart sémantique entre les concepts ou les groupes de concepts. D'autre part, nous avons utilisé une taxonomie de domaine établie au laboratoire COGIT de l'IGN [Abadie et Mustière, 2008].

A titre illustratif, nous présentons sur la Figure 68 un extrait d'une taxonomie représentant le thème « *points remarquables du relief* ».

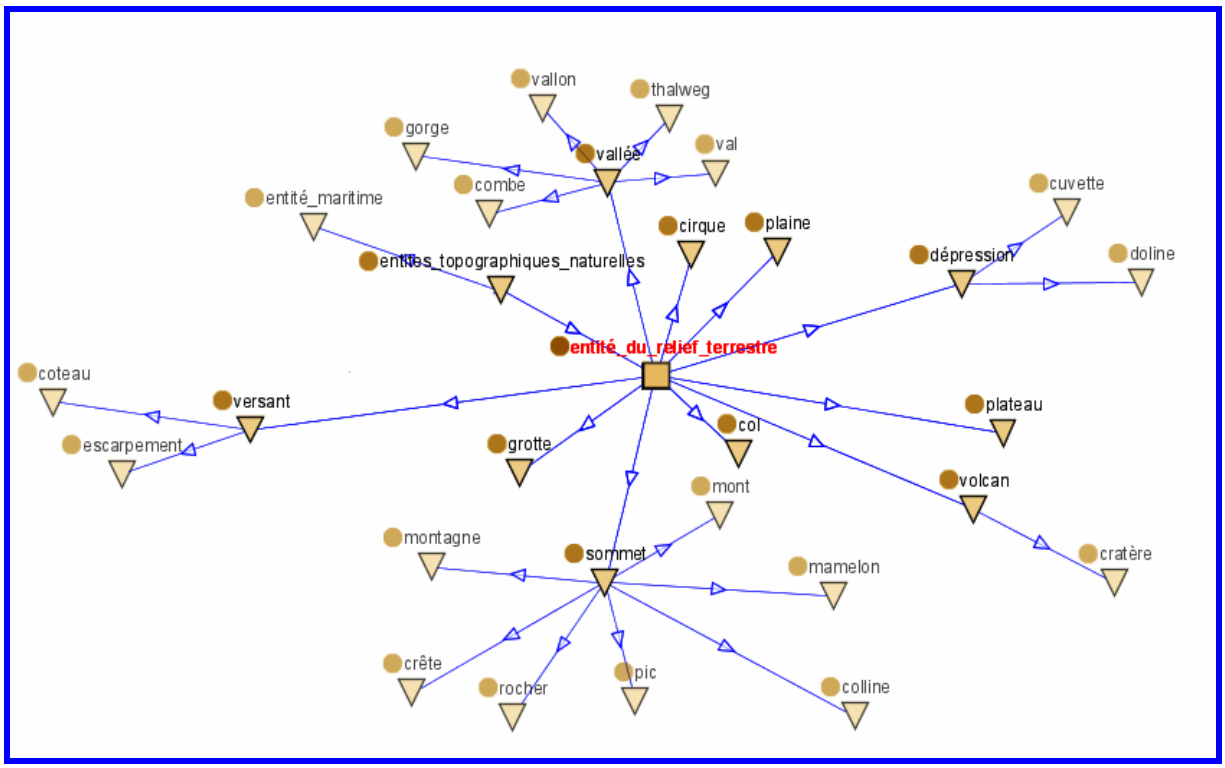


Figure 68. Extrait de la taxonomie réalisée pour les points remarquables du relief par [Abadie et Mustière, 2008]

La comparaison des deux études montre d'une part que la taxonomie est utilisable, car globalement elle est cohérente avec les experts, et d'autre part que les mesures de distance sémantique méritent cependant d'être améliorées. Nous reviendrons plus en détail sur ces deux études dans la partie expérimentation du chapitre D.

C.2.2.2 Initialisation des masses de croyance

Afin de comparer les différentes natures, nous avons utilisé la distance sémantique d_S définie par [Wu et Palmer, 1994], d'une part en raison de sa simplicité et d'autre part en raison de son unanimité dans la communauté. Par exemple, la distance sémantique entre un « pic » et un « sommet » est de 0,2, tandis que la distance sémantique entre un « pic » et un « col » est de 0,66. Cela montre qu'un pic est plus proche d'un sommet que d'un col.

La représentation des connaissances pour le critère sémantique est illustrée sur le Tableau 4. Le critère sémantique est basé sur trois types de connaissance : un seuil empirique T, les distances sémantiques calculées à partir des données comme nous l'avons décrit ci-dessus, et les règles. Nous proposons d'utiliser un seuil empirique déterminé grâce aux notes fournies par les experts qui nous ont guidés pour fixer la limite à partir de laquelle deux concepts ne peuvent plus être considérés comme ressemblants au sens sémantique du terme, conformément à l'Annexe 2. Nous appelons règles, les connaissances qui sont utilisées pour définir les masses de croyance en analysant les deux autres connaissances à savoir les distances sémantiques et le seuil T.

Le critère sémantique n'est pas le critère le plus discriminant, il nécessite la représentation de conditions nécessaires mais pas suffisantes. Etant donnée la spécificité des objets géographiques, il est souvent possible qu'il existe un grand nombre de candidats de même nature que l'objet *obj1* en cours d'analyse. Nous rappelons que notre approche consiste à analyser chaque candidat indépendamment des autres. Ainsi, dans l'hypothèse où plusieurs candidats aient la même nature, c'est-à-dire une distance sémantique égale à 0, et que nous attribuons une masse de croyance importante à l'hypothèse « le candidat C_i est l'homologue de l'objet *obj1* » et cela pour chacun des candidats, nous avons un fort conflit lors de la combinaison des candidats et donc aucune décision ne peut être prise. Ce conflit est tout à fait normal, puisque nous soutenons sans discrimination tous les candidats qui ont la même nature.

Afin que cela ne se produise pas, nous proposons une modélisation prudente, c'est-à-dire que même si deux objets ont la même nature, nous n'affirmons pas avec une grande certitude que les deux objets sont homologues.

La courbe illustrée sur la première ligne signifie que nous considérons que si la distance sémantique entre l'objet obj_1 et le candidat C_i est égale à 0, la masse de croyance attribuée à l'hypothèse C_i : « le candidat C_i est l'homologue de l'objet obj_1 » est égale à 0,5 dans notre exemple, donc nous n'attribuons pas de forte croyance à ce candidat. Plus la distance sémantique tend vers le seuil S, plus cette hypothèse devient peu plausible, et la masse de croyance diminue. Enfin, si la distance sémantique est supérieure au seuil T, nous croyons que le candidat C_i n'est pas le bon candidat, mais afin de ne pas l'éliminer en raison de l'imperfection de la distance sémantique, nous attribuons une masse de croyance très faible à l'hypothèse $appC_i$, mais non nulle. Nous pouvons cependant donner une masse non négligeable de 0,5 dans la mesure où grâce au cadre de discernement, nous ne comparons pas tous les objets, mais le sous-ensemble des objets probables.

La courbe illustrée sur la deuxième ligne du Tableau 4 représente la masse de croyance attribuée à l'hypothèse $\neg appC_i$: « le candidat C_i n'est pas l'homologue de l'objet obj_1 ». Rappelons que l'hypothèse $\neg appC_i$ fait partie du cadre de discernement, étant composée de toutes les hypothèses du cadre de discernement sauf l'hypothèse $appC_i$. Si la distance sémantique est égale à 0, cette hypothèse est impossible, la masse de croyance étant partagée

d'une manière égale entre l'hypothèse $appC_i$ et l'ignorance. Plus la distance sémantique s'accroît, plus l'hypothèse $\neg appC_i$ devient plausible, la masse de croyance augmentant proportionnellement avec la distance. A partir du seuil S, cette hypothèse est crédible : la masse de croyance est importante, et le doute n'est pas nul, c'est-à-dire que la masse de croyance attribuée à l'ignorance est égale à 0,1. Cette modélisation a été réalisée dans le but de prendre en compte des erreurs potentielles dans la nature des objets géographiques ou lorsque les classifications des deux jeux de données à apparier sont hétérogènes et donc difficiles à comparer et surtout afin de prendre en compte l'imperfection de la distance sémantique utilisée.

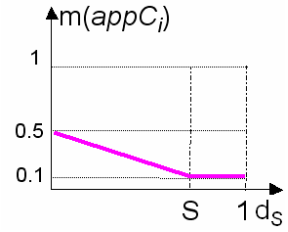
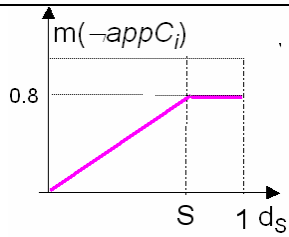
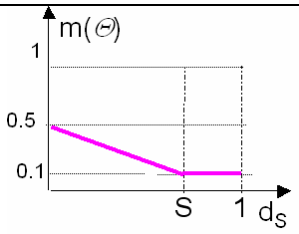
Hypothèse	Critère Sémantique
$appC_i$	
$\neg appC_i$	
Θ	

Tableau 4. Représentation des connaissances pour le critère sémantique

Enfin, la dernière courbe représente la masse de croyance attribuée à l'ignorance. Comme nous pouvons le constater, cette courbe est similaire à celle représentant l'hypothèse $appC_i$. L'idée de base de cette modélisation est la suivante : si deux objets sont proches du point de vue sémantique, il y a des chances qu'ils soient homologues, mais nous avons un doute, et si deux objets sont éloignés, ils ne sont pas homologues, mais nous avons également un doute, qui est plus faible que dans le cas précédent. Cela signifie que dans le processus d'appariement, nous devons représenter des conditions nécessaires mais pas suffisantes. Ainsi, pour que deux objets soient homologues il est nécessaire que les attributs « nature » des objets soient proches, c'est-à-dire qu'ils aient une distance sémantique faible, mais tous les objets ayant la même nature ne sont pas homologues.

C.2.3 Connaissances sur la toponymie ou sur les noms des objets géographiques

C.2.3.1 Problématique

Pour certains thèmes tels que le découpage administratif, le réseau routier ou le réseau hydrographique, l'attribut qui désigne le nom du lieu ou le nom de l'objet géographique peut être considéré comme un identifiant. Théoriquement, une ville a le même nom dans toutes les bases de données, une route a le même numéro de route. Cependant, dans la réalité il y a de nombreuses raisons qui font que des objets homologues peuvent avoir des toponymes différents :

- les erreurs de frappe lors de la saisie des données,
- la variabilité linguistique issue des différences de prononciation, par exemple « Munhoa » et « Monhoa »,
- 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,
- l'utilisation du nom officiel et du nom d'usage pour la même entité géographique,
- l'absence du nom dans les jeux de données, par exemple il se peut que pour le thème routier seulement les routes importantes ont un numéro de route,
- la mise à jour décalée des bases de données géographiques. Par exemple, le nom d'une entité géographique a changé, la mise à jour a été faite dans une BDG mais pas dans l'autre,
- les nomenclatures différentes, par exemple dans une base de données géographiques française, une autoroute est notée *A11*, tandis que dans une base de données géographiques européenne cette autoroute est notée *E50* suivi d'un autre numéro.

Pour toutes ces raisons, l'enjeu est de définir une mesure qui permette d'évaluer la ressemblance entre les toponymes ou entre les noms d'objet, et de réaliser une représentation explicite des connaissances basées sur la toponymie capable de gérer les imperfections présentes dans les données et surtout l'absence de nom ou de toponyme.

C.2.3.2 Initialisation des masses de croyance

Le nom des lieux ou les noms des objets sont des informations importantes dans un jeu de données, elles peuvent être considérées comme un identifiant. Cependant, supposons que cette information existe dans les jeux de données, à travers un attribut. Il existe des cas où la valeur de l'attribut est plus ou moins remplie. Ainsi, pour remédier au manque d'information, nous avons défini deux critères, un *critère toponymique* adapté surtout aux noms de lieu et il est utilisé lorsque la valeur de l'attribut est toujours remplie, et un *critère nom d'objet* qui peut être utilisé même lorsque la valeur de l'attribut n'est pas toujours remplie, ce dernier critère étant discret et adapté aux noms d'objet géographique ou aux numéros de route par exemple.

Critère toponymique

Comme nous pouvons le remarquer dans le Tableau 5, les courbes sont différentes de celles du critère de distance ; elles expriment le fait que nous sommes moins confiants en ce critère lorsque l'écart toponymique est important. Pour cela, nous proposons de diminuer la masse de croyance attribuée à l'hypothèse $\neg appC_i$ et d'augmenter la masse de croyance attribuée à l'ignorance.

La courbe sur la première ligne du Tableau 5 illustre la masse de croyance attribuée à l'hypothèse $appC_i$, « le candidat C_i est l'homologue de l'objet obj_1 ». Elle exprime le fait que plus deux toponymes se ressemblent, c'est-à-dire que la distance toponymique est faible, plus l'hypothèse $appC_i$ est crédible. Ainsi, si la distance toponymique est nulle, c'est-à-dire s'il s'agit du même toponyme, nous croyons que les deux objets sont homologues en attribuant à l'hypothèse $appC_i$ une masse de croyance égale à l'unité. Sinon, la masse de croyance diminue proportionnellement avec la distance toponymique. Lorsque la distance toponymique est supérieure à un seuil S (par exemple 30% des lettres ne se ressemblent pas), nous croyons que cette hypothèse est presque impossible, mais afin de ne pas éliminer le candidat, nous attribuons à l'hypothèse $appC_i$ une masse de croyance très faible, $m(appC_i)=0,1$.

Les courbes illustrées sur la deuxième et la troisième ligne de la première colonne du Tableau 5 sont sensiblement identiques. Si la distance d_T est inférieure au seuil S, les masses de croyance dans les hypothèses $\neg appC_i$, « le candidat C_i n'est pas l'homologue de l'objet obj_1 » et Θ, « on ne sait pas » augmentent de la même manière proportionnellement avec la distance d_T. Par contre, si la distance d_T est supérieure au seuil S, les masses de croyance attribuées aux hypothèses $\neg appC_i$ et Θ sont égales à 0,5 et 0,4, pour toute distance d_T. Cette représentation illustre le fait que lorsque les deux toponymes sont identiques, il est crédible que les deux objets soit homologues, tandis que s'ils ne le sont pas, nous n'avons aucune certitude que les deux objets ne sont pas homologues. Ceci explique pourquoi nous avons donné autant d'importance à l'ignorance.

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 Gaulle » et « Place de l'Etoile » ou lorsque nous avons deux toponymes exprimés en dialectes différents.

D'une manière générale nous pouvons affirmer que le critère toponymique a un poids important lorsque les toponymes sont identiques ou se ressemblent fortement, parce que nous considérons que le toponyme est presque un identifiant. Cette hypothèse est pertinente lorsque nous utilisons un cadre de discernement basé sur la distance. Sinon, cette hypothèse ne serait pas valable, puisqu'il existe plusieurs entités ayant le même toponyme sur le territoire français. Lorsque les toponymes se ressemblent très peu ou pas du tout, le critère a un poids moins important, laissant la possibilité aux autres critères de se prononcer d'une manière plus décisive.

Nous remarquons qu'en attribuant à l'ignorance une certaine importance, la représentation explicite des connaissances liées à la toponymie est prudente, c'est-à-dire que si nous sommes sûrs, nous nous prononçons pour une hypothèse, sinon, nous restons impartiaux. Ce type de représentation dite prudente a comme conséquence un faible conflit lors de la combinaison du critère toponymique avec les autres.

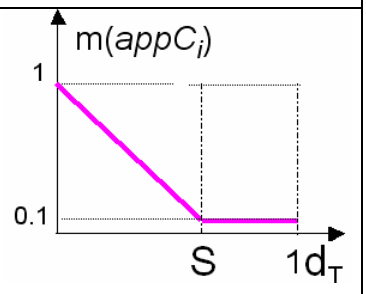
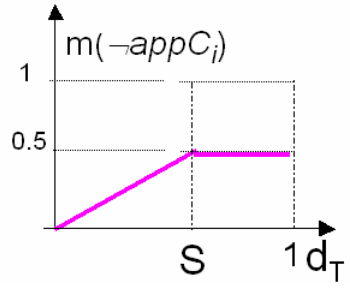
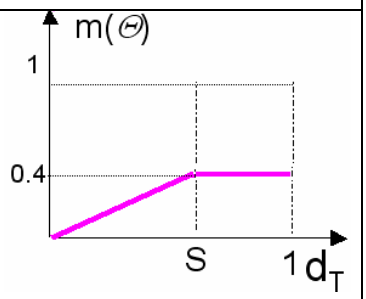
Hypothèse	Critère Toponymique
$appC_i$	
$\neg appC_i$	
Θ	

Tableau 5. Représentation des connaissances pour le critère toponymique

Critère nom d'objet

Comme nous l'avons déjà dit ci-dessus, le critère nom d'objet est un critère d'abord adapté aux données qui présentent des incomplétudes et aux numérotations, pour lesquelles il n'y a pas de sens à évaluer l'écart entre les lettres. Il s'agit donc des données pour lesquelles les attributs désignant les noms des objets géographiques ne sont pas toujours remplis, notamment les réseaux routiers ou les réseaux hydrographiques. Le critère nom d'objet est aussi adapté à des entités géographiques composées de plusieurs objets géographiques. Par exemple, une route est composée de plusieurs tronçons de route.

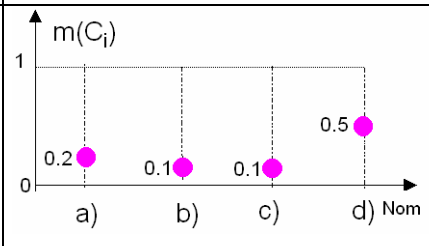
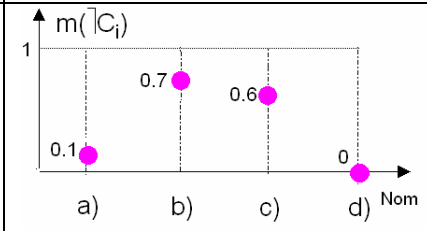
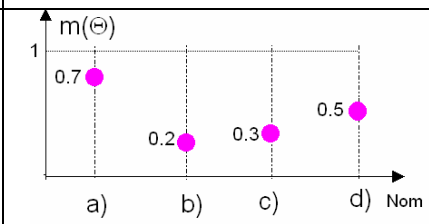
Même si l'attribut n'est pas toujours rempli ou s'il est rempli mais des différences existent, il contient de l'information qui peut nous aider à apparier les objets géographiques. Ces différences sont dues, entre autres, aux rythmes de mise à jour différents des bases de données géographiques ou à l'existence de plusieurs nomenclatures, par exemple pour les réseaux routiers.

Ce critère est basé sur des connaissances liées à l'importance des objets géographiques. Plus précisément, si par exemple pour un couple d'objets en cours d'analyse aucun ne possède de nom, il y a plus de chances que les deux objets soient homologues que pour un couple d'objets pour lequel un objet possède un nom et l'autre pas. Notre réflexion s'appuie sur l'hypothèse que les objets sans nom sont moins importants que ceux ayant un nom, sous

réserve, bien évidemment, qu'il ne s'agisse pas d'une erreur de saisie. Ainsi, si nous revenons à notre exemple antérieur des deux couples d'objets, sous l'hypothèse que nous venons de formuler, le premier couple est composé de deux objets importants, tandis que le deuxième est composé d'un objet important et d'un autre moins important.

Toutes ces raisons font que ce critère est moins important que les autres et surtout qu'il est moins discriminant car plusieurs candidats à l'appariement peuvent avoir le même nom. Il est défini en fonction de l'attribut censé représenter le nom d'un objet géographique. Pour ce critère l'ignorance a un poids significatif lorsque la valeur de l'attribut n'est pas remplie.

La première colonne du Tableau 6 illustre le jeu de masses pour ce critère, correspondant aux quatre cas illustrés dans la deuxième colonne du Tableau 6:

Hypothèse	Critère nom d'objet
$appC_i$	
$\neg appC_i$	
Θ	

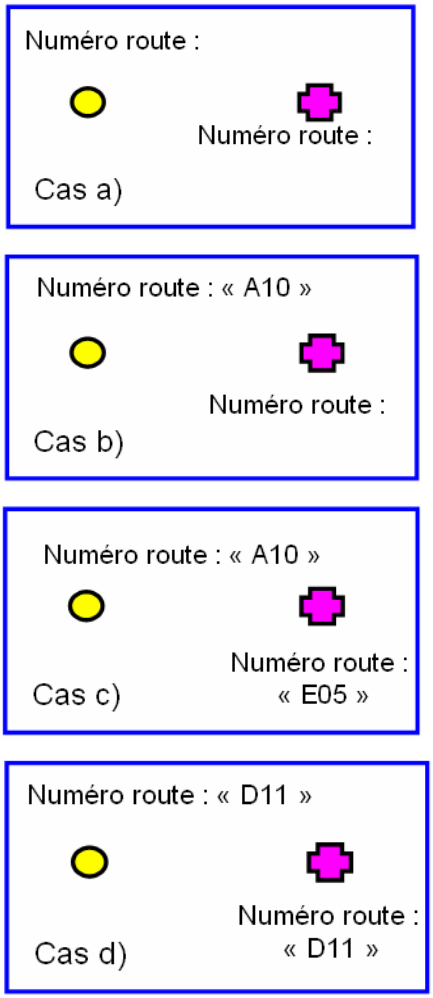


Tableau 6. Représentation des connaissances pour le critère nom d'objet

- cas a) : la valeur de l'attribut n'est pas remplie pour les deux objets, le candidat C_i et l'objet obj_1 en cours d'analyse. Dans ce cas, le critère ne peut pas prendre de décision ; il attribue une masse de croyance importante à l'ignorance, m(Θ)=0,7. Le complément de l'ignorance est divisé entre les hypothèses $appC_i$, « le candidat C_i est l'homologue de l'objet obj_1 » et $\neg appC_i$, « le candidat C_i n'est pas l'homologue de l'objet obj_1 ». Le fait

que la masse attribuée à l'ignorance ne soit pas égale à l'unité est dû justement à l'hypothèse que nous avons faite que, sauf erreur, deux objets qui n'ont pas de nom, ont la même importance.

- cas b) : la valeur de l'attribut est remplie seulement pour un des deux objets. Dans ce cas, nous croyons que les deux objets ne sont pas homologues. Par conséquent, nous attribuons une valeur importante à l'hypothèse $\neg appC_i$. Afin de prendre en compte le cas d'un oubli, il existe un doute exprimé à travers la masse de croyance, même faible, attribuée à l'ignorance.
- cas c) : les deux objets possèdent des noms différents. Dans ce cas, nous croyons que les deux objets ne sont pas homologues et nous attribuons une masse de croyance importante à cette hypothèse. Afin de gérer les cas où deux objets représentent la même réalité mais ont des noms différents à cause des nomenclatures ou des langues différentes, le critère ne rejette pas complètement l'hypothèse qu'ils sont homologues. De ce fait, $m(appC_i) = 0,1$.
- cas d) : les objets ont le même nom. Dans ce cas, il est très crédible que les deux objets soient homologues. Il existe des entités du monde réel qui sont représentées dans les bases de données par plusieurs objets géographiques. Par exemple une route est représentée par plusieurs tronçons de route. En conséquence, dans le processus d'appariement il s'avère que le cas où plusieurs candidats ont le même nom arrive fréquemment. Ainsi, afin de représenter de telles connaissances, nous attribuons une masse de croyance égale à 0,5 à l'hypothèse «le candidat C_i est l'homologue de l'objet obj_1 » et à l'ignorance. Dans ce cas, il est très crédible que les deux objets soient homologues.

C.2.4 Connaissances sur le voisinage

Les objets d'une base de données géographiques ont des relations spatiales qui sont décrites à travers la topologie, les mesures de distance ou d'orientation, la densité, etc. Par exemple, il est possible de décrire un objet géographique par rapport à un autre : une autoroute est à l'est de la ville, la ville de Paris se trouve à x km de Rome, etc. Ces informations sont implicites et sont utilisées dans l'analyse des données ou dans les processus qui manipulent les données.

C.2.4.1 Problématique

La topologie joue un rôle très important dans le processus d'appariement des réseaux. Les réseaux tels que le réseau routier ou le réseau hydrographique se caractérisent par la connexité de leurs éléments. Il nous semble donc utile d'exploiter cette information dans le cas des réseaux.

Un processus d'appariement basé sur la topologie et qui utilise des réseaux peut améliorer la qualité du processus. Il peut être alors efficace lorsque la topologie des bases de données géographiques est sans erreur. Cependant, un processus d'appariement basé seulement sur la topologie ne gère pas les erreurs de topologie. Par exemple, il peut y avoir d'une part des erreurs internes à une base de données, par exemple lorsque deux routes proches ne sont pas connectées entre elles ou lorsqu'une route n'est connectée à aucune autre route dans la base, et d'autre part des incohérences dans l'organisation topologique des deux bases de données géographiques : des routes connectées dans un jeu de données mais pas dans l'autre ou des routes connectées mais pas de la même manière. Les causes de ces erreurs peuvent être les rythmes différents de mise à jour, l'erreur pendant la saisie des données, une mauvaise

généralisation si une base de données est issue d'un processus de généralisation, etc. Dans ce dernier cas, la difficulté consiste à identifier la source de l'erreur et à déterminer quelle organisation topologique est juste.

Dans la suite de cette partie nous donnons à titre d'illustration un exemple d'utilisation de l'information topologique dans le processus d'appariement. Dans cet exemple nous utilisons les relations topologiques dites de voisinage (au sens de l'adjacence). Ces dernières sont issues des données géographiques sur lesquelles ont été déterminées les relations topologiques de connexion, c'est-à-dire les nœuds et les arcs adjacents. Ainsi, un arc peut avoir ou non des arcs incidents au nœud initial et des arcs incidents au nœud final.

Le principe de ce critère est le suivant : deux objets A et B se ressemblent si A a des relations topologiques avec ses voisins comparables avec les relations topologiques de l'arc B avec ses voisins.

C.2.4.2 Initialisation des masses de croyance

Afin de faire une analyse globale de la base de données avec laquelle l'appariement d'un objet dépend de l'appariement de ses voisins, et les relations entre les objets géographiques, sont prises en compte, nous avons défini un critère de voisinage. Pour y parvenir, le processus d'appariement est un peu plus complexe que celui décrit dans la partie C.1, et se décompose en deux grandes parties. Dans un premier temps, le processus d'appariement est mis en œuvre en utilisant par exemple n critères basés sur la géométrie ou les attributs des objets géographiques. Puis, dans un second temps, les résultats de la première étape sont analysés pour initialiser les masses de croyance du critère de voisinage, et le processus final est mis en œuvre avec les n+1 critères, y compris le critère de voisinage. Notons que nous pouvons boucler au besoin. Le processus est illustré sur la Figure 69.

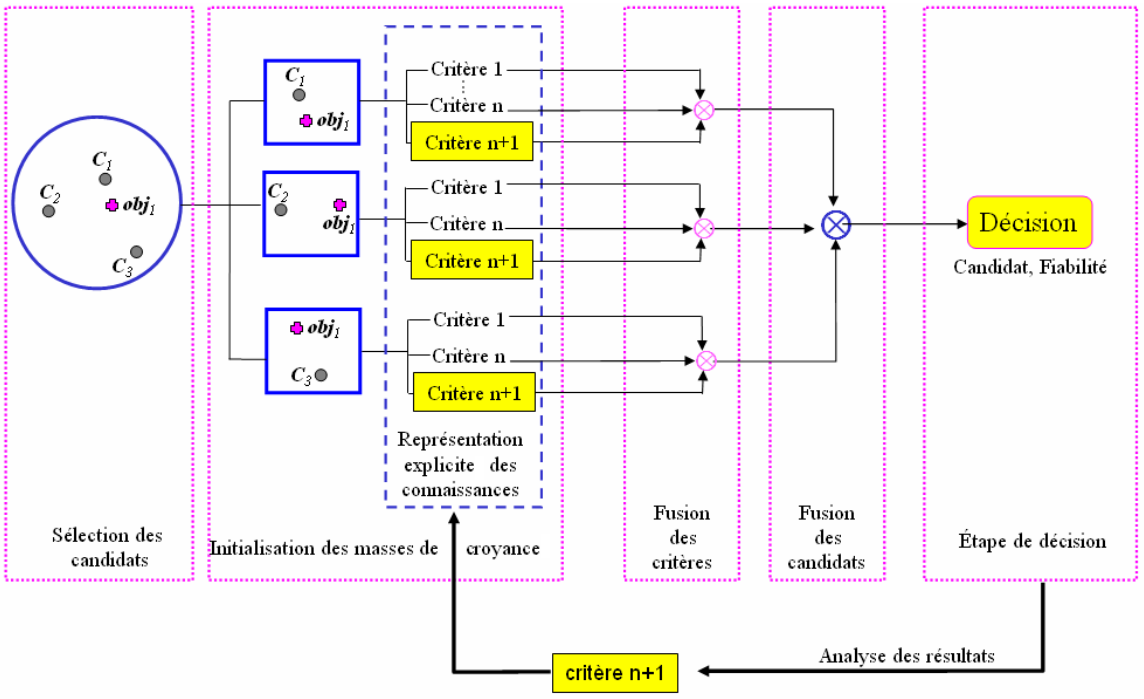


Figure 69. Processus d'appariement basé sur des résultats d'appariement

Ce critère de voisinage est adapté surtout pour les données organisées en réseau, qui se caractérisent par le fait que les objets géographiques ont des relations spatiales entre eux. Il est de même adapté pour l'appariement 1 : n, c'est-à-dire que N objets *obj1*, *obj2*, …*objN* d'un jeu de données plus détaillé, appelé JD1, correspondent à un seul objet dans l'autre jeu de données moins détaillé, appelé JD2. D'autres critères de voisinage, étudiant par exemple les décalages de position systématiques [Samal *et al.*, 2004] pourraient être définis selon le même principe.

Nous détaillons ci-dessous un exemple typique d'analyse des résultats d'appariement qui différencie d'une part les objets appariés et d'autre part ceux qui n'ont pas été appariés en première passe du processus d'appariement. Afin de mieux comprendre la manière dont nous modélisons le critère voisinage, nous reprenons l'idée de base du processus. Pour chaque objet *obj1* appartenant à JD1, nous sélectionnons des candidats à l'appariement dans le jeu de données moins détaillé. Comme nous l'avons décrit auparavant, chaque candidat C_i, i=1..N, est analysé indépendamment des autres candidats. Dans le but de modéliser le critère voisinage, d'abord nous cherchons à savoir si l'objet *obj1* a été apparié dans le premier processus. Si c'est le cas, nous utilisons le critère voisinage-objets appariés, sinon, nous utilisons le critère voisinage-objets non-appariés. Les deux critères sont décrits ci-après.

Critère voisinage-objets appariés

Pour les objets appariés lors de la première passe du processus, la modélisation des masses de croyance du critère voisinage est illustrée dans le Tableau 7 à gauche. Supposons que nous sommes en train d'analyser l'objet *obj1* appartenant à JD1 et le candidat C_i appartenant à la base de données moins détaillée JD2.

Nous analysons également si le candidat C_i est apparié lors de la première passe. Si c'est le cas, ses n objets homologues sont groupés en groupes connexes. Nous appelons groupe connexe un ensemble d'arcs connectés entre eux. Notons que cette analyse est faite seulement si l'objet *obj1* fait partie des objets homologues du candidat C_i. Si un seul groupe G_1 a été identifié, le groupe est évalué comme étant sûr : cas d) de la première colonne du Tableau 7. Cela veut dire que les objets appartenant à ce groupe ont été bien appariés en première passe. Sinon, si plusieurs groupes sont trouvés, nous analysons la manière dont les voisins du candidat C_i sont appariés lors de la première étape, et surtout si ses voisins sont appariés avec les voisins des groupes connexes. Nous avons distingué quatre cas illustrés dans la deuxième colonne du Tableau 7.

- cas a) : aucun voisin du candidat C_i n'est apparié aux voisins du groupe G_1. Dans ce cas, nous croyons que le candidat n'est pas l'homologue de l'objet *obj1* en cours d'analyse. Nous supposons dans ce cas qu'un sur-appariement s'est produit lors du précédent processus d'appariement. Afin de corriger cette possible erreur, nous attribuons à l'hypothèse $\neg appC_i$ une masse de croyance importante, car nous croyons à travers l'information de voisinage que l'appariement réalisé en première passe est peu probable.
- cas b) : un seul voisin du candidat C_i est apparié à un seul voisin du groupe G_1. Dans ce cas, nous croyons faiblement que le candidat C_i n'est pas l'homologue de l'objet *obj1* en cours d'analyse. Par rapport au cas précédent, dans cette situation, nous diminuons la masse de croyance attribuée à l'hypothèse $\neg appC_i$ et nous attribuons une masse de croyance faible mais non nulle à l'hypothèse $appC_i$,

- cas c) : plusieurs voisins du candidat C_i sont appariés à plusieurs voisins du groupe G_1, mais pas tous les voisins. Dans ce cas, du fait que les niveaux de détail des deux bases de données sont différents, une comparaison fiable du nombre de voisins du groupe G_1 et des voisins du candidat C_i n'est pas faisable. Nous attribuons alors une masse de croyance élevée à l'ignorance et le complément est partagé entre les hypothèses $appC_i$ *et* $\neg appC_i$.
- cas d) : tous les voisins de *obj1* sont appariés avec tous les voisins du groupe G_1. Dans ce cas, nous sommes fermement convaincus que le candidat C_i est l'homologue de l'objet *obj1*. Ainsi, nous attribuons à l'hypothèse $\neg appC_i$ une masse de croyance importante, avec un faible doute.

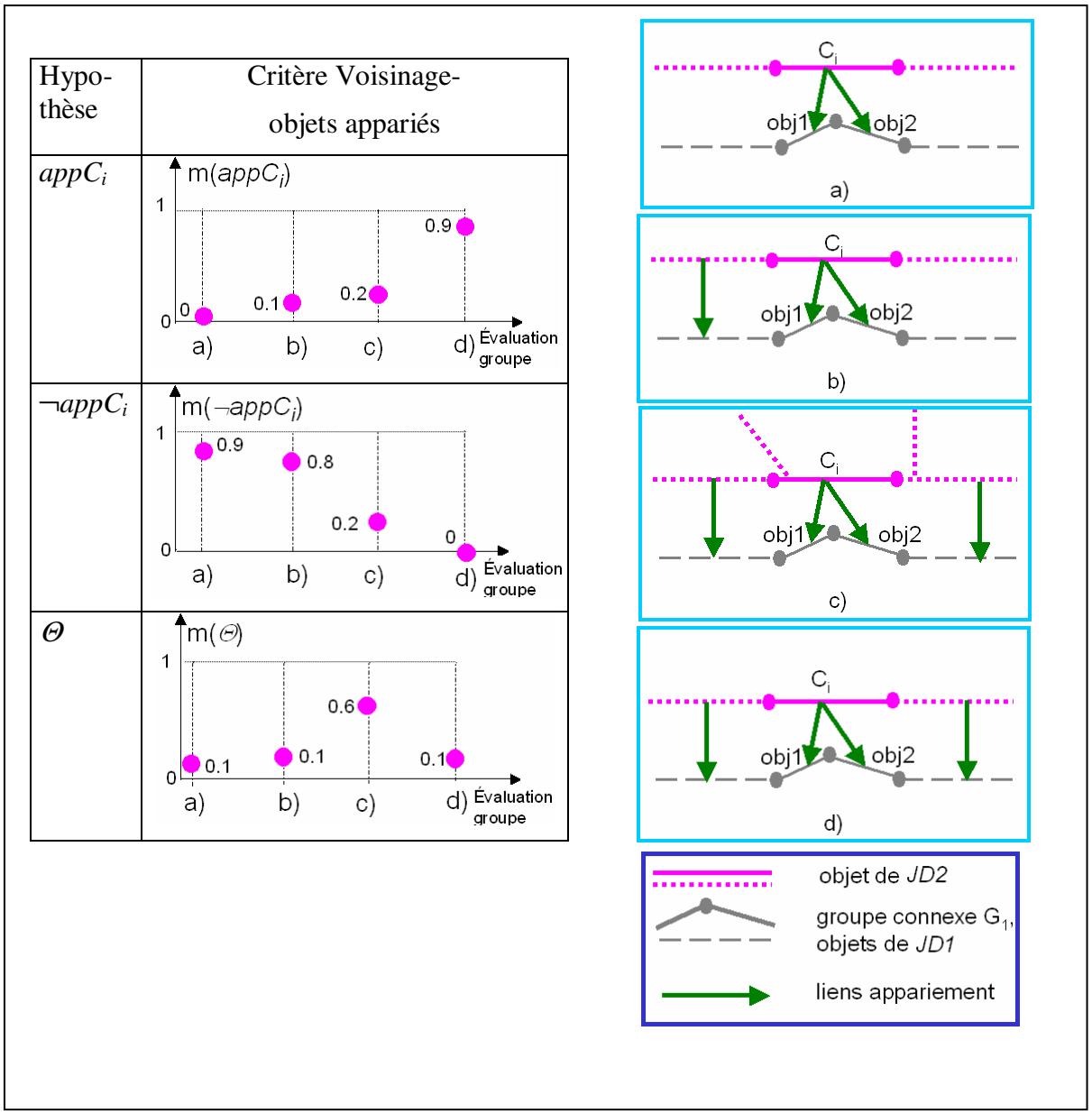


Tableau 7. Représentation des connaissances pour le critère voisinage-objets appariés (à gauche) dans les quatre cas définis à droite.

Critère voisinage, objets non-appariés

Si l'objet *obj1* n'a pas été apparié lors de la première passe du processus nous utilisons le critère voisinage-objets non-appariés. Nous évaluons l'objet *obj1* en analysant ses voisins pour savoir s'ils ont été appariés ou pas. Ainsi, pour l'objet *obj1* nous définissons un attribut « évaluation des voisins » qui renseigne sur le nombre de voisins appariés ou non appariés. Ensuite, en fonction de la valeur de cet attribut nous initialisons les masses de croyance du critère voisinage-objets non appariés (deuxième colonne du Tableau 8). Les voisins d'un objet considérés ici sont seulement ceux qui ont la même direction que l'objet *obj1* en cours d'analyse à une tolérance près. Afin de ne pas engendrer de sur-appariements, nous avons traité seulement le cas où il existe deux voisins dans la même direction.

Ainsi :

- cas a) : les deux voisins n'ont pas été appariés. Cela conforte l'idée que l'objet *obj1* n'est pas apparié. Etant donné que notre processus ne permet pas d'initialiser directement l'hypothèse NA, c'est-à-dire « l'objet *obj1* n'est pas apparié », nous attribuons une masse de croyance élevée à l'hypothèse $\neg appC_i$, dont l'hypothèse NA fait partie.
- cas b) : les voisins ont été appariés à des homologues différents, et parmi eux il y a le candidat C_i que nous sommes en train d'analyser. Cela signifie qu'il existe une forte possibilité que l'objet *obj1* soit aussi apparié avec le candidat C_i. Dans ce cas, nous croyons que l'objet *obj1* peut être apparié soit avec le candidat C_i soit avec son voisin, le candidat C_{i-1}. En conséquence, nous attribuons à l'ignorance une masse de croyance importante et le complément est partagé entre les hypothèses $appC_i$ *et* $\neg appC_i$,
- cas c) : seulement un voisin a été apparié et celui-ci est le candidat C_i. Dans ce cas nous croyons qu'il y a une possibilité que l'objet *obj1* soit apparié aussi à l'homologue de son voisin, c'est-à-dire le candidat C_i. La masse de croyance est partagée entre l'hypothèse $appC_i$ et l'ignorance.
- cas d) : enfin, les deux voisins ont été appariés et avec le même homologue, et celui-ci est le candidat C_i. Ce cas conforte l'idée qu'il existe une forte possibilité que l'objet soit aussi apparié à l'homologue de ses voisins, c'est-à-dire le candidat C_i.

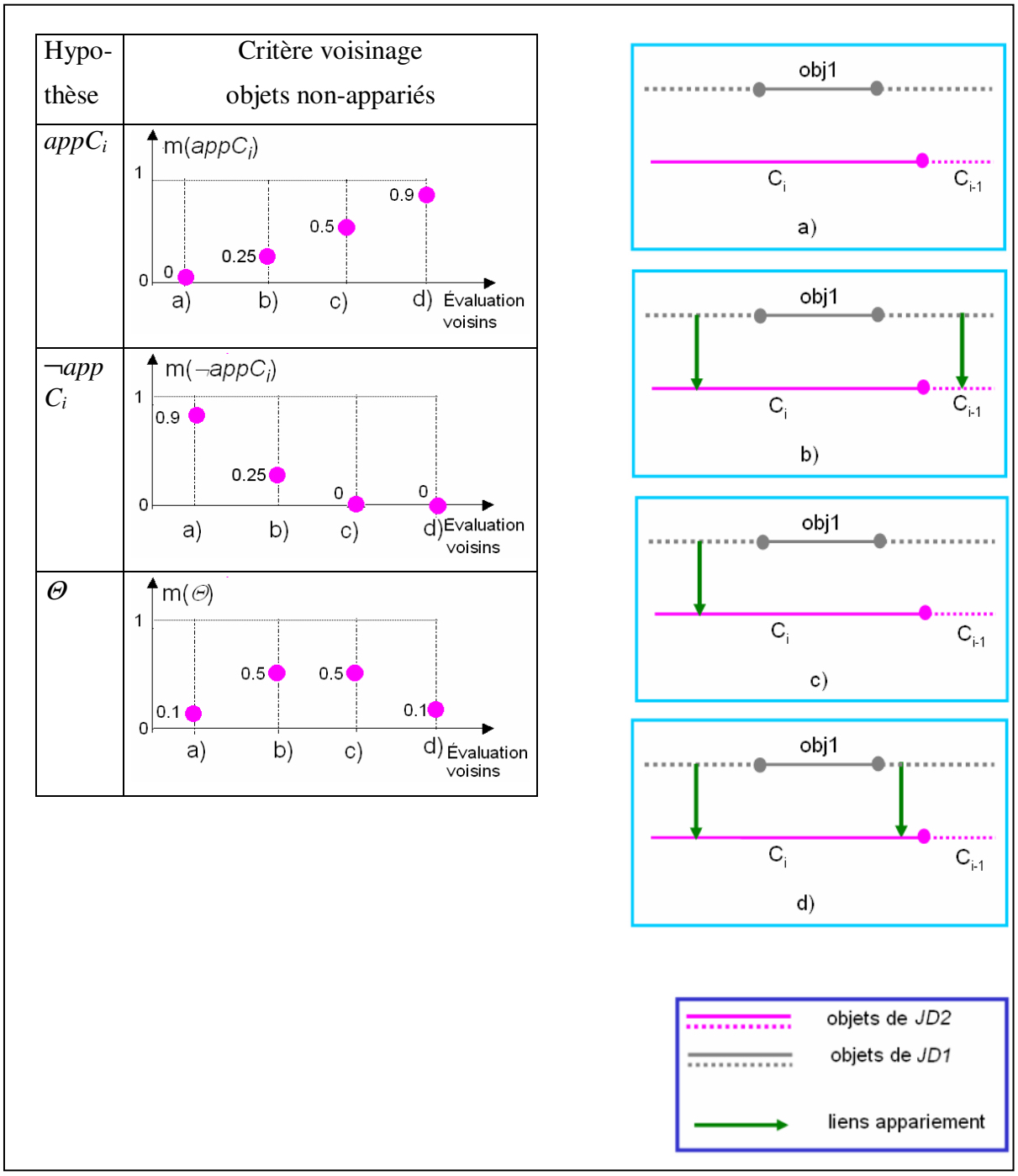


Tableau 8. Représentation des connaissances pour le critère voisinage-objets non-appariés (à gauche) dans les quatre cas définis à droite.

C.3 Conclusion

Dans ce chapitre nous avons présenté notre approche d'appariement de données géographiques basée sur la théorie des fonctions de croyance, qui permet de modéliser d'une manière explicite les imperfections présentes dans les données et dans les connaissances.

Nous avons défini un processus d'appariement de données géographiques composé de cinq étapes principales : la sélection des candidats, l'analyse des candidats et l'initialisation des masses de croyance, c'est-à-dire la représentation des connaissances à travers des critères d'appariement, la fusion des critères d'appariement, la fusion des candidats et la décision.

Dans le but d'avoir une analyse plus globale et ainsi de corriger d'éventuelles erreurs d'appariement, nous pouvons introduire dans notre processus d'appariement des critères d'appariement liés à la notion de voisinage. Le critère de voisinage est instancié à partir des résultats d'une première passe du processus d'appariement, puis le processus est relancé avec ce nouveau critère. Ainsi, nous pouvons imaginer que le processus est répété jusqu'au moment où la convergence du processus est obtenue, c'est-à-dire que les résultats ne changent plus. Ce point n'a pas été étudié en détail. Il s'agit donc d'une piste qu'il faudrait approfondir et expérimenter.

Les critères d'appariement s'appuient sur des connaissances qui peuvent provenir des spécifications du contenu des bases de données géographiques à apparier (comme par exemple les seuils), des données elles-mêmes (par exemple les différentes distances calculées à partir de la géométrie ou les attributs des objets géographiques) ou encore des experts (par exemple les règles).

Nous considérons que notre processus d'appariement est adaptable et évolutif, c'est-à-dire que le processus ne se borne pas à quelques critères seulement mais qu'il est possible d'en rajouter autant que nous en avons besoin et aussi de les adapter en fonction des données utilisées. Dans ce chapitre, nous avons donné à titre d'exemple quelques critères d'appariement typiques qui peuvent être utilisés dans le processus d'appariement ainsi que la représentation explicite des connaissances spécifique à chacun d'entre eux : le critère d'écart de position, le critère d'orientation, le critère sémantique, le critère toponymique, le critère de voisinage topologique.

L'initialisation des jeux de masses, c'est-à-dire la représentation explicite des connaissances, est une étape primordiale. Comme nous avons pu le remarquer à travers des exemples typiques, les courbes représentant les jeux de masses peuvent être différentes d'un critère à l'autre. Ceci s'explique par le fait que la modélisation de chaque critère d'appariement s'appuie sur différents types de connaissance et que le degré de fiabilité est différent. Cette flexibilité, c'est-à-dire la représentation explicite de connaissances différentes, est un avantage clé de notre approche.

Cependant, il s'avère que la représentation explicite des connaissances n'est pas si immédiate et qu'à ce stade elle nécessite une bonne connaissance des spécifications et des données, ainsi que des connaissances d'experts. Afin de rendre le processus d'appariement plus générique et plus accessible aux utilisateurs, nous proposons dans le chapitre E plusieurs solutions d'aide à la modélisation des critères.

CHAPITRE D
Expérimentations et évaluation

D Expérimentations et évaluation

Dans ce chapitre nous présentons les expérimentations que nous avons mises en œuvre afin d'étudier la validité de notre approche d'appariement de données que nous avons décrite au chapitre C. D'abord nous présentons la mise en œuvre du processus d'appariement de données, en décrivant le modèle de données que nous avons défini ainsi que le prototype utilisé, c'est-à-dire la plate-forme GeOxygene et l'interface d'appariement.

Ensuite nous exposons les deux applications que nous avons réalisées : l'appariement des points remarquables du relief et l'appariement des réseaux routiers. Enfin, après l'évaluation des résultats, nous présentons une analyse détaillée des résultats et nous concluons.

D.1 Mise en oeuvre

D.1.1 Prototype: plate-forme GeOxygene

Notre processus d'appariement a été mis en œuvre dans la plate-forme open-source GeOxygene [Badard et Braun, 2004]. Cette dernière est la plate-forme de travail du laboratoire COGIT de l'IGN depuis 2002. L'objectif de cette plate-forme est de fournir un cadre ouvert, modulaire et interopérable pour le développement des applications de recherche du laboratoire COGIT liées à la représentation multiple, à l'analyse de données topographiques dans les domaines des risques naturels et des phénomènes territoriaux et à l'accessibilité des données (par exemple l'aide à la restructuration des données, à la création des légendes, à la création des cartes, etc).

L'interopérabilité de la plate-forme GeOxygene est assurée grâce au fait qu'elle s'appuie sur les standards de l'ISO[4] et de l'OGC[5].

Elle est articulée autour du réseau (Internet et Intranet). Les différents composants de l'architecture sont décrits en Figure 70. Le noyau représente le cœur de la plate-forme GeOxygene et il est présenté plus en détail ci-après. D'autres éléments composent la plate-forme, tels que l'interface de développement (Eclipse) et un atelier de génie logiciel qui permet la réalisation d'un modèle conceptuel de données (comme par exemple le logiciel Objecteering). L'importation des données stockées au format des SIG commerciaux, comme par exemple le format shape, dans le SGBD utilisé par la plate-forme Oracle ou PostGis, est réalisée grâce à un traducteur de données (FME).

[4] ISO signifie International Organization for Standardization
[5] OGC signifie Open Geospatial Consortium

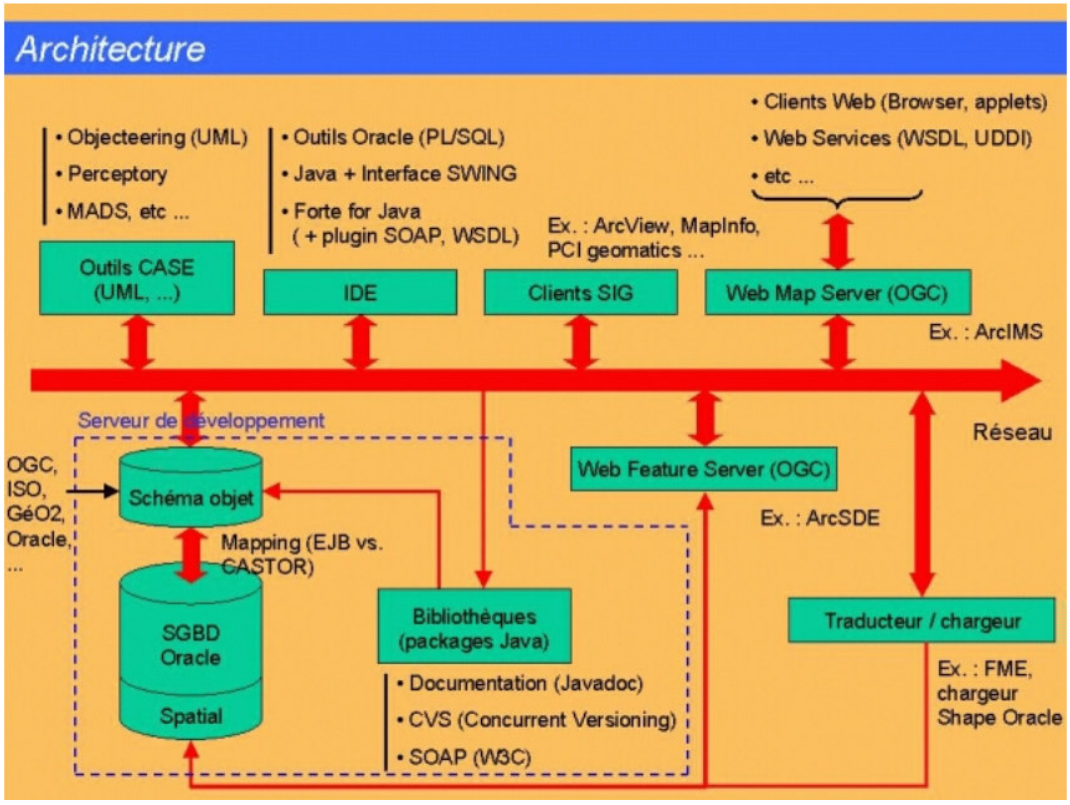


Figure 70. L'architecture de la plate-forme GeOxygene d'après [Badard et Braun, 2003]

Nous présentons brièvement le noyau de la plate-forme GeOxygene puisqu'il est à la base de nos développements. Les principaux éléments du noyau sont les suivants :

- le modèle orienté-objet : ce modèle prend en compte les aspects géométrique, topologique et sémantique des données géographiques. Il s'appuie sur les standards développés par l'ISO et l'OpenGIS (normes 19107, 19109) et est implémenté en JAVA, langage orienté objet.
- le système de gestion de bases de données (SGBD) : il permet le stockage des données. Actuellement les SGBD relationnels utilisés sont Oracle et PostGis.
- le lien objet-relationel (mapping) : ce lien stocké dans un fichier XML, permet de faire la liaison entre le modèle implémenté en Java qui est complètement orienté-objet et les tables du SGBD qui sont des tables relationnelles. Le mapping consiste donc à décrire au moyen d'un module open-source OJB[6] de la fonction Apache[7], quelle classe correspond à quelle table, par exemple la classe Java « TronçonHydrographique » correspond à la table « TRONÇON-HYDROGRAPHIQUE ».
- les bibliothèques d'opérateurs : les opérateurs géographiques, tels que le calcul d'une triangulation ou un appariement de données, sont codés dans des bibliothèques séparées afin d'assurer l'indépendance des développements. Ces opérateurs sont soit issus de bibliothèques Open Source, soit développés au laboratoire en fonction des besoins.

[6] OJB signifie ObJect relational Bridge
[7] Apache est un logiciel de serveurs HTTP produit par l'Apache Software Foundation.

D.1.2 Interface d'appariement de données

De nombreuses applications qui manipulent des données géographiques ont été développées dans la plate-forme GeOxygene : appariement de données [Devogèle, 1997 ; Bel Hadj Ali, 2001 ; Mustière 2006], détection d'incohérence entre les données [Sheeren, 2005], enrichissement des MNT en utilisant des données géographiques [Bonin et Rousseaux, 2005], etc. Afin de visualiser les résultats obtenus de ces applications, la plate-forme GeOxygene a un lien avec le visualisateur JUMP [Ramsey, 2004] sous la forme d'un plug-in.

Nous présentons plus en détail notre environnement de travail, c'est-à-dire l'interface d'appariement de données géographiques développée sur la plate-forme GeOxygene. Cette interface a été mise en place par le laboratoire COGIT lors du projet [BDCARTO-BDTOPO, 2005] et améliorée lors d'un stage au laboratoire COGIT [Teng, 2006]. La visualisation des résultats d'appariement est une vraie problématique en raison de la complexité des données géographiques. Généralement, l'interface est basée sur un affichage multi-fenêtres. Ainsi, en fonction de notre besoin nous pouvons utiliser deux, trois ou quatre fenêtres.

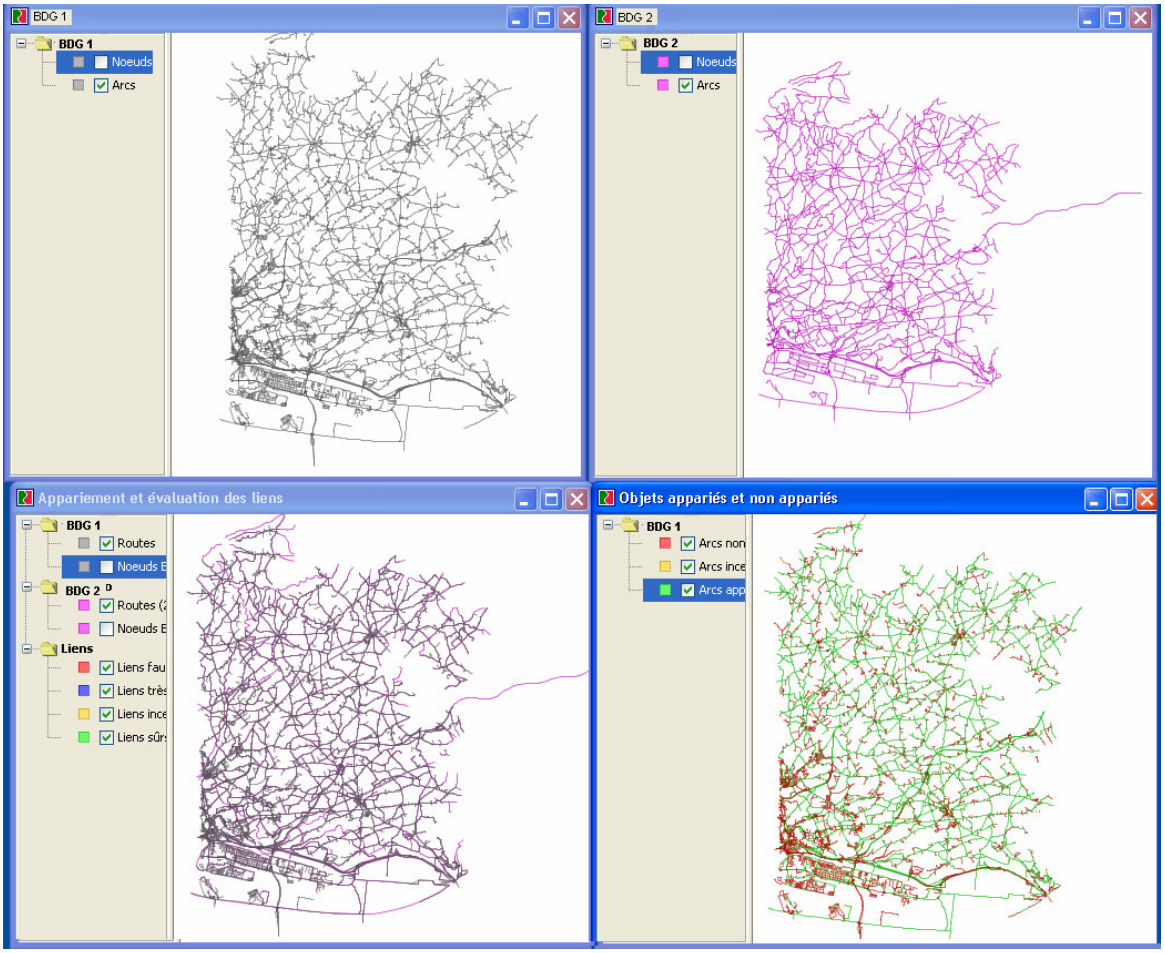


Figure 71. Interface de l'appariement de données géographiques, en mode multi-fenêtres

Supposons que nous avons deux bases de données à apparier, BDG_1 et BDG_2. L'interface de visualisation des résultats d'appariement de données est représentée en Figure 71. Les bases de données sont affichées en haut, BDG_1 à gauche et BDG_2 à droite. En bas à gauche, nous avons les deux bases de données BDG_1 et BDG_2 superposées ainsi que les liens

d'appariement illustrés par des traits entre les objets homologues. Un zoom sur cette fenêtre est montré en Figure 72.

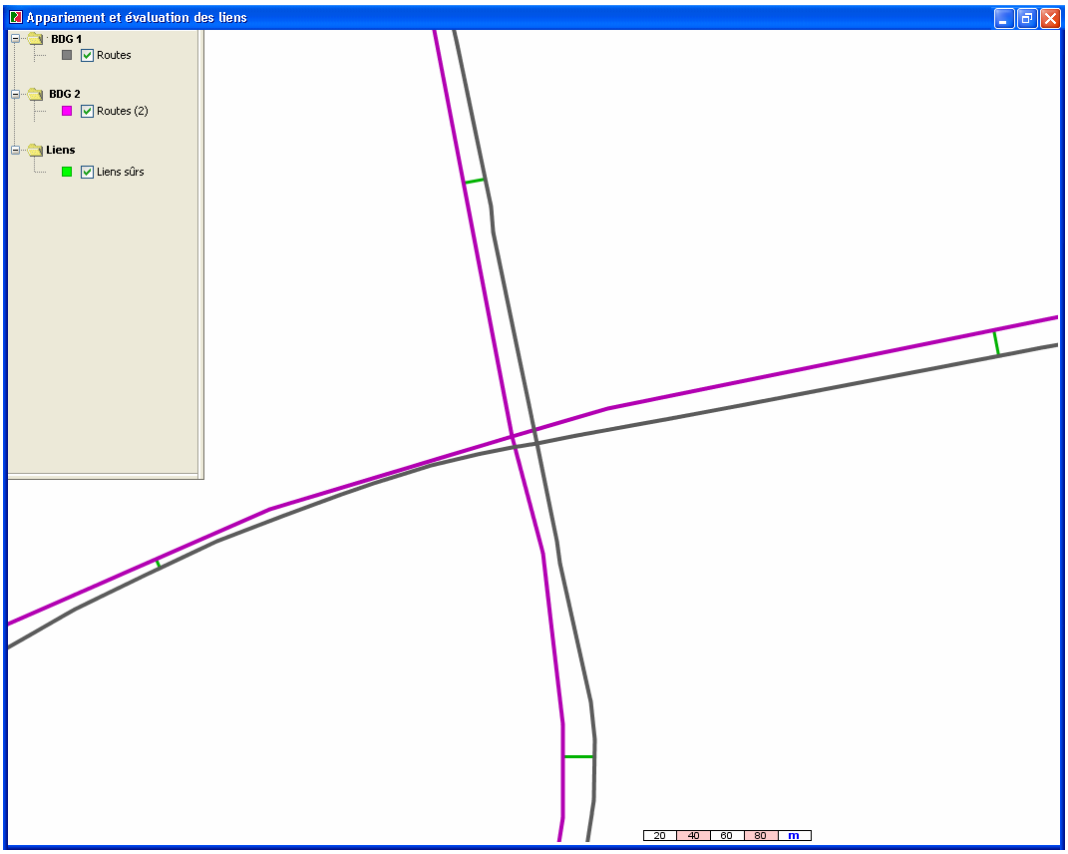


Figure 72. Exemple d'affichage des résultats d'appariement de données entre deux lignes

Le bilan des objets d'une base de données, c'est-à-dire les objets appariés et non-appariés, appartenant soit à BDG_1 soit à BDG_2 (en occurrence il s'agit de BDG_1), peut être visualisé dans la fenêtre en bas à droite. Dans l'exemple ci-dessus, les objets qui ne sont pas appariés sont affichés en rouge, et les objets appariés sont affichés en vert ou en jaune selon l'évaluation du lien d'appariement. Cette évaluation permet de mieux visualiser les éventuelles erreurs d'appariement, c'est-à-dire les objets qui ont peut-être été appariés par erreur [BDCARTO-BDTOPO, 2005].

Cette interface d'appariement permet également une évaluation interactive des résultats d'appariement, c'est-à-dire que nous pouvons analyser les résultats d'appariement, en les parcourant un par un et en zoomant à chaque fois sur le résultat. De plus, lorsque nous nous focalisons sur un couple d'objets appariés dans le but de l'évaluer, nous pouvons faire afficher les attributs des objets appariés et donc les comparer (Figure 73). Ainsi, l'évaluation interactive peut être plus efficace.

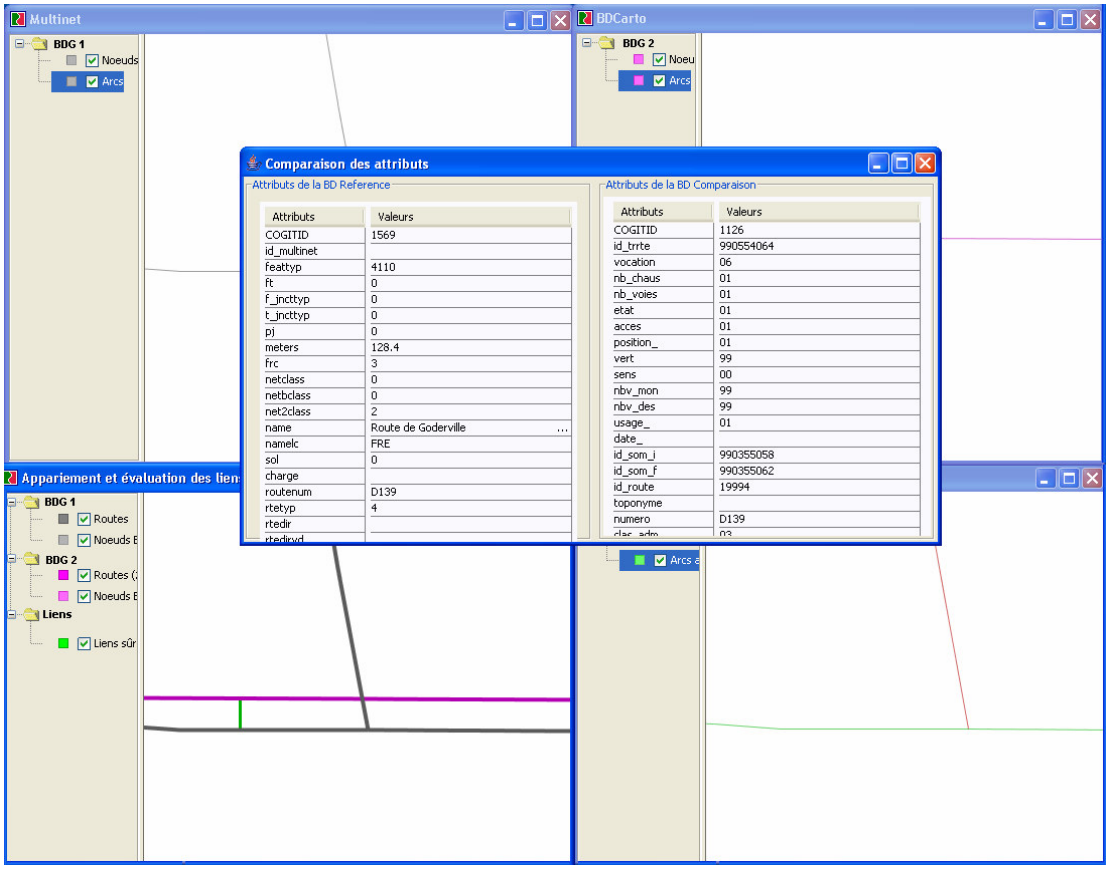


Figure 73. Exemple d'analyse des résultats en utilisant la comparaison des attributs

D.1.3 Schéma conceptuel de données

Le schéma conceptuel de la plate-forme GeOxygene contient plusieurs paquetages : Spatial (les classes liées à la géométrie et à la topologie des objets géographiques), SRC (les classes relatives aux systèmes de coordonnées), Geoschema (contient les différents schémas définis par l'utilisateur), etc.

Nous présentons d'abord les principales classes d'objets qui nous intéressent et qui ont été utilisées pour la mise en œuvre de notre processus d'appariement. Il s'agit essentiellement des classes d'objets contenues dans le paquetage Spatial. Ensuite, le schéma conceptuel de données que nous avons réalisé est illustré et détaillé.

D.1.3.1 Classes de GeOxygene utilisées

La classe mère de tous les objets géographiques dans la plate-forme GeOxygene est la classe *FT_Feature*. Par conséquent, toutes les classes géographiques, telles que « Points remarquables du relief », « Tronçon de route », « Tronçon Hydrographique », etc. en héritent. Cette classe est une classe abstraite, c'est-à-dire qu'elle ne peut pas être instanciée. La classe *FT_Feature* est reliée à la classe *GM-Object* qui représente la géométrie (Figure 74). Ainsi, chaque objet géographique possède une géométrie qui peut être une primitive de base à savoir point, ligne, surface ou une agrégation de primitives de base. La classe *GM_Object* est une classe mère pour plusieurs classes représentant la géométrie d'un objet géographique. La classe *ElementCarteTopo* représente la classe relative à la topologie (nœud, arc, face). Un

élément de la carte topologique a également une géométrie. Cette famille de classes n'est pas issue des normes ISO. Elle a été développée au COGIT pour faciliter la mise en œuvre de processus s'appuyant fortement sur la topologie.

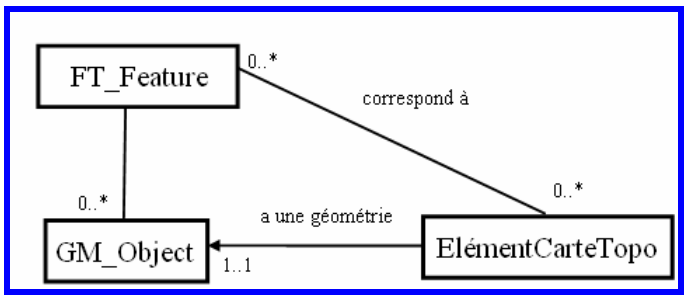


Figure 74. Les classes d'objets relatives à la géométrie des objets géographiques

La partie du schéma de données liée à l'appariement de données est illustrée en Figure 75. La classe *FT_FeatureCollection* est une agrégation de la classe *FT_Feature* qui peut porter un index spatial. La classe *DataSet* est la classe mère pour tout jeu de données. Elle représente un jeu de données, que ce soit une zone d'une base de données ou seulement un thème, et elle est composée d'un ensemble d'objets de la classe *Population*. Le lien d'agrégation récursive permet la définition des relations entre les jeux de données. La classe *Population* contient tous les objets d'une classe héritant de la classe *FT_Feature*. Par exemple, si nous voulons apparier deux jeux de données représentant le réseau routier d'un département, nous avons deux populations « populationJD_1 » et « populationJD_2 ».

La classe *EnsembleDeLiens* est une agrégation de liens d'appariement contenant les résultats d'appariement de deux jeux de données. Un objet *Lien* définit la relation entre les objets homologues. La classe *Lien* hérite également de la classe *FT_Feature*. Cette solution a été choisie afin qu'un objet de la classe *Lien* ait une géométrie qui soit sa représentation graphique.

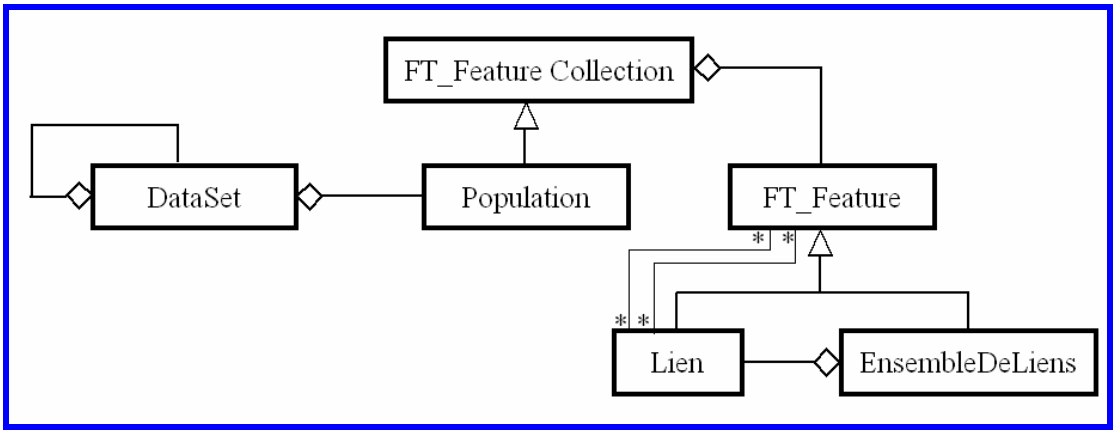


Figure 75. Classes d'objets de base de la plate-forme GeOxygene [Guide utilisateur]

La topologie dans la plate-forme GeOxygene est définie à travers la carte topologique [David, 1988]. Comme l'illustre la Figure 76, la carte topologique est composée de plusieurs classes : *Nœud*, *Arc*, *Face*, *Groupe*. Globalement, la carte topologique permet de modéliser des relations topologiques telles que : un arc a un nœud initial et un nœud final, une face est

composée d'arcs, un arc a une face à gauche et une face à droite, etc. La classe *Groupe* représente une agrégation de nœuds, d'arcs et de faces ; elle est utilisée dans le cadre de l'appariement de données. Par exemple un carrefour représenté par des arcs et des nœuds peut être modélisé par un objet de type *Groupe*. Une carte topologique peut être aussi un réseau, c'est-à-dire la carte topologique sans faces ou même une carte spaghetti, c'est-à-dire sans relations topologiques (dans ce cas le terme de carte topologique est utilisé par abus de langage). Par exemple, pour apparier le thème du bâtiment il n'est pas nécessaire de créer une topologie, alors que pour le thème routier l'instanciation de la topologie peut être utile.

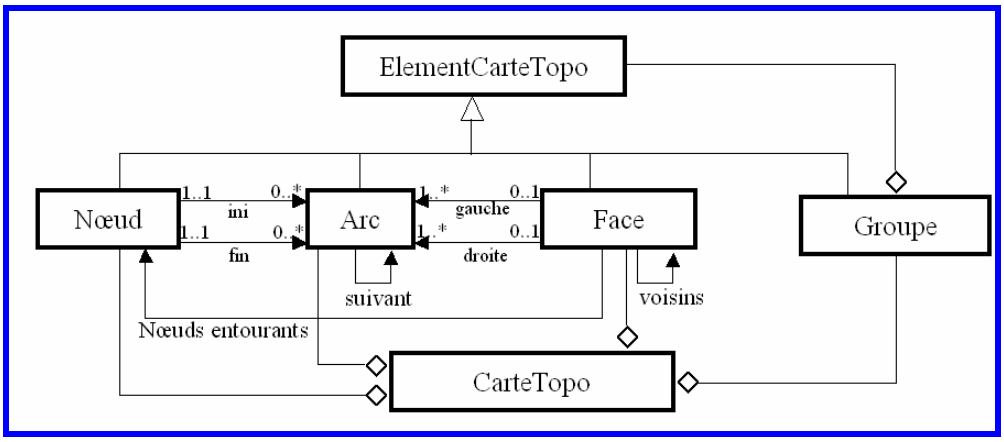


Figure 76. Schéma conceptuel de la Carte Topologique défini dans la plate-forme GeOxygene [David, 1988]

D.1.3.2 Schéma des données proposé pour l'appariement

Notre processus d'appariement doit être capable d'apparier des données géographiques. Afin de l'implémenter pour le valider, nous avons d'abord défini un schéma de données (Figure 80). Les résultats d'appariement sont stockés dans la classe *EnsembleDeLiens*.

Le schéma de données est composé de deux grands groupes de classes : les classes géographiques, qui contiennent des objets ayant une géométrie, et les classes spécifiques à la théorie des fonctions de croyance, qui contiennent des objets sans géométrie. Nous présentons de façon plus approfondie les classes composant ces deux groupes avec leurs attributs.

Les classes géographiques

Nous avons défini quatre classes géographiques, qui héritent de la classe *FT_Feature* (Figure 77) :

- Les classes *Ref* et *Comp* sont les jeux de données à apparier,
- Les classes *ObjetRef* et *ObjetComp* : dans le processus d'appariement nous choisissons le sens de l'appariement, c'est-à-dire que pour les objets d'un jeu de données nous cherchons des objets homologues dans l'autre jeu de données. Pour une meilleure compréhension, nous appelons le jeu de données dont nous parcourons les objets un par un afin de trouver des homologues le jeu de données de « référence », et l'autre jeu de données dans lequel nous cherchons des candidats à l'appariement le jeu de données de « comparaison ».

Signalons que les termes de référence et de comparaison n'ont aucun lien avec la qualité des jeux de données.

Le choix du sens de l'appariement dépend des données à apparier et de la cardinalité des liens souhaités. Un appariement dans les deux sens peut être également mis en place.

Un objet de la classe *ObjetRef* possède deux attributs : commentaire et écartPremierDeuxiemeMax. Le premier renseigne sur le résultat de l'appariement, c'est-à-dire si l'objet a été apparié ou pas et aussi la cause d'un non-appariement. La cause d'un non-appariement peut être qu'il n'a pas de candidats, que les sources sont en conflit ou encore que l'hypothèse NA (« non-apparié ») a été choisie. Le deuxième attribut renseigne sur la fiabilité de l'appariement de cet objet, c'est-à-dire l'auto-évaluation. La valeur de l'attribut correspond à l'écart entre la valeur maximale de la fonction de décision attribuée à l'hypothèse choisie et la deuxième valeur de la fonction de décision attribuée à une autre hypothèse.

Un objet de la classe *ObjetRef* permet de définir le cadre de discernement, c'est-à-dire les candidats à l'appariement. Pour chaque objet de la classe *ObjetRef* il y a un objet *CadreDeDiscernement.*

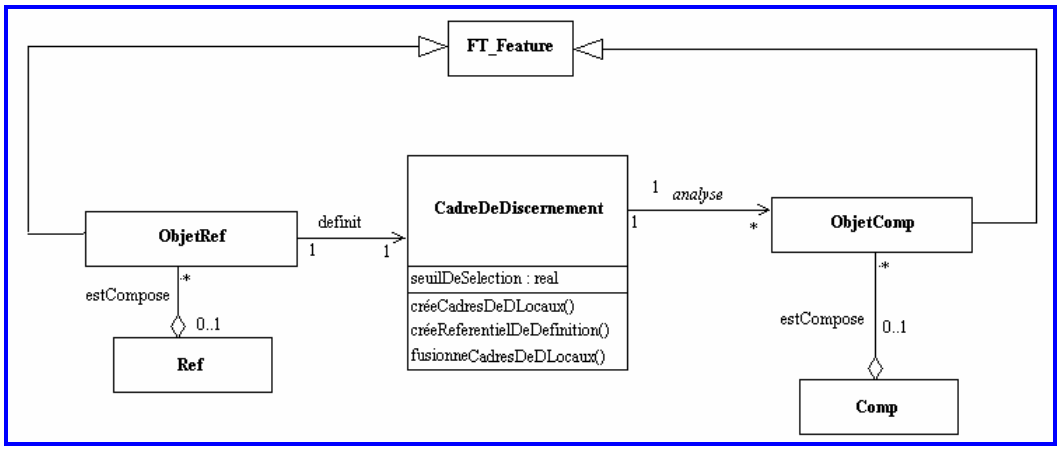


Figure 77. Extrait du modèle conceptuel de données : les classes géométriques

Les classes spécifiques à la théorie des fonctions de croyance

Les classes spécifiques à la théorie des fonctions de croyance sont les classes liées à la définition des candidats (*CadreDeDiscernement*, *RéférentielDeDéfinition)* et des hypothèses (*Hypothèse*, *HypothèseSimple*, *HypothèseRéférentielDeDéfinition*, *CadreDeDiscernement-Local)*, les classes permettant la modélisation des critères d'appariement (*Critère*, *CritèreEcartDistance*, *CritèreSémantique*, etc.) et enfin la classe qui permet la prise de décision (*EnsembleHypothèses)*.

- La classe *CadreDeDiscernement* est associée aux candidats à l'appariement : C_1, C_2,… C_N, c'est-à-dire les instances de la classe *ObjetComp*. Chaque objet de la classe *ObjetRef* définit un seul et unique cadre de discernement. L'attribut *seuilDeSelection* représente le seuil de sélection des candidats, c'est-à-dire la distance maximale jusqu'à laquelle un *objetComp* peut être apparié avec un *objetRef* en cours d'analyse. Cette classe contient

aussi les méthodes qui déterminent les cadres de discernement locaux et le référentiel de définition ainsi que la méthode qui permet la combinaison des candidats,

- *RéférentielDeDéfinition* : elle modélise la combinaison des hypothèses et sert à définir les hypothèses possibles pour un appariement d'un objet. La relation entre les candidats et les hypothèses associées est faite par le lien entre la classe *RéférentielDeDéfinition* et la classe *HypothèseRéférentielDeDéfinition.* Par conséquent, si lors de la décision une hypothèse est choisie, par exemple $appC_2$, nous pouvons récupérer le candidat en correspondance, en occurrence le candidat C_2, pour pouvoir créer ensuite le lien d'appariement,
- *CadreDeDiscernementLocal* : cette classe permet l'analyse individuelle de chaque candidat (voir Figure 78). Un cadre de discernement local contient uniquement les hypothèses définies par l'approche des sources spécialisées, c'est-à-dire $\{appC_i, \neg appC_i$ et $\Theta\}$.
  Sur un objet *CadreDeDiscernementLocal* deux opérations sont possibles : l'initialisation des masses des croyance pour les hypothèses $appC_i$, $\neg appC_i$ et Θ et la fusion des critères. La première opération permet la détermination de toutes les masses de croyance relatives au candidat en cours d'analyse, issues de tous les critères d'appariement qui se sont prononcés individuellement, c'est-à-dire $m_1(appC_i)$, $m_1(\neg appC_i)$, $m_1(\Theta)$, $m_2(appC_i)$, $m_2(\neg appC_i)$et $m_2(\Theta)$, etc. La deuxième opération consiste à fusionner les masses de croyance issues de chaque critère d'appariement.

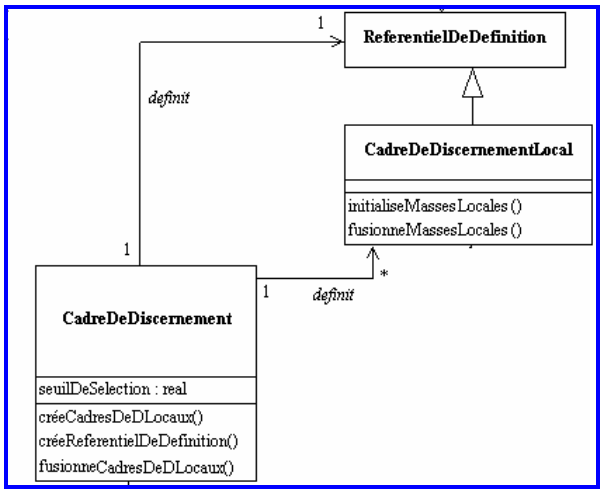


Figure 78. Extrait du modèle conceptuel de données : analyse individuelle de chaque candidat

Exemple d'instanciation

Supposons que nous cherchons à apparier un objet obj1 et qu'après l'étape de sélection, nous avons deux candidats à l'appariement C_1 et C_2. Dans ce cas de figure l'instanciation de notre modèle est la suivante : un objet de la classe ObjetRef définit un objet de type CadreDeDiscernement qui contient trois éléments $\{C_1, C_2, NA\}$ et qui définit un objet de la classe ReferentielDeDefinition. Ce dernier possède les éléments $\{C_1, C_2, NA, (C_1, C_2), (C_1, NA), (C_2, NA)$ et $(C_1, C_2, NA)\}$. Etant donné que nous avons deux candidats, nous avons deux objets instanciés de la classe CadreDeDiscernementLocal. Pour chaque objet de cette

dernière classe nous initialisons d'abord les masses de croyance, puis nous les fusionnons pour chaque critère.

- *Hypothèse* : elle modélise toutes les hypothèses possibles, qui sont soit des hypothèses simples, par exemple $appC_i$, qui signifie que l'objet *objetRef* est apparié avec le candidat C_i, soit des propositions (union d'hypothèses simples), par exemple $\{appC_1, appC_2, NA\}$ qui signifie que l'objet est soit apparié soit avec le candidat C_1, soit apparié avec le candidat C_2, soit pas apparié du tout. Elle est la classe mère des classes *HypothèseSimple* et *HypothèseRéférentielDeDéfinition*,
- *HypothèseSimple* : contient uniquement les hypothèses simples telles que $appC_1$, $appC_2$, $appC_3$, etc.,
- *HypothèseRéférentielDeDéfinition :* contient toutes les propositions possibles, c'est-à-dire $\{\{appC_1\}, \{appC_2\}, \{appC_1, appC_2\}, \{NA\}, \{appC_1, appC_2, NA\}...\{\Theta\}\}$,

  Un extrait des trois classes est présenté en Figure 79.

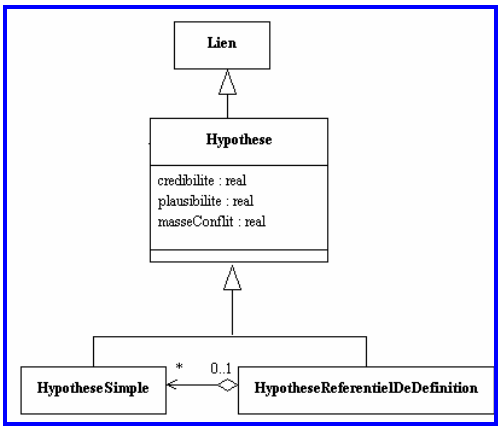


Figure 79. Extrait du modèle conceptuel de données : définition des hypothèses

- *Critère* : cette classe est la classe mère de toutes les classes modélisant un critère d'appariement. Ce schéma de données contient seulement les classes relatives aux critères que nous avons utilisés dans le processus d'appariement de données. Bien évidemment d'autres critères, c'est-à-dire d'autres classes, peuvent être rajoutés. Toutes ces classes contiennent quatre opérations (initialiseMasseAppCi (), initialiseMasseAppPasCi (), initialiseMasseIgnorance ()) qui permettent d'initialiser les masses de croyance pour les hypothèses $appC_i$, $\neg appC_i$ et Θ conformément aux courbes présentées dans le chapitre C. A part les classes *CritèreNomDObjet* et *CritèreTopologique*, chaque classe possède au minimum un attribut qui renseigne sur les seuils illustrés sur les courbes du chapitre C. Notons que ces classes contiennent une seule instance par processus d'appariement,
- *MasseDeCroyance* : elle représente la masse de croyance, associée à une hypothèse, instance de la classe *Hypothèse*,
- *EnsembleHypothèses* : cette classe hérite de la classe *EnsembleDeLiens* et elle contient toutes les hypothèses retenues lors d'un processus d'appariement de données. La décision est prise, en fonction de notre besoin, en s'appuyant sur les mesures de crédibilité, de plausibilité ou de probabilité pignistique.

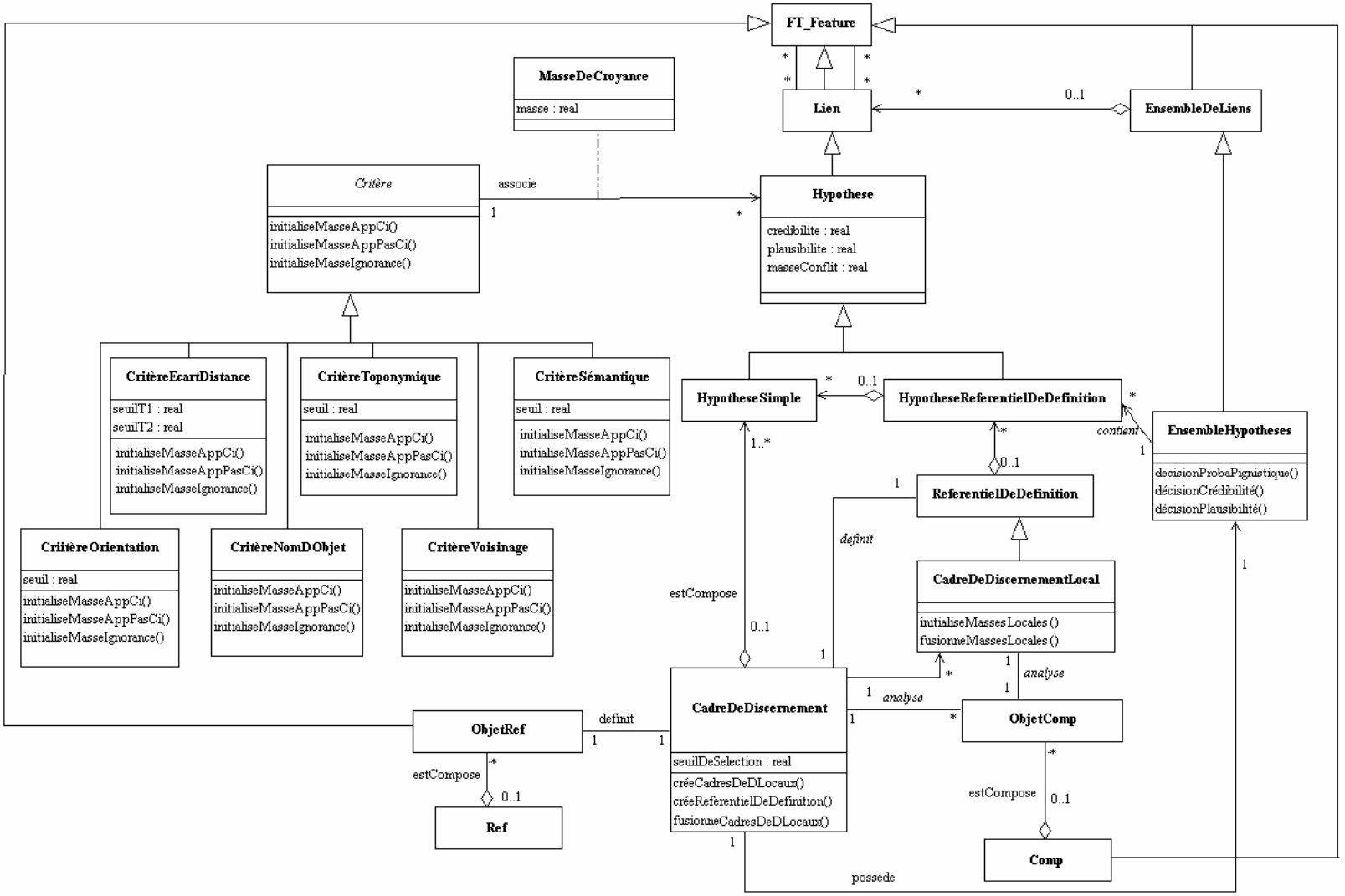


Figure 80. Modèle conceptuel de données du processus d'appariement basé sur la théorie des fonctions de croyance

D.2 Etude des points remarquables du relief

La première étude que nous avons réalisée afin de déterminer la validité de notre approche concerne l'appariement de points remarquables du relief. Après la présentation des jeux de données que nous avons utilisés, nous montrons les résultats que nous avons obtenus. Enfin, une évaluation des résultats et une discussion sont proposées à la fin de cette sous-partie.

D.2.1 Présentation des données

Les données que nous avons utilisées sont issues de deux bases de données géographiques de l'IGN, la BDCARTO® et la BDTOPO®, et concernent les points remarquables du relief. Les deux bases de données géographiques présentent des niveaux de détail différents et elles ont été produites pour répondre à des besoins différents.

La BDCARTO® a été produite pour faire de la cartographie à l'échelle de 1 : 100 000 ou 1 : 250 000 et pour réaliser des analyses à des niveaux régionaux et départementaux. La précision de la base est de un à plusieurs décamètres. Elle a été également créée pour des études de réseaux, elle s'intéresse donc à la manière dont les objets sont connectés. Les données géographiques de cette base de données proviennent des images SPOT pour le thème de l'occupation du sol et des cartes à l'échelle de 1 : 50 000 pour les autres thèmes. La base de données est régulièrement mise à jour à partir de diverses sources.

La BDTOPO®, quant à elle, devenue le référentiel à grande échelle, est une base de données de précision métrique [BDTOPO, 2001]. Elle prend en compte les détails topographiques et donne une position précise des objets. Elle est utilisée pour réaliser des cartes topographiques à l'échelle de 1 : 25 000 et pour faire des études à grande échelle. Les données proviennent principalement d'images stéréoscopiques aériennes et d'enquêtes complémentaires faites sur le terrain (opération de complètement).

La différence de niveau de détail entre les deux bases de données géographiques est illustrée en Figure 81.

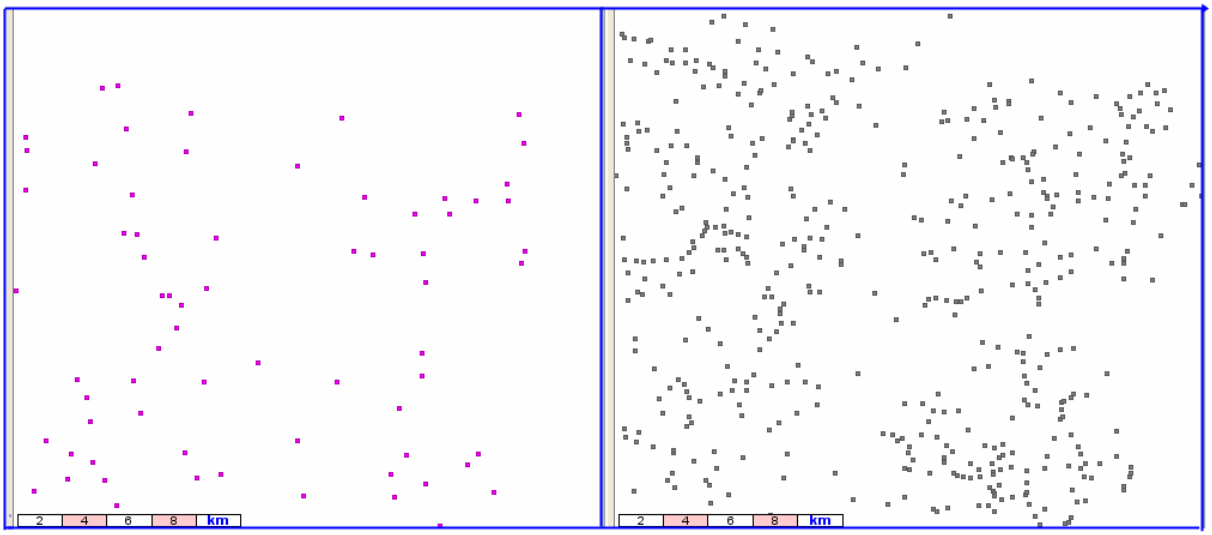


Figure 81. Représentation des points remarquables du relief dans la BDCARTO (à gauche) et la BDTOPO (à droite)

Nous avons utilisé dans nos expérimentations cinq jeux de données représentant cinq départements français : l'Aude (11), les Pyrénées-Atlantiques (64), les Pyrénées-Orientales

(66), la Haute-Garonne (31), le Rhône (69). Précisons que nous avons mis au point notre approche d'appariement de données sur le département 64, et qu'ensuite nous l'avons appliquée sans aucune modification aux autres départements pour l'évaluer. Ces jeux de données ont été choisis en raison du caractère montagneux des départements. Ces départements présentent une hétérogénéité assez représentative du territoire français avec des zones de plaine, de montagne ou de bord de mer. Le département 69 a été choisi afin de tester notre approche sur une zone montagneuse différente des quatre autres départements qui se trouvent dans les Pyrénées.

Dans le Tableau 9 nous présentons pour chaque département la surface et la pente du relief ainsi que le nombre d'objets géographiques présents dans chaque jeu de données, chaque objet géographique représentant un point remarquable du relief. La pente pour chaque département représente la valeur moyenne des pentes moyennes des communes du département. La valeur moyenne de la pente et l'écart type illustrent l'hétérogénéité du relief du département. Ainsi, plus la valeur moyenne est élevée, plus le département est montagneux et plus l'écart type est élevé, plus le relief est hétérogène. A titre d'exemple, la pente moyenne pour un département plat, par exemple la Loire Atlantique est de 26 et pour un département vallonné, par exemple, la Savoie est de 361.

JD1 et JD2 signifient les jeux de données extraits respectivement de la BDCARTO et de la BDTOPO. Un point remarquable du relief, qu'il appartienne à la BDTOPO ou à la BDCARTO, possède deux attributs : la nature de l'objet et le toponyme.

	Dépt 11 Aude		**Dépt 31** Haute-Garonne		**Dépt 64** Pyrénées-Atlantiques		**Dépt 66** Pyrénées-Orientales		**Dépt 69** Rhône	
	JD1	**JD2**	**JD1**	**JD2**	**JD1**	**JD2**	**JD1**	**JD2**	**JD1**	**JD2**
Surface planimétrique (km²)	6 139		6 309		7 645		4 116		3 249	
Pente moyenne (‰) / écart-type	166/ 116		138/ 139		151/ 126		241/ 161		145/ 81	
Nb objets	307	3358	110	504	365	1965	360	2145	90	504

Tableau 9. Nombre d'objets dans les jeux de données utilisés dans les expérimentations, et les caractéristiques des départements couverts par les jeux de données

Nous montrons à titre d'illustration sur la Figure 82 les classes d'objets de la BDCARTO et BDTOPO qui représentent les points remarquables du relief.

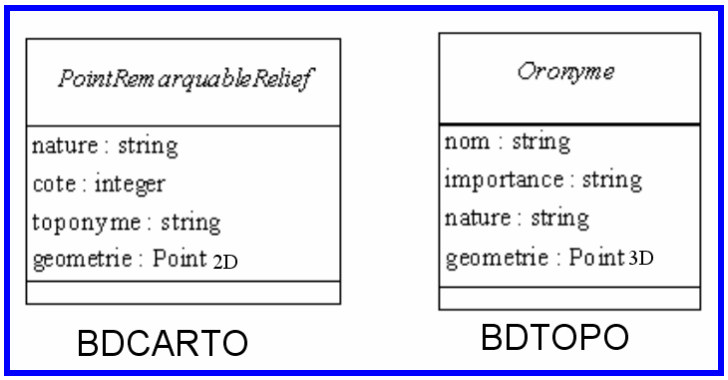


Figure 82. Classes d'objets de la BDCARTO et de la BDTOPO représentant les points remarquables du relief

La localisation des données est imprécise d'une part par définition, la limite entre une vallée et une montagne n'étant par exemple pas parfaitement définie, et d'autre part par le choix de représentation géométrique, une vallée ou une plage sont représentées par un point. De plus, au sein d'un même jeu de données, la précision de la localisation est différente, puisque des concepts tels que le pic ou le sommet, sont représentés de la même manière que des concepts qui ont une plus grande étendue tels que la vallée ou la plaine.

Les points remarquables du relief possèdent un toponyme. Cette information est très importante et peut être considérée comme un identifiant. Cependant, les objets homologues peuvent avoir différents toponymes, en particulier en raison :

1. des écarts issus des différences de prononciation, par exemple « Munhoa » et « Monhoa »,
2. de l'imprécision, quand des entités possédant le même toponyme ont des localisations différentes, par exemple « Vallée du Valentin » dans le département 64,
3. de l'utilisation du nom officiel et du nom d'usage pour la même entité géographique,
4. enfin il est possible que la même entité du monde réel possède un toponyme en français dans un jeu de données et un toponyme saisi dans la langue locale dans l'autre base. Ce cas de figure se rencontre surtout dans les régions proches d'une frontière avec un autre pays ou dans les régions où plusieurs langues existent.

La mise en correspondance de l'attribut nature qui ont le même intitulé (« cirque » dans la BDTOPO, « cirque » dans la BDCARTO) nous semble immédiate. Cependant, dans la réalité tous les objets qui ont la même nature ne sont pas homologues, et parfois des objets qui ont des natures légèrement différentes peuvent être homologues. Ceci est dû au niveau de détail de la classification dans les deux bases de données. Par exemple dans la BDCARTO il y a des concepts qui sont regroupés et représentés avec la même valeur de l'attribut nature (par exemple « sommet, crête, colline »), alors que dans la BDTOPO ils sont bien séparés, c'est-à-dire « sommet », « crête », « colline ». De plus, la valeur de l'attribut nature peut designer un seul concept dans la base, et un regroupement de concepts dans l'autre base. Par exemple un objet géographique de nature « rocher » désigne un chaos, un éboulis, un pierrier ou bien un rocher.

Même si les deux bases de données ont été produites par le même producteur, il s'avère que la classification n'est pas toujours la même dans les deux bases de données. Le concept de

« cluse » (« une gorge transversale dans un pli anticlinal, c'est-à-dire un pic dont la convexité est tournée vers le haut » d'après Larousse) se trouve dans la BDTOPO rattaché au concept de « gorge » et dans la BDCARTO au concept de « col, passage ». Cette information est présente uniquement dans les spécifications des bases de données. Par conséquent, il s'avère que pour des cas difficiles, nous devons gérer l'incertitude, c'est-à-dire autoriser d'apparier des objets de natures différentes mais proches.

Il existe aussi souvent des confusions relatives à la nature des objets, lorsqu'il s'agit de concepts qui sont très proches d'un point de vue sémantique. Les concepts de pic et de sommet sont différents : un pic représente « un sommet pointu d'une montagne » [BDTopoPays, 2002] tandis qu'un sommet représente « un point haut du relief caractérisé par un profil abrupt ». Pourtant, il arrive que la même entité du monde réel soit de nature « pic » dans un jeu de données et de nature « sommet » dans l'autre jeu de données. Cette confusion est liée sûrement aux points de vue différents des opérateurs qui ont réalisé la saisie de la nature des objets, mais aussi à la définition des deux concepts à travers les adjectifs « pointu » et « abrupt » pouvant faire référence à la même réalité.

Utiliser seulement un critère basé sur une mesure de proximité 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, puisque par exemple un col et un sommet peuvent avoir les mêmes toponymes.

D.2.2 Tests

Afin de tester notre approche d'appariement de données pour les points remarquables du relief, nous avons utilisé le critère d'écart de position, le critère toponymique et le critère sémantique. Nous n'avons pas utilisé l'altitude, puisque cette information n'est pas présente dans la BDCARTO. Une possibilité est d'utiliser la base de données altimétrique (BDALTI) de l'IGN et donc de définir un critère basé sur cette information.

L'appariement des points remarquables du relief consiste, pour chaque objet du jeu de données moins détaillé (BDCARTO), à chercher d'abord des candidats potentiels à l'appariement dans le jeu de données plus détaillé (BDTOPO). Etant donnée la spécificité des données, nous cherchons des liens d'appariement de cardinalité 1 : 1, c'est-à-dire un objet géographique de la BDCARTO correspond à un objet géographique de la BDTOPO.

Les candidats à l'appariement ont été sélectionnés en utilisant un seuil qui dépend en général de la précision des jeux de données à traiter, de la connaissance de l'erreur moyenne sur la position des éléments présents dans ces bases et de la nature des objets. Un seuil de sélection élevé permet de prendre en compte l'imprécision dans les données et de sélectionner tous les homologues potentiels. Ainsi, nous sommes assurés que tous les homologues potentiels ont été sélectionnés, et que le cadre de discernement est exhaustif. La raison pour laquelle nous avons choisi de fixer le seuil de sélection en fonction de la nature des objets est la réduction de la complexité algorithmique. En effet, si nous avons de nombreux candidats, la complexité algorithmique augmente, puisque d'une part le nombre d'hypothèses possibles augmente exponentiellement avec le nombre de candidats, et d'autre part nous analysons chaque candidat indépendamment des autres et ensuite nous fusionnons les candidats.

Pour les objets géographiques représentant une entité du monde réel ponctuelle, tels que les sommets, les cols, les pics, nous avons choisi un seuil de 1 km, tandis que pour les objets géographiques représentant une grande étendue par exemple une vallée ou une plage, nous avons choisi un seuil de 10 km.

D.2.2.1 Paramétrage des courbes pour l'initialisation des masses de croyance

Comme nous l'avons vu au chapitre C, l'initialisation des masses de croyance nécessite de calculer des distances et de déterminer les fonctions de croyance. Les fonctions de croyance ayant été définies au chapitre C, il ne reste plus qu'à définir les seuils. Nous discutons dans cette sous-partie du paramétrage des courbes spécifiques à chaque critère utilisé, c'est-à-dire des distances et des seuils que nous avons déterminés afin de réaliser nos expérimentations.

Critère d'écart de position

Pour mesurer l'écart de position entre deux objets géographiques représentant deux points remarquables du relief appartenant à deux jeux de données différents, nous avons utilisé la distance euclidienne planimétrique. Les courbes utilisées sont celles qui ont été illustrées au chapitre C (paragraphe C.2.1.1). Nous les rappelons ici dans le Tableau 10.

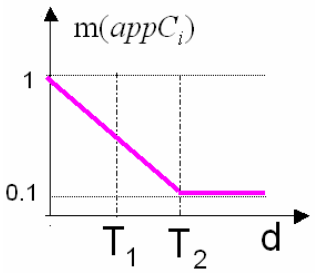
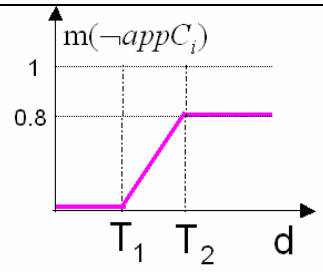
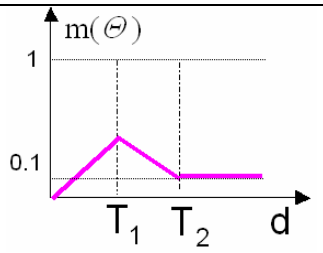
Hypothèse	Critère d'écart de position
$appC_i$	
$\neg appC_i$	
Θ	

Tableau 10. Représentation des connaissances pour le critère d'écart de position

En ce qui concerne les seuils utilisés dans les courbes qui modélisent le critère d'écart de position, T_1 représente la précision des données et a été fixé expérimentalement à 400 m, et T_2 est égal au double de T_1, c'est-à-dire 800 m. Ainsi, si deux objets se trouvent à une distance supérieure au double de la précision des données, nous croyons qu'ils ne sont pas homologues.

Critère sémantique

Pour étudier la pertinence de l'évaluation de la ressemblance sémantique à partir d'une ontologie, nous avons testé deux méthodes. La première consiste à demander à des experts d'attribuer des notes entre les valeurs de l'attribut nature qui sont des concepts ou des groupes de concepts de nos bases de données [Abadie *et al.*, 2007]. La deuxième méthode consiste à calculer des distances sémantiques entre les concepts ou les groupes de concepts présents dans les deux jeux de données à partir d'une taxonomie de domaine [Abadie et Mustière, 2008].

Nous avons mis en place un test qui consiste à demander à des experts de fournir des notes pour évaluer la ressemblance entre les concepts. Les notes sont comprises dans l'intervalle [0, 1], où 0 signifie une ressemblance sémantique totale et 1 signifie qu'il n'y a aucune ressemblance. Le test réalisé ainsi que les résultats obtenus sont détaillés dans l'Annexe 2. Nous les résumons néanmoins ci-dessous.

Ce test a montré que les experts fournissent une mesure intuitive en comparant les propriétés des concepts telles que la forme ou la localisation, en plus des relations de spécialisation. Prenons par exemple les concepts « dune » et « colline ». Les deux concepts ont physiquement la même forme mais n'ont pas les mêmes propriétés, c'est-à-dire que la dune est un monticule de sable alors que la colline est une petite élévation de terre. Les experts les évaluent comme proches d'un point de vue sémantique en raison de leurs aspects semblables. Les résultats montrent également l'importance du point de vue de chacun des experts et les ambiguïtés que celui-ci peut engendrer. Par exemple, le concept de cirque est vu par quelques experts comme une montagne et par d'autres comme une dépression. Cependant, ces ambiguïtés nous poussent à des réflexions plus amples et sont exploitables par exemple pour la validation d'une ontologie de domaine.

La plus grande difficulté pour les experts concerne le regroupement des concepts, tels que « dune, plage », « sommet, crête, colline » ou « cuvette, dépression ». Ainsi, autant la paire « plaine, plateau »-« plaine-plateau » est facile à évaluer parce que les deux paires contiennent les mêmes concepts, autant la paire « dune, plage »-« dune, isthme » est beaucoup plus complexe. La même difficulté est rencontrée lorsqu'ils doivent évaluer un regroupement de concepts avec chaque concept composant le groupe, comme par exemple « sommet, crête, colline »-« sommet », « sommet, crête, colline »-« colline », etc. Face à ces problèmes et n'ayant pas de contrainte de notation, les experts ont noté de façon hétérogène.

Globalement, les notes fournies par les experts ont l'avantage de leur finesse et l'inconvénient qu'elles ne sont pas homogènes, elles varient d'un expert à l'autre, sauf pour les cas évidents.

La deuxième méthode d'évaluation de la ressemblance sémantique testée entre les concepts s'appuie sur une taxonomie de domaine obtenue au laboratoire COGIT par extraction automatique des spécifications des bases de données géographiques [Abadie et Mustière, 2008]. Ainsi, afin de comparer les différentes natures, nous avons utilisé une distance sémantique d_S [Wu et Palmer, 1994], présentée dans la partie A.4.2.3.

Afin de gérer la difficulté de comparaison causée par le regroupement de concepts, nous avons choisi de comparer les concepts composant le groupe et de garder la valeur maximale obtenue. Par exemple, pour deux groupes « dune, plage » et « dune, isthme » nous calculons les différentes valeurs de similarité « dune »-« dune », « dune »-« plage », « isthme »-« dune », et « isthme »-« plage », et nous ne conservons finalement que la valeur de similarité « dune »-« dune ». Nous remarquons en Figure 83 et en Figure 84 que le concept « isthme » n'apparaît pas dans notre taxonomie. Par conséquent, ce choix a été retenu dans le but de

favoriser les appariements entre objets géographiques dont les valeurs des attributs *nature* possèdent au moins un terme en commun.

Après avoir comparé théoriquement les distances obtenues par les deux méthodes, nous avons remarqué que la taxonomie est utilisable parce que globalement elle est cohérente avec les experts. Cette cohérence montre l'intérêt de la taxonomie puisqu'elle est plus facile à mettre en œuvre. La taxonomie présente donc un grand avantage, car il est difficile de trouver un ensemble représentatif d'experts disponibles pour évaluer la sémantique des concepts pour chaque thème de la base de données. De plus, cette cohérence a été validée au moyen des tests d'appariement que nous avons réalisés sur deux départements : 64 et 69. Plus précisément, nous avons apparié les jeux de données en utilisant les mesures données par les experts et les mesures calculées automatiquement à partir de la taxonomie de domaine. Les résultats obtenus sont identiques. Ceci montre aussi que, comme nous l'espérions, grâce à la combinaison de critères, notre approche est peu sensible aux variations des mesures qui ne sont pas trop importantes. Précisons que la taxonomie est valable surtout parce qu'elle a été définie à partir des spécifications des bases de données que nous souhaitons apparier.

Par conséquent, nous avons décidé d'utiliser la taxonomie de domaine pour évaluer la sémantique des concepts. Ce choix a été favorisé en raison de l'automatisation de l'étape de calcul des distances sémantiques, par le fait que nous disposons déjà d'une telle taxonomie et que les tests comparatifs que nous avons faits montrent une cohérence entre la taxonomie et les tests auprès des experts.

La taxonomie de domaine utilisée pour les points remarquables du relief est composée de deux grandes classes : entité du relief terrestre (voir la Figure 83) et entité maritime (voir la Figure 84).

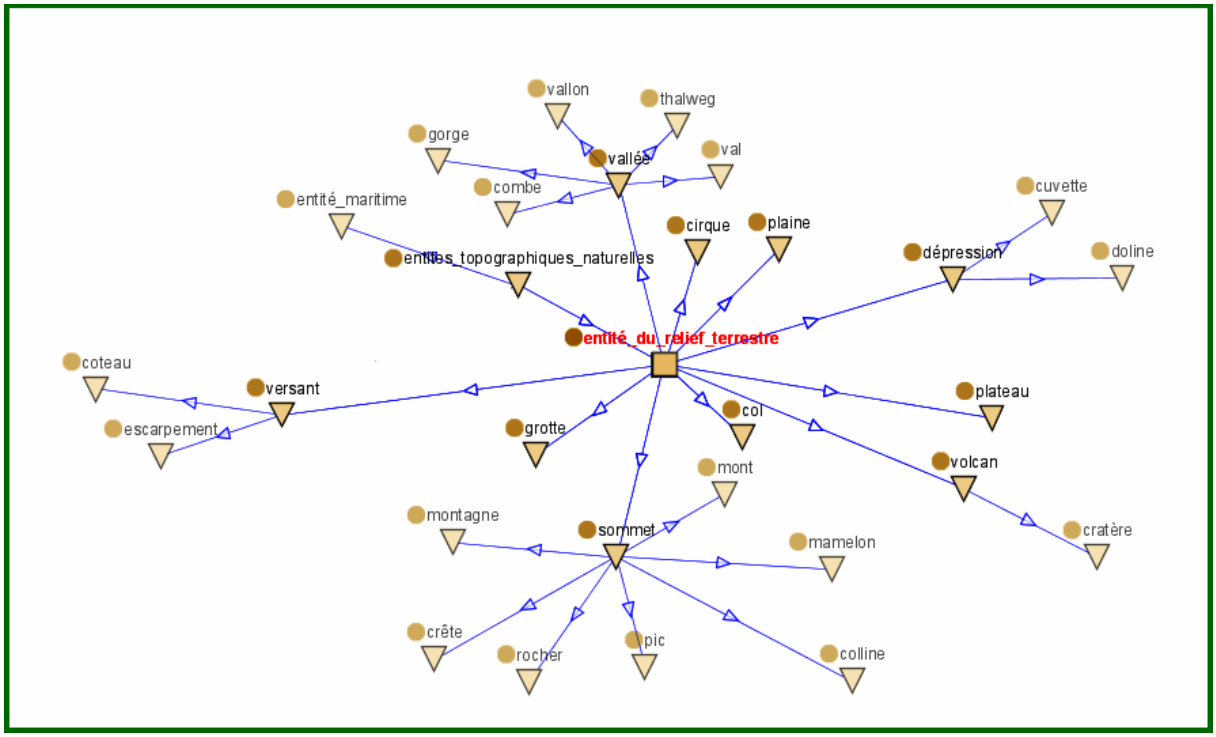


Figure 83. Taxonomie pour la classe entité du relief terrestre d'après [Abadie et Mustière, 2008]

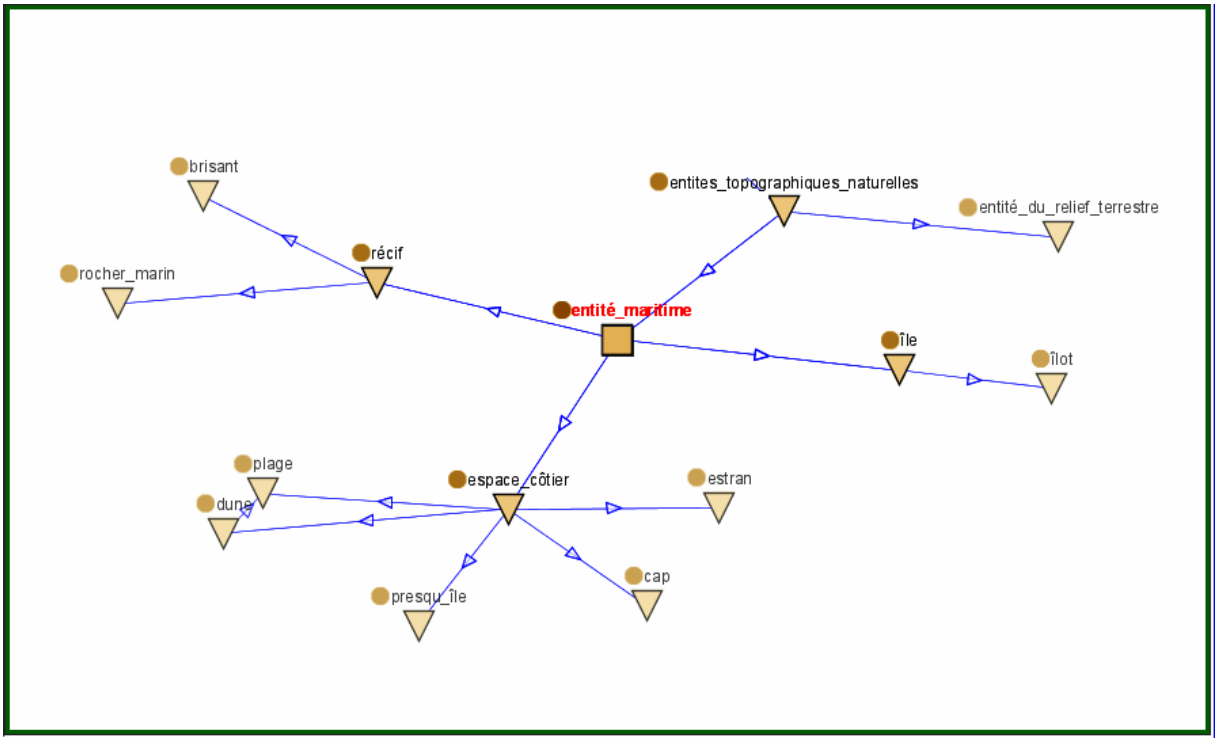


Figure 84. Taxonomie pour la classe entité maritime d'après [Abadie et Mustière, 2008]

Les courbes utilisées sont celles qui ont été définies au chapitre C. Nous les rappelons ici dans le Tableau 11.

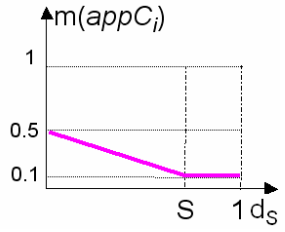
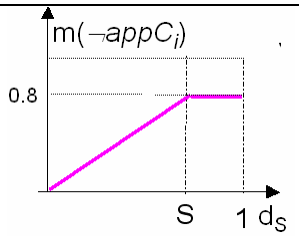
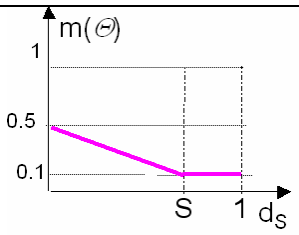
Hypothèse	Critère Sémantique
$appC_i$	
$\neg appC_i$	
Θ	

Tableau 11. Représentation des connaissances pour le critère sémantique

Le seuil S défini dans les courbes qui modélisent le critère sémantique a été fixé à 0,6, c'est-à-dire qu'à partir d'une distance sémantique supérieure à 0,6, deux objets géographiques ont très peu de chances d'être homologues. A titre d'exemple, pour appréhender ce seuil, notons que la distance sémantique d_S (sommet, montagne) est égal à 0,4, d_S (« pic », « sommet ») est égal à 0,57, d_S (« pic »-« escarpement ») est égal à 0,8.

Critère toponymique

Le toponyme d'un point remarquable du relief signifie le nom du lieu. Le critère toponymique a un poids élevé lorsque deux objets géographiques possèdent le même toponyme.

Afin de comparer deux toponymes, nous avons choisi la mesure de similarité définie par [Samal *et al.*, 2004] dans la partie A.4.2.2. Nous rappelons la mesure ici :

$$coefficient_{Similarité} = \frac{\sum_{i=1}^{nombre\ de\ mots_1} valeurMax_i}{(nombre\ de\ mots_1 + nombre\ de\ mots_2)/2}$$

Nous avons retenu cette distance d'une part en raison de sa simplicité et de sa rapidité de calcul, et d'autre part en raison des caractéristiques des données utilisées. Le cas où un toponyme du jeu de données JD_1 est composé d'un toponyme du jeu de données JD_2 correspondant suivi d'un autre toponyme, est fréquent dans les données représentant les points remarquable du relief, par exemple « col de peyrelue ou port vieux de sallent » dans JD_1 et « col de peyrelue » dans JD_2.

Nous donnons à titre d'illustration un exemple de calcul de distance entre deux toponymes.

	col	peyrelue	port	vieux	sallent
col	**1**	0,125	0,625	0	0,25
peyrelue	0,125	**1**	0,25	0,25	0,125

Tableau 12. Matrice « mot-mot » pour deux toponymes et les mesures de similarité entre les mots

Les valeurs du Tableau 12 illustre la similarité entre les mots composant les deux toponymes, après avoir éliminé les mots de liaison. Cette dernière s'appuie sur la distance de Levenshtein et est calculée de la manière suivante :

$$similarité(mot_1, mot_2) = 1 - \frac{d_{Levenshtein}}{max(longueur\ du\ mot1, longueur\ du\ mot2)}$$

A partir de ces valeurs de similarité, le coefficient de similarité proposé par [Samal *et al.*, 2004] est égal à :

$$coefficient_{Similarité} = \frac{1+1}{(2+5)/2} = 0{,}57$$

Rappelons que notre approche est basée sur des mesures de distance. Par conséquent, la distance toponymique est le complément du coefficient de similarité, c'est-à-dire 0,43.

La distance toponymique que nous avons choisie permet de déterminer si l'écart est dû à une variabilité linguistique, à une erreur de frappe, ou si les toponymes sont semblables. Par contre, elle ne suffit pas à lever l'ambiguïté due à l'imprécision ou celle due à l'utilisation de noms différents comme le nom officiel et le nom d'usage. Afin de remédier à ce manque, nous proposons une modélisation plus souple pour le critère toponymique, afin de ne pas rejeter un candidat s'il possède un toponyme complètement différent de l'objet en cours d'analyse. Cette modélisation consiste à donner moins de poids à l'hypothèse $\neg appC_i$ et plus de poids à l'ignorance.

Les courbes utilisées sont celles qui ont été définies dans le chapitre C, section C.2.3.2. Nous les rappelons ici dans le Tableau 13.

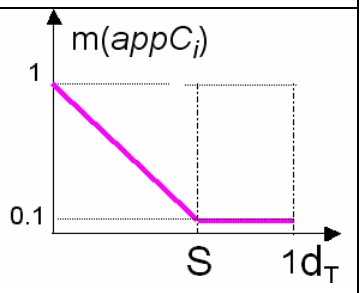
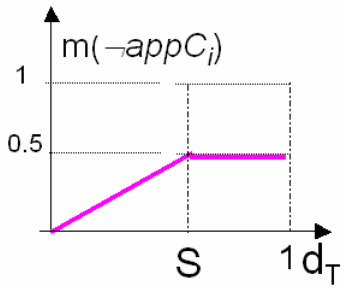
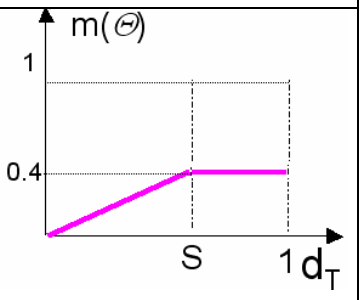
Hypothèse	Critère Toponymique
$appC_i$	
$\neg appC_i$	
Θ	

Tableau 13. Représentation des connaissances pour le critère toponymique

Le seuil S défini dans les courbes qui modélisent le critère toponymique a été fixé à 0,7, c'est-à-dire qu'environ 30% des lettres ne se ressemblent pas.

D.2.2.2 Résultats des expérimentations: exemples

Nous montrons dans cette sous-partie quelques résultats d'appariement que nous avons obtenus. Nous avons également testé d'autres approches d'appariement de données ponctuelles. Il s'agit de l'approche proposée par [Beeri *et al.*, 2004] et de celle qui a été implémentée par le laboratoire COGIT, lors d'une étude interne à l'IGN [BDCARTO-BDTOPO, 2005]. Nous avons expliqué plus en détail dans le chapitre A l'approche proposée

par [Beeri *et al.*, 2004] qui, rappelons-le, s'appuie uniquement sur la géométrie des objets et est basée sur des mesures de probabilités calculées à partir de la distance euclidienne.

En ce qui concerne l'approche de laboratoire COGIT, elle est basée sur un enchaînement des deux critères basés sur la géométrie et la toponymie. Ainsi, deux objets sont appariés si la distance de Levenshtein est au maximum 1 ou si un toponyme d'un jeu de données commence par celui de l'autre jeu de données. Dans le cas où, après la comparaison des toponymes, plusieurs candidats existent, le choix de l'objet à apparier est réalisé en fonction de la distance euclidienne, i.e. n'est conservé que le candidat le plus proche.

Pour mieux comprendre les illustrations des résultats d'appariement obtenus, précisons que dans le cas où les résultats que nous illustrons ne sont pas identiques aux résultats obtenus avec au moins une des autres méthodes, nous le signalons et nous l'illustrons par une image qui montre les résultats obtenus par cette autre méthode. Lorsque le résultat est similaire, aucun commentaire n'est fait.

Sur les figures suivantes montrant les résultats, les objets de la BDCARTO sont représentés par des cercles pleins roses et les objets de la BDTOPO sont représentés par des croix grises. Les courbes bordeaux représentent les courbes de niveau.

Un exemple typique d'appariement sans ambiguïté de points remarquables du relief est illustré en Figure 85. L'objet ayant le toponyme « utzigagna » de la BDCARTO est bien apparié avec son homologue dans la BDTOPO. Dans ce cas, les objets homologues sont proches l'un de l'autre, ont le même toponyme et la même nature (sommet).

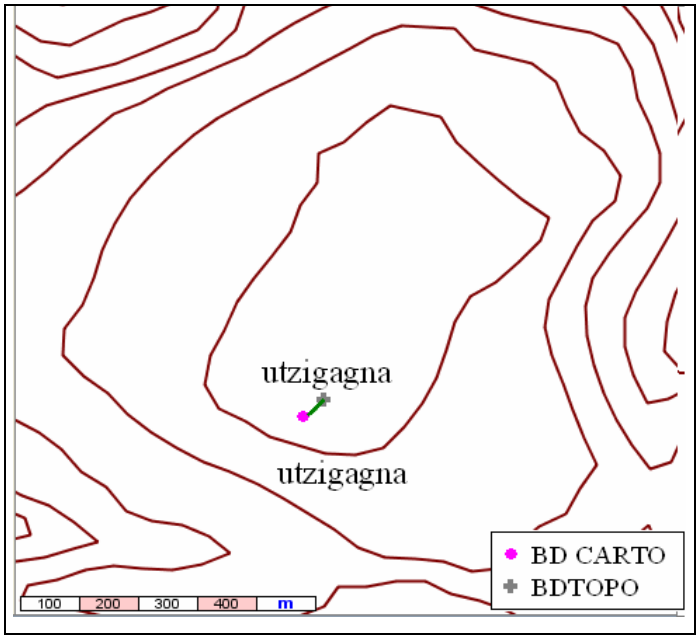


Figure 85. Exemple typique de résultat d'appariement des points remarquables du relief

Le résultat d'appariement illustré en Figure 86 à gauche montre que notre processus d'appariement n'apparie pas nécessairement à l'objet le plus proche. L'objet de la BDCARTO « pic de cortaplana » est bien apparié avec son homologue dans la BDTOPO appelé « pic de cortaplana » et non pas avec l'objet le plus proche. Si nous utilisons l'algorithme de [Beeri *et al.*, 2004] (voir la Figure 86 à droite), l'objet « pic de cortaplana » est apparié à l'objet le plus proche de la BDTOPO, « porte sainte engrâce ».

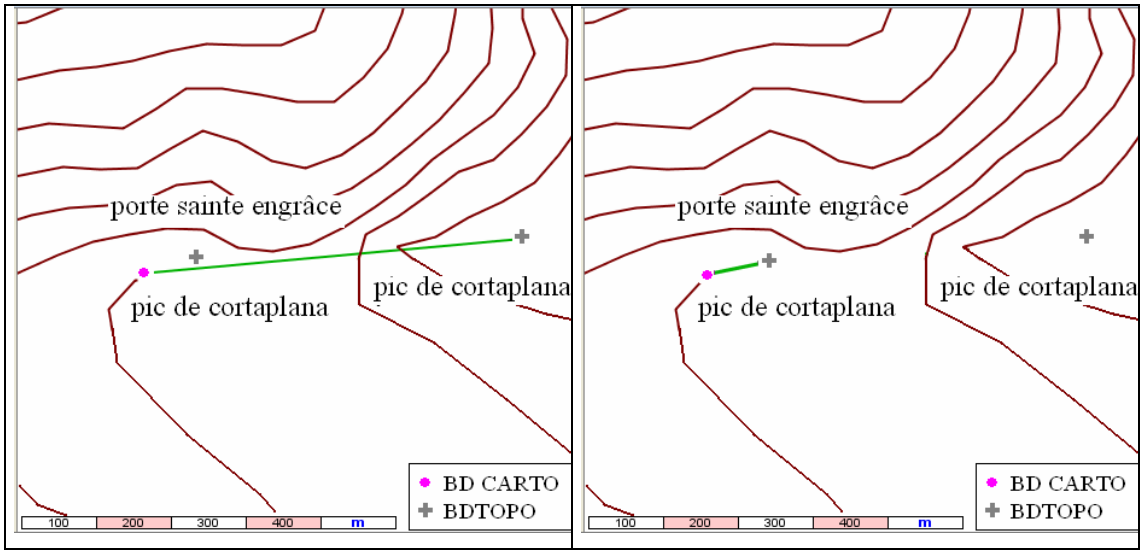


Figure 86. Résultat d'appariement illustrant le fait que le processus n'apparie pas au plus proche objet (à gauche) et résultat d'appariement basé sur l'approche géométrique de [Beeri *et al.*, 2004] (à droite)

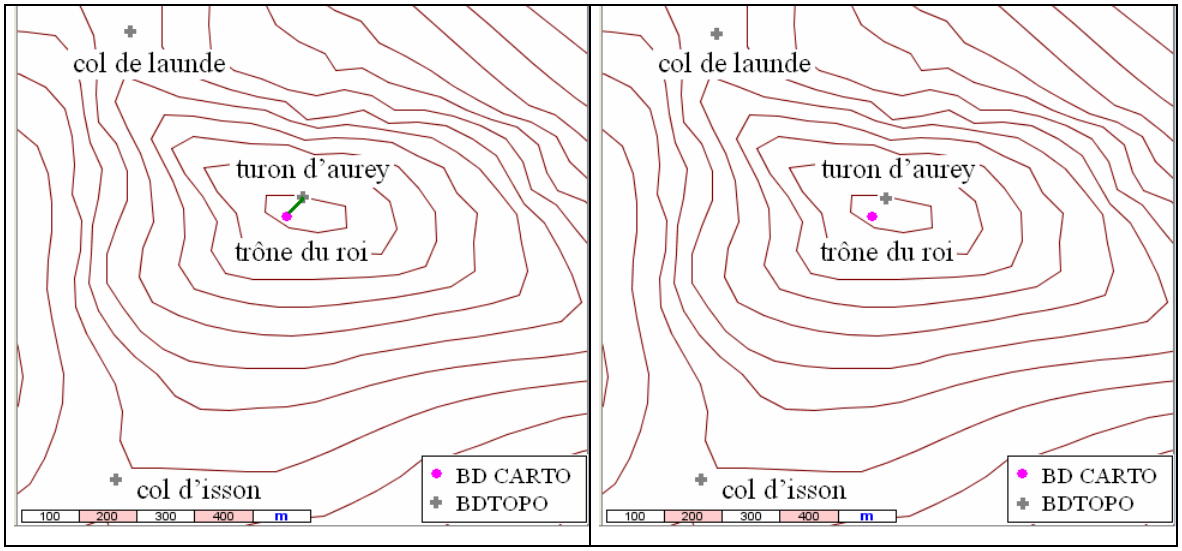


Figure 87. Résultat d'appariement illustrant le fait que le processus apparie deux objets ayant des toponymes différents (à gauche), et résultat d'appariement basé sur l'approche de [BDCARTO-BDTOPO, 2005] (à droite)

Nous illustrons sur la Figure 87 à gauche un exemple de résultat d'appariement où notre processus apparie deux objets ayant des toponymes différents. L'objet appartenant à la BDCARTO possède un toponyme en français : « trône du roi » tandis que son homologue dans la BDTOPO possède un toponyme écrit en béarnais : « turon d'aurey ». Ce résultat est possible grâce à la flexibilité des courbes définies pour le critère toponymique, c'est-à-dire que même si les toponymes sont différents, le candidat, en occurrence l'objet « turon d'aurey », n'est pas rejeté.

L'algorithme de [Beeri et al., 2004] donne le même résultat, tandis que l'algorithme de [BDCARTO-BDTOPO, 2005] n'apparie pas les deux objets en raison de la différence importante entre les toponymes des deux objets homologues.

En Figure 88 à gauche, nous montrons un exemple de résultat d'appariement qui illustre que notre processus apparie deux objets de nature « vallée » très éloignés, situés à 8,5 km l'un de l'autre. Ce résultat est possible d'une part grâce à la sélection des candidats à l'appariement en fonction de la nature des objets géographiques, c'est-à-dire que pour un objet de nature « vallée » le seuil de sélection est plus élevé que pour un objet de nature « sommet ». D'autre part, le résultat est possible grâce à la modélisation des courbes. En effet, en regardant l'écart de distance, nous croyons que les deux objets ne sont pas homologues, mais nous ne rejetons pas l'hypothèse contraire, même si elle est très faible. Les deux objets ayant la même nature (condition nécessaire et suffisante), nous sommes partagés entre l'hypothèse que les deux objets sont homologues et l'ignorance. Par contre, le fait que les deux objets ont le même toponyme nous amène à croire avec certitude que les deux objets sont homologues. Par conséquent, la fusion des trois critères fait que l'appariement est possible.

Le même résultat est obtenu en utilisant l'approche de [BDCARTO-BDTOPO, 2005], puisque elle est basée principalement sur la toponymie. Par contre, le résultat est différent si nous utilisons l'algorithme de [Beeri *et al.*, 2004], comme le montre la Figure 88 à droite. L'objet « vallée d'aspe » est apparié à l'objet le plus proche, « pène de lamounédère », qui est de nature « crête ».

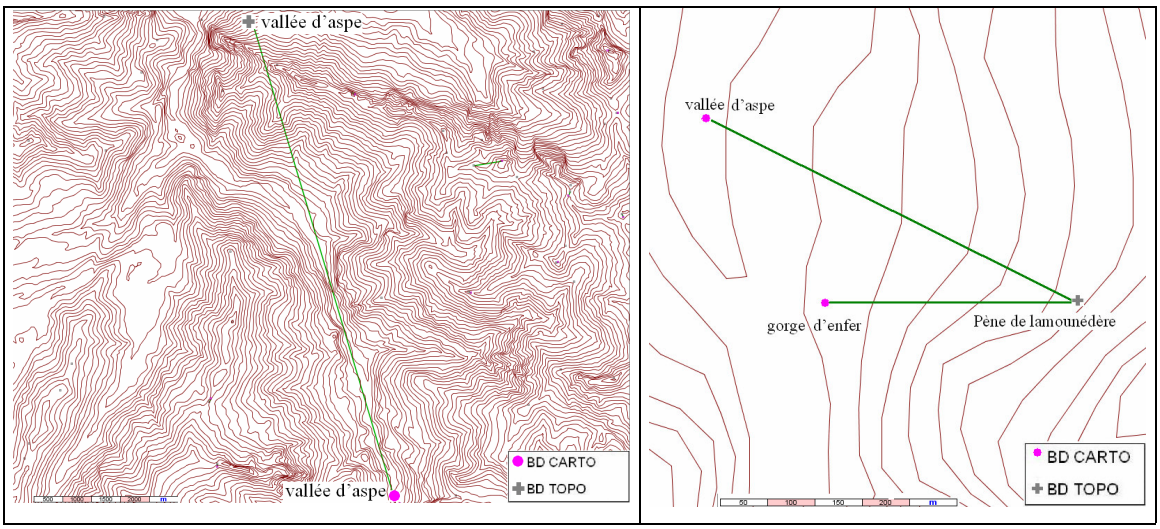


Figure 88. Exemple de résultats d'appariement d'objets très éloignés : avec notre processus (à gauche), et avec l'algorithme de [Beeri *et al.*, 2004] (à droite)

Nous montrons dans l'exemple illustré en Figure 89 l'intérêt de définir l'hypothèse NA, c'est-à-dire « l'objet n'est pas apparié » dans le processus d'appariement. Dans cet exemple, l'objet « pic de louesque » de la BDCARTO de nature pic n'a pas été apparié par notre processus. Ce résultat est juste, parce que nous sommes en limite de département et que l'objet homologue dans la BDTOPO, ayant la même nature et le même toponyme, se trouve dans l'autre département, et ne fait donc pas partie de notre jeu de données.

Comme nous pouvons le voir dans l'exemple, cet objet a trois candidats, qui ne sont pas très proches, portent des toponymes relativement différents et sont de nature différente : l'objet « soum de louesque » est de nature « sommet », tandis que les deux autres objets « passe de bourroux » et « col de louesque » sont de nature « col ». La fusion des critères et des candidats fait qu'aucun candidat ne s'est distingué et donc l'hypothèse NA a été choisie.

Notons que ce lien d'appariement a été évalué par notre processus comme étant incertain, du fait que les valeurs des probabilités pignistiques attribuées aux hypothèses relatives aux trois candidats et à l'hypothèse NA sont proches.

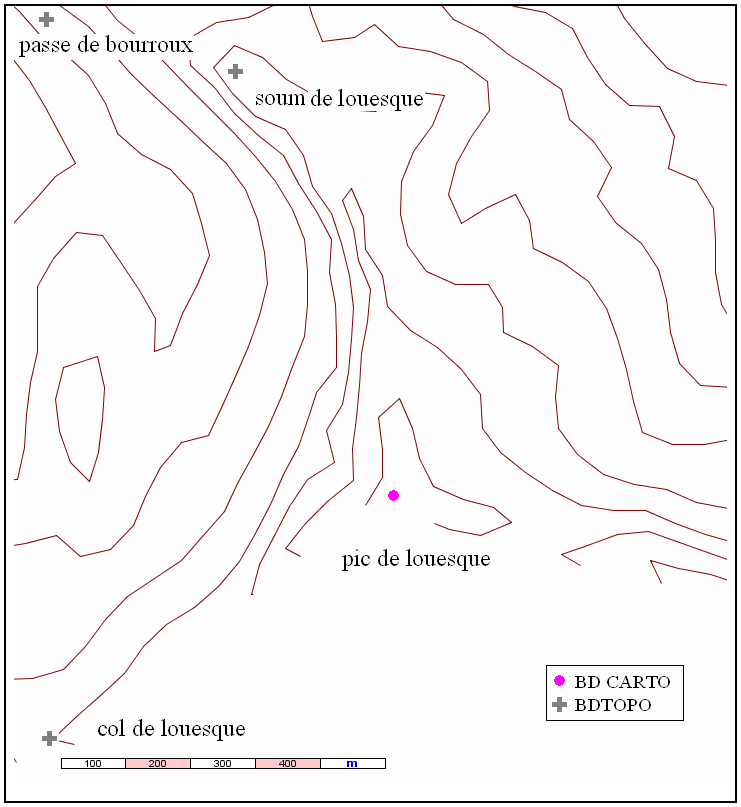


Figure 89. Exemple de résultat d'appariement lorsque l'objet n'est pas apparié

Dans les exemples suivants, nous montrons l'intérêt de fusionner plusieurs critères d'appariement.

Sur la Figure 90 à gauche, nous montrons un résultat d'appariement obtenu en utilisant seulement le critère d'écart de position et le critère toponymique. Le processus d'appariement n'arrive pas à apparier les deux objets homologues : « l'escarpu » de la BDCARTO et « l'escarpu ou pic de louesque » de la BDTOPO en raison des écarts de position et de distances toponymiques importants. En revanche, lorsque nous ajoutons le critère sémantique, le résultat de l'appariement est amélioré (voir la Figure 90, à droite). En effet, les deux objets ayant la même nature sont appariés, même si la différence entre l'hypothèse qui soutient ce candidat, en occurrence l'objet « l'escarpu ou pic de louesque », et l'hypothèse NA est faible.

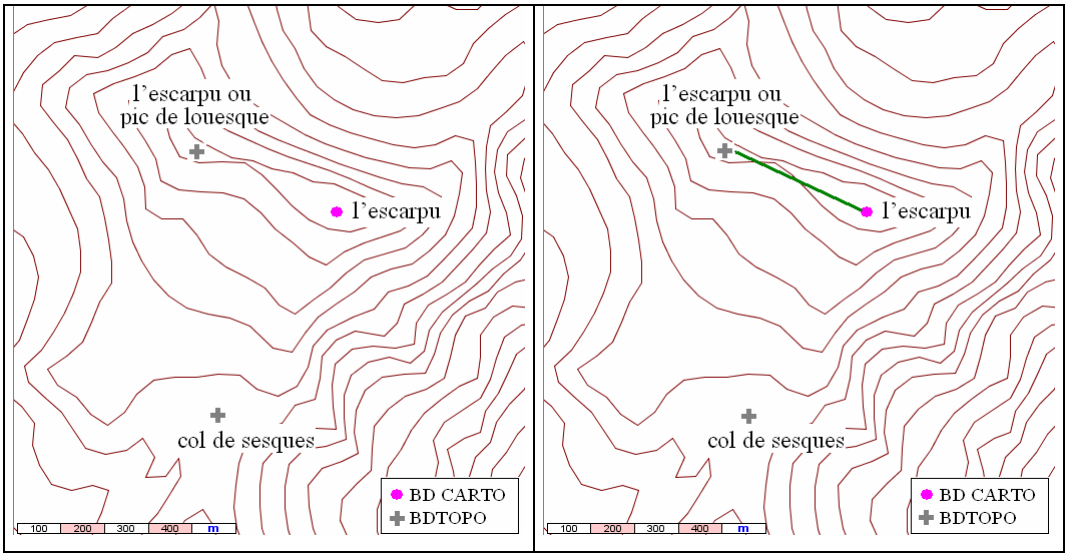


Figure 90. Exemple de résultat d'appariement obtenu en utilisant deux critères (à gauche) et trois critères (à droite)

Afin d'avoir une vue d'ensemble sur l'importance de l'utilisation de plusieurs critères, nous présentons, à travers des matrices de confusion des données appariées, basées sur l'attribut « Nature » des objets BDCARTO et BDTOPO, l'intérêt d'utiliser le critère sémantique. Le Tableau 14 représente la matrice de confusion déterminée à partir des données appariées en utilisant le critère d'écart de position et le critère toponymique, tandis que le Tableau 15 illustre la matrice de confusion calculée pour des données appariées en se basant sur les trois critères : d'écart de position, toponymique et sémantique.

Dans la matrice de confiance les appariements marqués en vert sont qualifiés de justes et ceux marqués en rouge sont qualifiés de faux. Par exemple dans le Tableau 14, nous voyons que 11 objets de la BDCARTO de nature « Cap, pointe » sont appariés à raison à 10 objets de nature « Cap » et à tort à 1 objet de nature « Dune, Isthme, Plage » de la BDTOPO.

BD Carto \ BD Topo	Cap	Cirque	Col	Versant	Dépression	Dune, Isthme, Plage	Ile	Pic	Plaine ou plateau	Récifs	Rochers, Escarpement	Crête, Sommet, Montagne	Vallée, Gorges	Volcan	Non-appariés
Cap, pointe	11	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Cirque	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Col, passage	0	0	106	0	0	0	0	0	0	0	0	1	1	0	0
Coteau, falaise	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0
Cuvette, dépression	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Dune, plage	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
Espace marin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ile	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Pic	0	0	2	0	0	0	0	52	0	0	1	8	0	0	4
Plaine, plateau	0	0	0	0	0	0	0	0	4	0	0	1	0	0	0
Récifs	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rocher	0	0	0	0	0	0	0	1	0	0	12	0	0	0	0
Sommet, crête, colline	0	0	2	0	0	0	0	1	0	0	0	120	2	0	8
Vallée	0	0	0	0	0	0	0	0	0	0	0	0	4	0	10
Volcan, cratère	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tableau 14. Matrice de confusion pour les données appariées pour le département 66 en utilisant deux critères

Nous remarquons que lorsque les trois critères sont utilisés, les résultats d'appariement sont améliorés.

D'une part, nous pouvons remarquer que de nombreuses erreurs d'appariement (par exemple un objet de nature « sommet » est apparié avec un objet de nature « col » ou un objet de nature « col » est apparié avec un objet de nature « vallée ») sont corrigées lorsque le critère sémantique est rajouté au processus d'appariement. D'autre part, il existe des cas où l'ajout du critère sémantique a comme conséquence la définition de nouveaux liens d'appariement justes. Dans le Tableau 15 nous pouvons voir que 122 objets de la base BDCARTO de nature « sommet, crête, colline » sont appariés avec des objets de la base BDTOPO de nature « crête, sommet, montagne », contrairement à 120 dans le Tableau 14. Il s'agit d'objets qui ont des homologues assez éloignés, des toponymes relativement différents, mais qui sont de même nature.

BD Carto \ BD Topo	Cap	Cirque	Col	Versant	Dépression	Dune, Isthme, Plage	Ile	Pic	Plaine ou plateau	Récifs	Rochers, Escarpement	Crête, Sommet, Montagne	Vallée, Gorges	Volcan	Non-appariés
Cap, pointe	11	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Cirque	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Col, passage	0	0	106	0	0	0	0	0	0	0	0	0	0	0	2
Coteau, falaise	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0
Cuvette, dépression	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Dune, plage	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
Espace marin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ile	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Pic	0	0	0	0	0	0	0	52	0	0	0	8	0	0	6
Plaine, plateau	0	0	0	0	0	0	0	0	4	0	0	0	0	0	1
Récifs	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rocher	0	0	0	0	0	0	0	0	0	0	12	0	0	0	1
Sommet, crête, colline	0	0	0	0	0	0	0	1	0	0	0	122	0	0	10
Vallée	0	0	0	0	0	0	0	0	0	0	0	0	5	0	10
Volcan, cratère	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tableau 15. Matrice de confusion pour les données appariées du département 66 en utilisant trois critères

La Figure 91 montre un exemple où une erreur d'appariement est corrigée après l'ajout du critère sémantique. L'objet de la BDCARTO ayant le toponyme « méhatzé » de nature « sommet » est apparié avec l'objet de la BDTOPO « col de méhatzé » de nature « col » lorsque seuls les critères d'écart de position et toponymique sont utilisés. Cette erreur est possible en raison d'une part d'un faible écart de position, et d'autre part du fait que les toponymes se ressemblent. En revanche, lorsque le critère sémantique est ajouté, l'erreur est corrigée puisque l'écart sémantique entre les concepts « sommet » et « col » est très important.

Mentionnons que le manque des courbes de niveau sur la Figure 91 est dû au fait que les objets se trouvent en limite de département.

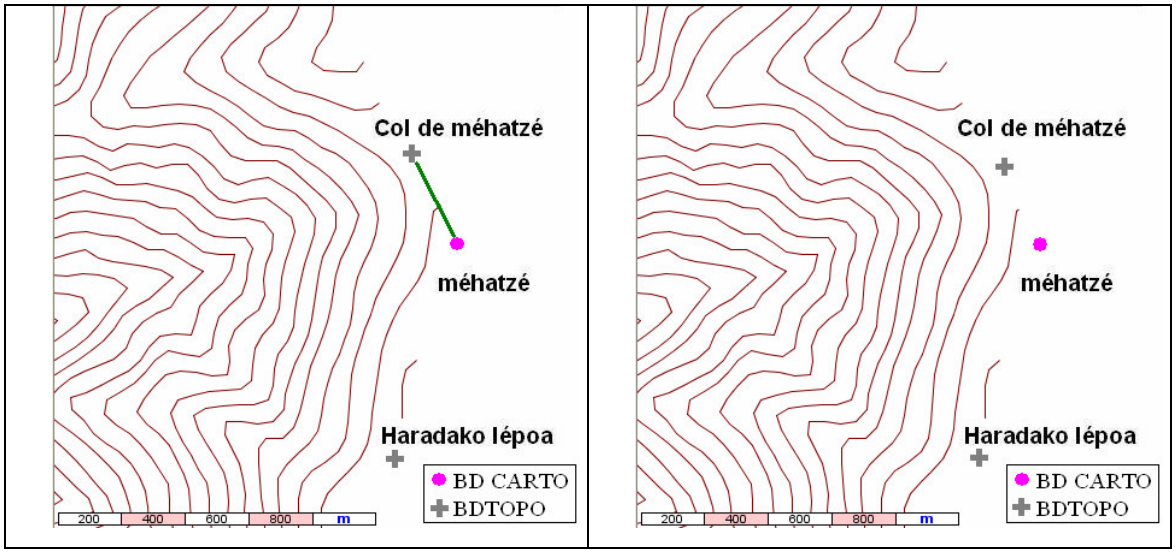


Figure 91. Exemple de résultat d'appariement obtenu en utilisant deux critères (à gauche) et trois critères (à droite)

Les exemples précédents ont illustré l'intérêt de notre approche. Celle-ci a néanmoins ses limites, que nous détaillons ci-dessous.

La cardinalité

Une première limite de notre approche est qu'elle définit des appariements de type 1 : 1. Nous donnons quelques pistes d'améliorations au chapitre E. Nous montrons en Figure 92 à gauche un résultat d'appariement qui est incomplet. L'objet de la BDCARTO « plages d'anglet » de nature « plage » est apparié avec l'objet de la BDTOPO « les corsaires » de nature « plage ». L'appariement est possible parce que les objets sont très proches et qu'ils ont la même nature. Dans cet exemple, nous remarquons en analysant la carte au 1 : 25 000 illustrée sur la Figure 92 à droite qu'il s'agit d'une agrégation d'entités dans la BDCARTO, c'est-à-dire que l'objet « plages d'anglet » représente l'ensemble des plages « les sables d'or », « les cavaliers », « la barre », etc. Dans la BDTOPO, toutes ces entités sont représentées indépendamment.

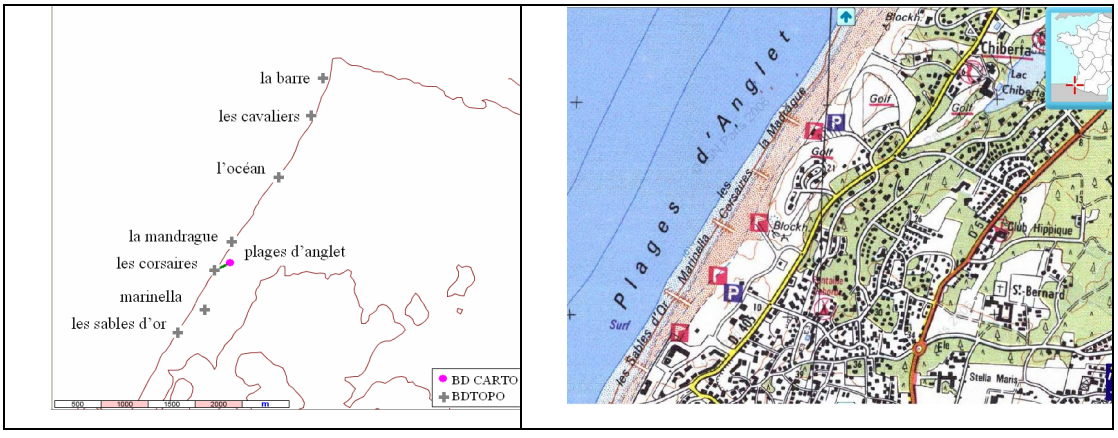


Figure 92. Exemple de sur-appariement (à gauche) et carte au 1 : 25 000 de la zone représentant notre exemple (à droite)

Le seuillage

Sur la Figure 93 nous illustrons un exemple de sous-appariement, c'est-à-dire un objet de la BDCARTO qui devait être apparié, mais qui ne l'est pas par erreur. Cette erreur est due au fait qu'aucun critère ne soutient le candidat à l'appariement. L'objet de la BDCARTO « gorges » a un candidat « gorges du bitet ». La distance entre les deux objets est de 490 m (donc supérieure à la précision de la base de données), la valeur de la nature des deux objets est « vallée » dans la BDCARTO et « gorge » dans la BDTOPO, et enfin comme nous pouvons le constater l'écart entre les deux toponymes est assez important. Par conséquent, la fusion des trois critères fait que l'hypothèse NA (non-apparié) est choisie. Signalons que l'appariement a été évalué comme incertain, c'est-à-dire que l'écart entre le premier maximum (la probabilité pignistique attribuée à l'hypothèse NA) et le deuxième maximum (la probabilité pignistique attribuée à l'hypothèse que l'objet « gorges » est apparié à l'objet « gorges du bitet ») est inférieure à 0,5. Concernant les deux autres approches que nous avons testées, l'approche de [Beeri *et al.*, 2004] n'apparie pas l'objet « gorges » à cause de l'écart de distance, tandis que l'approche de [BDCARTO-BDTOPO, 2005] apparie bien les objets « gorges » et « gorges du bitet » parce qu'elle s'appuie principalement sur la comparaison des toponymes et parce que le toponyme de l'objet de la BDTOPO commence avec le toponyme de l'objet de la BDCARTO. Par contre, l'appariement est évalué comme étant très incertain.

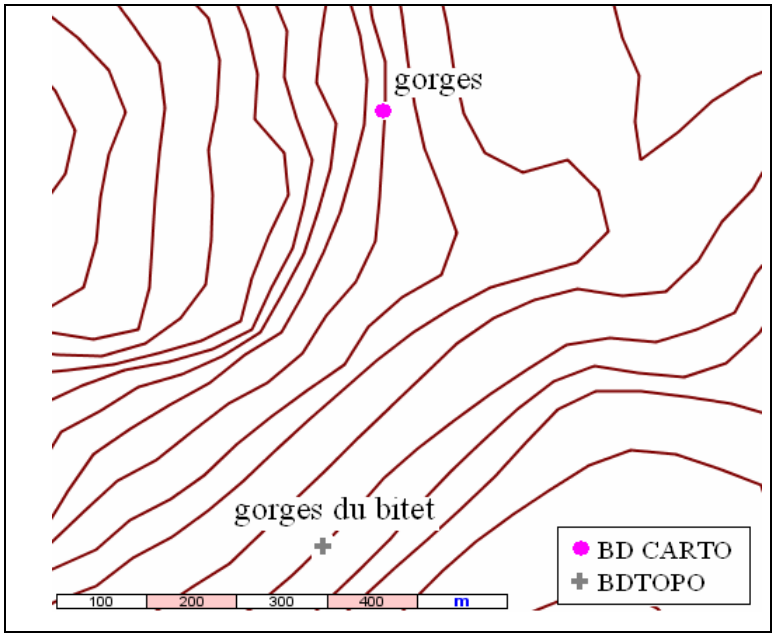


Figure 93. Exemple de sous-appariement

Ambiguïté d'interprétation

Sur la Figure 94, nous montrons un exemple de conflit de fusion, c'est-à-dire qu'après la fusion des candidats à l'appariement, la masse associée au conflit est égale à 1. Conformément à l'opérateur de Dempster, lorsque le conflit est égal à 1, la normalisation des masses de croyance ne peut pas être faite. Dans notre approche, afin d'être plus prudent, si le conflit est supérieur à 0,9, nous considérons qu'il y a un conflit et donc la normalisation n'est pas faite. Comme nous l'avons déjà dit, nous considérons que dans ce cas, l'appariement est difficile et qu'il nécessite l'intervention humaine. Ainsi, lors d'un conflit total, le processus d'appariement signale le conflit en émettant un avertissement.

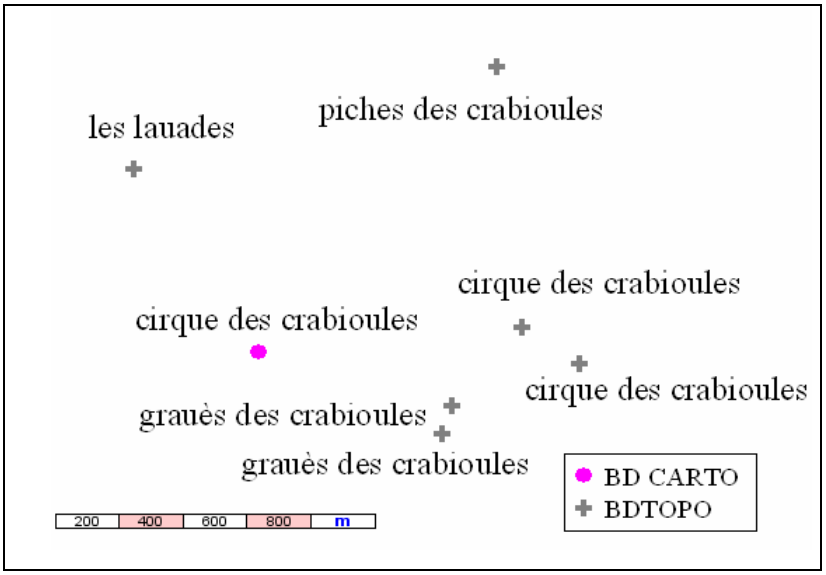


Figure 94. Exemple de conflit de fusion

Dans l'exemple illustré en Figure 94, nous pouvons remarquer que le conflit provient de la fusion des candidats. En effet, l'objet de la BDCARTO « cirque des crabioules » a six candidats à l'appariement dans la BDTOPO, qui sont assez proches. Parmi eux, il y en a deux de la même nature (« cirque ») et ayant le même toponyme (« cirque des crabioules »). Ainsi, lors de l'analyse indépendante des candidats, les trois critères d'écart de position, toponymique et sémantique soutiennent d'une manière sensiblement similaire les deux candidats de la BDTOPO « cirque des crabioules ». Cette analyse engendre un conflit lors de la fusion de ces deux candidats. De plus, le conflit est amplifié après la fusion des autres candidats, parce qu'il y encore un candidat « piches des crabioules » et deux candidats « grauès des crabioules » qui ont la même valeur de l'attribut nature (« dépression »), et qui sont relativement proches du point de vue de la distance géométrique et de l'écart toponymique.

Nous considérons que dans ce cas, c'est-à-dire deux objets dans la même base de données, très proches l'un de l'autre et ayant le même toponyme et la même nature, il est difficile de prendre une décision puisque la justesse des données dépend de la logique de saisie de l'opérateur. L'opérateur peut penser soit « le cirque est ici », soit « cet endroit est un cirque ». Dans le premier cas, un seul objet doit être saisi dans la base de données et donc le deuxième objet est une erreur de saisie, tandis que dans le deuxième cas deux ou plusieurs objets peuvent être saisis dans la base pour représenter la même entité. Notre processus a donc signalé ce problème par un conflit total et aucune décision n'est prise.

D.2.3 Evaluation quantitative

Nous présentons dans cette partie l'évaluation quantitative des résultats d'appariement obtenus avec notre processus d'appariement. Les résultats d'appariement ont été évalués d'une manière interactive, par département. L'évaluation concerne les objets de la BDCARTO.

D'abord, pour chaque département nous avons réalisé une référence, c'est-à-dire que nous avons fait un appariement manuel. Pour cela, nous avons utilisé les jeux de données géographiques appartenant aux bases de données BDCARTO et BDTOPO. Pour des cas difficiles pour lesquels nous ne pouvions pas réaliser un appariement manuel, d'autres informations ont été utilisées, telles que :

1. la base de données altimétriques, la BDALTI de l'IGN,
2. la base de données BDNYME de l'IGN représentant les toponymes des entités géographiques sur tout le territoire français,
3. les images aériennes,
4. les cartes à l'échelle de 1 : 25 000,
5. des experts de la cellule toponymique de l'IGN.

Une fois que l'appariement de référence a été réalisé, nous avons évalué nos résultats d'appariement automatique en nous basant sur celle-ci.

L'évaluation des résultats d'appariement par département est présentée dans le Tableau 16.

		Appariement manuel (nombre d'objets)	Appariement automatique (nombre d'objets)	Précision	Rappel
Département 11	Appariés	284	282 (1 à tort)	99,7%	99%
	Non-appariés	23	22 (0 à tort)	100%	95,6%
	Conflit total	-	3	-	-
Département 31	Appariés	96	95 (0 à tort)	100%	99%
	Non-appariés	14	14 (0 à tort)	100%	100%
	Conflit total	-	1	-	-
Département 64	Appariés	343	341 (1 à tort)	99,7%	99%
	Non-appariés	22	23 (2 à tort)	91%	95%
	Conflit total	-	1	-	-
Département 66	Appariés	329	329 (5 à tort)	98,4%	98,4%
	Non-appariés	31	28 (1 à tort)	96,4%	87%
	Conflit total	-	3	-	-
Département 69	Appariés	79	77 (0 à tort)	100%	97,4%
	Non-appariés	11	11 (0 à tort)	100%	100%
	Conflit total	-	2	-	-
Total	-	1232	1232	-	-

Tableau 16. Evaluation des résultats pour les départements 11, 31, 64, 66 et 69

La colonne désignée « conflit total » représente le nombre d'objets de la BDCARTO pour lesquels nous n'avons pas pu prendre de décision à cause du conflit total issu de la fusion des critères et des candidats. Rappelons que nous avons choisi de ne pas redistribuer le conflit, mais de le signaler, afin qu'un opérateur intervienne. Ce choix a été fait en raison du fait que nous considérons que si un conflit total existe, ceci est dû sûrement à une erreur ou à un cas géographiquement complexe qui mérite d'être traité manuellement. La référence est présentée

en nombre d'objets de la BDCARTO qui ont un homologue et en nombre d'objets qui n'ont pas d'homologue dans la BDTOPO.

L'évaluation proprement dite est présentée d'une part en nombre d'objets (la colonne Nombre d'objet) et d'autre part en termes de précision et de rappel (la colonne Précision et Rappel). La précision représente le nombre de liens justes par rapport au nombre de liens trouvés. Le rappel représente le nombre de liens justes trouvés par rapport au nombre de liens définis dans la référence. Notons qu'un objet non-apparié correspond à un lien de cardinalité 1 : 0.

Nous remarquons que pour les départements 31 et 69, nous avons obtenu une précision de 100% à la fois pour les objets qui ont un homologue et pour ceux qui n'en ont pas, c'est-à-dire que tous les liens de cardinalité 1 : 1 et 1 : 0 trouvés sont justes. Une moins grande précision est obtenue pour le département 66. Une précision de 98,4% pour les objets appariés signifie que parmi les liens d'appariement de cardinalité 1 : 1 trouvés, 1,6 % sont faux. En ce qui concerne les liens de cardinalité 1 : 0, la précision est de 96,4%.

En général, le rappel est inférieur ou égal à la précision. Un rappel inférieur est dû aux cas d'objets qui sont en conflit total. Le rappel le plus faible, 87%, a été obtenu pour le département 66 dans le cas des objets non-appariés. Ceci est dû d'une part au nombre d'objets en conflit total, et d'autre part au fait que cinq objets ont été mal appariés (quatre objets ont été sur-appariés et un objet a été sous-apparié).

Généralement, les erreurs d'appariement trouvées arrivent lorsqu'un objet de la BDCARTO a plusieurs candidats pour lesquels l'écart de distance est faible, les natures sont relativement proches et les toponymes sont différents. Ceci est dû surtout à la modélisation du critère toponymique qui est très important lorsque les toponymes sont identiques, mais qui n'a pas de poids lorsque les toponymes sont différents.

Concernant le nombre d'objets en conflit total, nous constatons que nous avons au maximum trois cas de conflit total par département. Les conflits totaux arrivent lorsque plusieurs candidats à l'appariement ont des caractéristiques similaires à la fois entre eux et avec l'objet en cours d'analyse. Par exemple, plusieurs candidats de la BDTOPO situés très proches les uns des autres ont la même nature et le même toponyme.

Chaque lien d'appariement est auto-évalué comme étant sûr ou incertain en fonction de la différence entre la valeur du premier maximum choisi, c'est-à-dire la probabilité pignistique maximale, et le deuxième maximum, c'est-à-dire la probabilité pignistique qui a la deuxième valeur. Nous rappelons que si la différence entre le premier et le deuxième maximum est supérieure à 0,5, alors nous considérons le lien comme sûr, dans le cas contraire il est considéré comme étant incertain. Précisons que tous les objets appariés ou non-appariés à tort ont été identifiés comme incertains. Parmi les objets bien appariés ou non-appariés il en existe qui ont été classifiés comme incertains, alors qu'ils sont sûrs. Ainsi, pour les points remarquables du relief, notre processus est pessimiste.

Afin d'avoir une vue d'ensemble sur l'évaluation des liens d'appariement, nous présentons pour chaque département un histogramme avec en abscisse la valeur de la différence entre le premier et le deuxième maximum et en ordonnée le nombre d'occurrences, c'est-à-dire le nombre de liens. Les histogrammes sont illustrés en Figure 95.

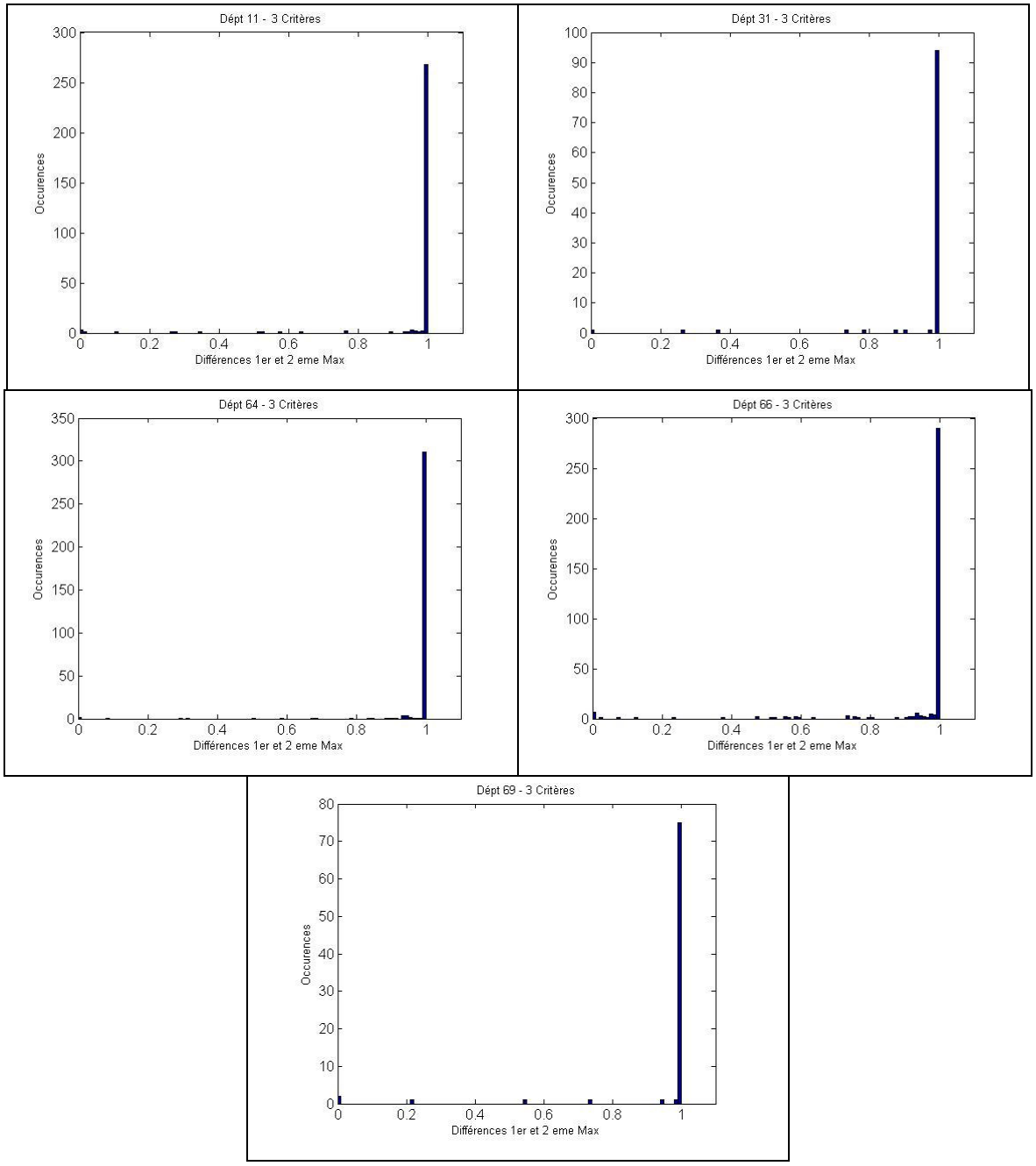


Figure 95. Histogrammes illustrant l'évaluation des liens pour les départements 11 (en haut à gauche), 31 (en haut à droite) 64 (au milieu à gauche), 66 (au milieu à droite) et 69 (en bas)

Nous remarquons que pour le département 11 par exemple, nous avons plus de 250 liens d'appariement pour lesquels la différence entre le premier et le deuxième maximum est égale à 1. Cela signifie que l'objet homologue choisi par le processus se distingue parmi tous les autres candidats à l'appariement. Dans ce cas les liens sont considérés comme sûrs. Le nombre de liens pour lesquels la différence entre le premier et le deuxième maximum est inférieure à 0,5 est très faible (9 liens). Dans ce cas, il n'existe pas de candidat qui se démarque parmi les autres et donc le processus d'appariement émet un avertissement.

La différence entre le premier et le deuxième maximum est égale à 0 lorsqu'un conflit total apparaît. Par exemple, nous observons que pour le département 31 nous avons un cas de conflit total. Un autre avertissement est lancé par notre processus pour l'identifier.

Globalement, nous remarquons que pour tous les départements, plus de 95% des liens d'appariement sont évalués comme étant sûrs. Plus précisément, le nombre de liens incertains par département est le suivant : 8 pour le département 11, 3 pour le département 31, 5 pour le département 64, 14 pour le département 66 et 3 pour le département 69.

Nous présentons dans le Tableau 17 une étude comparative entre les résultats obtenus par notre processus d'appariement (deuxième ligne du Tableau 17) et ceux obtenus par les algorithmes d'appariement de [BDCARTO-BDTOPO, 2005] (troisième ligne du Tableau 17) et [Beeri *et al.*, 2004] (quatrième ligne du Tableau 17). Cette comparaison a été faite uniquement pour le département 64. La lecture du Tableau 17 se fait de la même manière que pour le Tableau 16, présenté ci-dessus.

		Référence (nombre d'objets)	Nombre d'objets	Précision	Rappel
Notre approche	Appariés	343	341 (1 à tort)	99,7%	99%
	Non-appariés	22	23 (2 à tort)	91%	95%
	Conflit total	-	1	-	-
[BDCARTO-BDTOPO, 2005]	Appariés	343	329 (1 à tort)	99,6%	95,6%
	Non-appariés	22	36 (14 à tort)	61%	100%
[Beeri *et al.*, 2004]	Appariés	343	341 (20 à tort)	94,1%	93,5%
	Non-appariés	22	24 (11 à tort)	54%	59%

Tableau 17. Evaluation des résultats pour le département 64 en utilisant trois approches d'appariement

En analysant le Tableau 17, nous observons que les résultats d'appariement de données issus de notre processus et ceux issus du processus de [BDCARTO-BDTOPO, 2005] sont similaires pour les objets appariés, c'est-à-dire ceux qui ont un homologue dans la BDTOPO, les deux processus obtenant une précision d'environ 99%.

En revanche, notre processus donne une meilleure précision pour les objets non-appariés, c'est-à-dire les objets qui n'ont pas d'homologue dans la BDTOPO, grâce à l'utilisation des trois critères d'appariement et de l'hypothèse NA (non-apparié) que nous avons ajoutée dans l'ensemble des hypothèses possibles. L'algorithme de [BDCARTO-BDTOPO, 2005] a une faible précision pour les objets non-appariés parce qu'il ne permet pas de gérer l'imperfection au niveau de toponymes, l'approche étant binaire. Si l'écart toponymique entre deux objets est inférieur à un seuil, alors les objets sont appariés, sinon ils ne le sont pas. L'algorithme de [Beeri *et al.*, 2004], quant à lui, a une faible précision pour les objets non appariés parce qu'il n'utilise ni la toponymie, ni la sémantique (ainsi il a tendance à apparier aux objets les plus proches).

Etude de la sensibilité du processus aux seuils de sélection des candidats

Le processus est dépendant du seuil de sélection, dans le sens où s'il est trop petit, alors il peut y avoir des objets de la BDCARTO qui n'ont pas de candidat et donc ne sont pas appariés à tort. Par contre, si le seuil est très élevé, alors les résultats d'appariement sont

similaires, et les seules différences entre les résultats d'appariement obtenus sont liées au temps de calcul (plus il y a de candidats, plus le processus est long) et aux valeurs de confiance attribuées aux liens d'appariement. Par exemple, plus il y a de candidats, plus, au niveau de la décision, l'écart entre deux hypothèses simples est petit et donc l'écart entre le premier et le deuxième maximum est faible. Ceci signifie qu'aucun candidat ne se détache particulièrement.

Le seuil de sélection des candidats a été déterminé expérimentalement.

Etude de la sensibilité du processus aux seuils choisis pour les distances

Afin d'étudier la sensibilité aux seuils déterminés pour les distances euclidienne, sémantique et toponymique, nous avons réalisé plusieurs tests en faisant varier leur valeur pour un seul critère à la fois.

L'évaluation des résultats ayant été faite manuellement, les tests ont été réalisés uniquement sur le département 31, pour lequel nous avons obtenu une précision de 100%, et qui contient un nombre limité d'objets dans la BDCARTO (100 objets).

Pour le critère d'écart de position, nous avons défini deux seuils T_2 et T_1. Rappelons que les résultats que nous avons présentés ci-dessus ont été obtenus avec les seuils T_2 = 800 m et $T_1 = T_2 /2$. Les résultats d'appariement sont similaires pour un seuil T_2 appartenant à l'intervalle]50 m – 1400 m]. Pour un seuil inférieur à 50 m, nous avons obtenu un lien d'appariement de moins (à tort), tandis que pour un seuil dans l'intervalle]1400 m – 3000 m] nous avons obtenu à tort deux liens d'appariement de plus. Pour un seuil supérieur à 3000 m, le seul changement concerne le nombre de cas de conflit total.

Nous avons aussi étudié la sensibilité aux seuils fixés pour les critères sémantique et toponymique. En sachant que les distances sémantique et toponymique sont comprises entre 0 et 1, nous avons fait varier le seuil de 0 à 1 avec un pas de 0,1. Nous n'avons constaté aucun changement dans les résultats obtenus pour le département 31. Par contre pour les autres départements une faible sensibilité est obtenue. Les résultats qui varient avec les seuils représentent des cas géographiquement complexes.

En conclusion, nous pouvons dire que notre processus ne dépend pas des seuils fixés pour les distances utilisées. A notre avis, cette sensibilité faible est obtenue grâce à l'initialisation prudente des masses, à la fusion des critères, et au fait que nous n'avons jamais éliminé de candidat, même si celui-ci avait peu de chance qu'il soit l'homologue de l'objet en cours d'analyse.

Enfin, nous présentons dans le Tableau 18 le temps de calcul par département pour notre processus d'appariement. Nous remarquons qu'en moyenne, le temps de calcul est de 0,3 secondes par objet. La rapidité du processus dépend d'une part du nombre d'objets dans le jeu de données et d'autre part de la nature des objets. Par exemple, un département qui contient de nombreux objets ayant une grande étendue, comme les vallées par exemple, nécessite plus de temps de calcul. En effet nous avons choisi un seuil de sélection élevé, ce qui implique un nombre important de candidats à traiter.

	Dép 11	Dép 31	Dép 64	Dép 66	Dép 69
Temps de calcul	1'18"	0'28"	1'58"	1'34"	0'46"

Tableau 18. Temps de calcul par département pour le processus d'appariement

Bilan des expérimentations sur les points remarquables du relief

Nous avons présenté dans cette partie les expérimentations que nous avons mises en oeuvre afin de valider notre approche d'appariement de données. Ces expérimentations ont été réalisées sur cinq jeux de données différents, représentants des points remarquables du relief.

Nous avons remarqué, à travers les résultats que nous avons présentés ci-dessus, que notre processus d'appariement de données géographiques donne des résultats satisfaisants, une précision moyenne de 99% et un rappel moyen de 97% étant atteints. Cependant, il reste quelques cas d'appariement que le processus d'appariement n'arrive pas à gérer d'une manière efficace. Dans le chapitre E, nous présentons plus en détail les limites du processus, ainsi que les perspectives que nous proposons pour y remédier.

D.3 Etude des réseaux routiers

La deuxième expérimentation que nous avons réalisée concerne les réseaux routiers. Afin d'étudier la flexibilité de notre approche, nous avons utilisé deux jeux de données produits par deux producteurs différents, ayant des niveaux de détail différents, et contenant des données sensiblement différentes en contenu.

Dans cette partie, nous présentons d'abord les données que nous avons utilisées. Ensuite, nous montrons les résultats que nous avons obtenus. Enfin, une évaluation des résultats est faite en termes de précision et de rappel.

D.3.1 Présentation des données

Nous avons utilisé pour nos expérimentations deux jeux de données issus de la base de données BDCARTO produite par l'IGN [BDCARTO, 2004] et de la base de données MultiNet produite par TeleAtlas [MultiNet, 2004]. La zone d'étude se situe dans le département de la Seine-Maritime (76). Les jeux de données que nous avons sélectionnés couvrent une surface d'environ 760 km^2 et représentent à la fois des zones urbaines et des zones rurales. Pour cette superficie, le jeu de données issu de la BDCARTO contient 2063 arcs (1363 km) et 1399 nœuds, tandis que le jeu de données issu de MultiNet contient 12 725 arcs (19357 km) et 10 924 nœuds.

Rappelons que la BDCARTO a été produite par l'IGN pour faire de la cartographie à l'échelle de 1 : 100 000 ou 1 : 250 000 et pour réaliser des analyses à des niveaux régionaux et départementaux. La précision de la base est de un à plusieurs décamètres. Le réseau routier fait partie des quinze thèmes de la BDCARTO. Les routes formant le réseau routier ont des informations attributaires telles que la classification, la vocation, le numéro de route, le nom, l'état physique de la route, le sens de circulation.

La base de données MultiNet a été produite par la compagnie privée TeleAtlas. Elle couvre une grande partie de l'Europe, les Etats-Unis d'Amérique et le Canada et une petite partie d'Asie avec une précision générale de douze mètres, allant jusqu'à 5 m dans des zones telles que les zones urbaines denses. La base de données géographiques contient treize thèmes dont le thème routier. Ce dernier représente principalement les routes et les rues, et est utilisé dans les applications de navigation. Les objets ont de nombreux attributs tels que la classification, le nom des rues, le numéro de route, le sens de circulation.

Les deux bases de données géographiques sont très différentes. Dans cette partie nous allons voir en quoi cette différence consiste.

Comme le montre la Figure 96, le jeu de données appartenant à MultiNet est plus détaillé que celui appartenant à la BDCARTO. Néanmoins, il existe des entités du monde telles que

les chemins et les sentiers qui sont présents dans la BDCARTO mais qui ne le sont pas dans MultiNet.

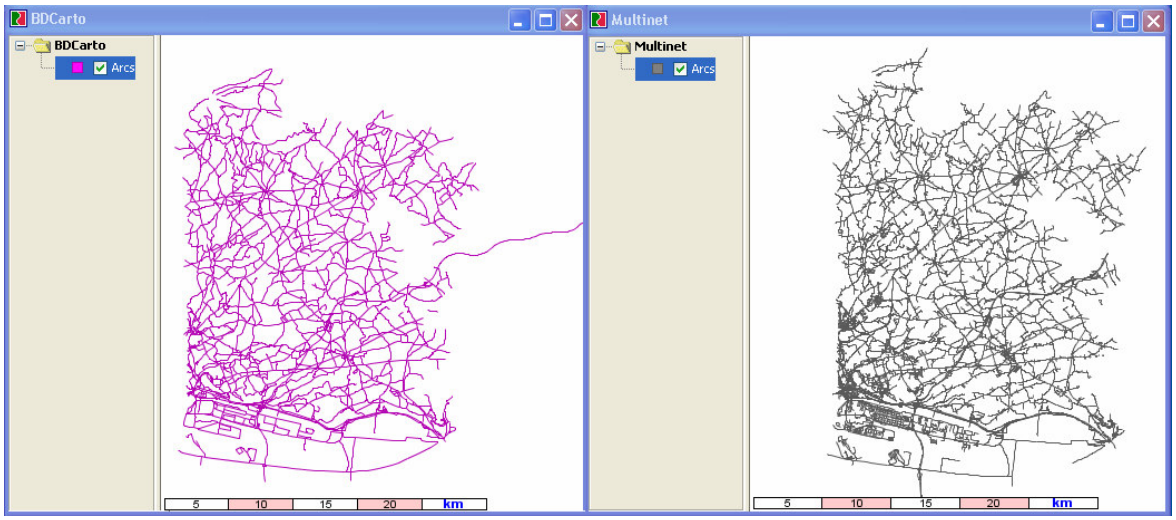


Figure 96. La zone d'étude utilisée : la BDCARTO (à gauche) et MultiNet (à droite)

La différence de détail n'est pas la seule différence entre les deux jeux de données. Il existe des différences de modélisation et de représentation : chaque base de données a sa propre représentation du monde réel en fonction de ses applications, sa perception du monde réel et ses objectifs.

Ainsi, pour les voies à chaussées séparées, conformément aux spécifications de la BDCARTO®, deux modélisations sont possibles : « si elles sont contiguës, la BDCARTO contient un tronçon à deux chaussées séparées ; si elles sont éloignées de plus de 100 m sur au moins un kilomètre de long, la BDCARTO contient deux tronçons (au moins) à une chaussée et en parallèle » [BDCARTO, 2004]. Au contraire, dans MultiNet, d'une manière générale, une route à chaussées séparées est représentée par deux tronçons parallèles représentant l'axe central de chaque chaussée. Un exemple qui illustre la représentation des autoroutes dans les deux jeux de données utilisés est montré en Figure 97. Nous remarquons sur cette figure que d'une part la même entité du monde réel est représentée différemment, et d'autre part le numéro de route est différent : « A29 » dans la BDCARTO et « E 44 » dans MultiNet.

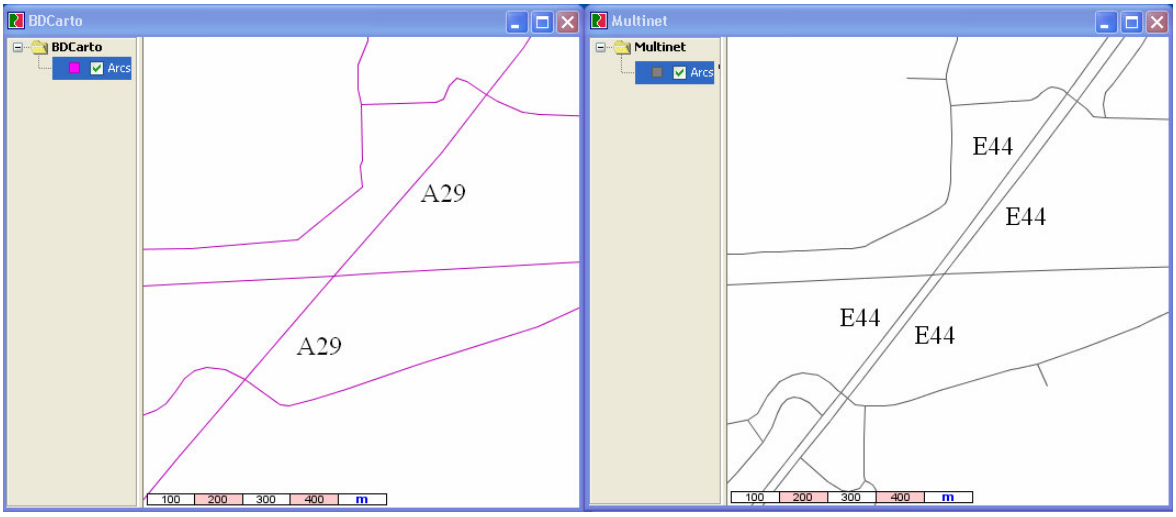


Figure 97. Représentation des autoroutes dans la BDCARTO (à gauche) et MultiNet (à droite)

Une autre différence importante concerne les ronds points et les carrefours complexes qui sont représentés par des arcs dans MultiNet et par des points appartenant à la classe *Nœuds du réseau routier* dans la BDCARTO (voir la Figure 98). Signalons que nous n'avons pas utilisé cette classe dans notre processus d'appariement. Seule la classe *Tronçons de routes* a été employée.

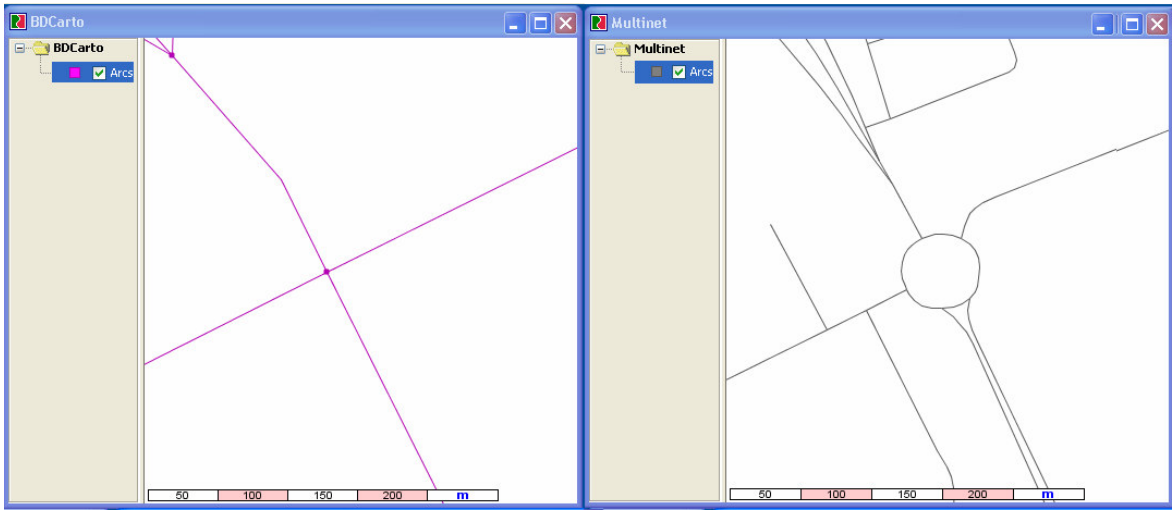


Figure 98. Représentation d'un rond-point dans la BDCARTO (à gauche) et dans MultiNet à (droite)

En ce qui concerne l'écart géométrique des deux objets homologues, nous avons constaté un écart planimétrique moyen de l'ordre de 10 m. Celui-ci peut atteindre une valeur plus importante, en particulier au niveau des carrefours complexes, jusqu'à 100 m.

Les deux jeux de données ont un attribut qui renseigne sur le numéro de route, par exemple D20, N15. Bien que cet attribut ne soit pas toujours rempli, nous considérons que c'est une information qui peut être exploitée dans le processus d'appariement. Pour le jeu de données de la BDCARTO, seulement 40% des tronçons de route ont un numéro de route rempli, tandis que pour celui issu de MultiNet, 77% des tronçons de route possèdent un numéro de route

rempli. De plus, la numérotation n'est pas toujours la même. Il s'agit surtout des autoroutes qui sont nommées différemment. La BDCARTO utilise la numérotation française tandis que dans MultiNet une numérotation européenne est employée. Un tel exemple de numérotation différente est illustré en Figure 97 : l'autoroute « A29 » dans la BDCARTO et « E44 » dans MultiNet.

La sémantique des tronçons de route est traduite par le classement administratif des routes. Les deux jeux de données que nous avons utilisés présentent un attribut qui renseigne sur cet aspect de la sémantique. Par contre, les deux classifications sont très différentes. La BDCARTO possède un attribut « type de classement » qui renseigne sur la classification administrative des routes. Selon cet attribut les routes sont classées dans quatre classes : les autoroutes, les routes nationales, les routes départementales, et les autres routes (rues, liaisons régionales, bretelles, carrefours dénivelés, etc.) [BDCARTO, 2004]. Dans la base de données MultiNet, il existe un attribut appelé « FRC » (Functional Road Class) qui renseigne sur la classification des routes en fonction de leur importance. L'attribut peut prendre l'une des valeurs suivantes : autoroute ou autre route prioritaire, route prioritaire moins importante que les autoroutes, autre route prioritaire, route secondaire, route locale d'importance élevée, route locale d'importance moins élevée, et autre route [MultiNet, 2004].

Les logiques de classification font qu'une comparaison immédiate entre les types de tronçon de route des deux jeux de données est très difficile, voire impossible. Une étape importante pour appréhender cette sémantique est d'étudier plus en détail les données elles-mêmes afin d'avoir une idée de la manière dont les attributs sont remplis. Par exemple, la valeur de l'attribut « autre route prioritaire » de MultiNet correspond dans les données à des routes européennes, nationales et départementales.

Pour apparier les deux jeux de données nous avons exploité les informations suivantes : la localisation des tronçons de route, leur orientation, le numéro de route, la nature des tronçons de route et la topologie du réseau routier. D'autres informations contenues dans les jeux de données auraient pu être exploitées, comme par exemple le sens de circulation, le nombre de voies ou le type de carrefour.

D.3.2 Tests

Afin de tester notre approche d'appariement de données pour les réseaux routiers, nous avons utilisé le critère d'écart de position, le critère d'orientation, le critère sémantique, le critère nom d'objet et le critère voisinage dont les principes sont présentés dans la partie C.3.4.2.

Le processus d'appariement est appliqué aux réseaux, c'est-à-dire que les données initiales (points et lignes) ont été d'abord transformées en une structure de réseau (nœuds, arcs, faces). Pour cela, nous avons utilisé la carte topologique présentée en Figure 76.

Pour apparier les deux réseaux, nous cherchons pour chaque arc appartenant au réseau plus détaillé (MultiNet) un arc homologue dans le réseau moins détaillé (BDCARTO). Ainsi, les candidats à l'appariement appartiennent au réseau le moins détaillé.

D.3.2.1 Paramétrage des courbes pour l'initialisation des masses de croyance

Nous discutons dans cette sous-partie du paramétrage des courbes spécifiques à chaque critère utilisé, c'est-à-dire des distances et des seuils que nous avons déterminés afin de réaliser nos expérimentations.

Sélection des candidats

Afin de sélectionner des candidats, nous nous sommes inspirés du processus d'appariement de [Mustière et Devogele, 2008]. Ainsi, l'étape de sélection des candidats est la suivante : pour chaque arc du jeu de données plus détaillé, nous cherchons des arcs candidats à l'appariement dans le jeu de données moins détaillé. Les candidats sélectionnés sont ceux qui se trouvent à une distance inférieure à un seuil fixé, ainsi que ceux qui se trouvent à une distance proche de la distance à l'arc le plus proche.

Plus précisément, supposons que pour l'arc A_1 du jeu de données le plus détaillé, nous cherchons les arcs candidats. La recherche des candidats est composée de deux étapes :

- d'abord, nous sélectionnons les arcs pour lesquels la distance entre chacun d'eux et l'arc A_1 est inférieure à un seuil S. A partir de cette première sélection, nous déterminons la distance qui correspond à l'arc candidat le plus proche, $d(A_1, A_{2\text{plus proche}})$.
- ensuite, nous filtrons les candidats afin de ne garder que ceux qui sont assez proches. Par conséquent, les candidats restants sont ceux qui se trouvent à une distance inférieure à la distance $d = \min (S, d(A_1, A_{2\text{plus proche}}) + d_{min}))$. La distance d_{min} représente la distance minimum au-dessous de laquelle l'écart de distance pour les arcs du réseau le moins détaillé n'a plus aucun sens.

La distance entre les arcs que nous avons utilisée est la demi-distance de Hausdorff, présentée au chapitre A. Le seuil S représente l'écart maximum de position entre les arcs homologues dans les deux jeux de données. La distance d_{min}, quant à elle, représente la précision géométrique du réseau le moins détaillé.

Dans nos tests, le seuil de sélection a été fixé assez largement à S = 100m, et la distance d_{min} a été fixée à 30m, qui est de l'ordre de grandeur de la précision de la base BDCARTO.

De cette manière, la sélection des candidats est pertinente : d'une part dans le cas où l'écart planimétrique est important, des candidats sont quand même sélectionnés grâce au seuil très large, d'autre part, dans les endroits où l'écart planimétrique avec au moins un candidat est très faible, elle permet d'éliminer des candidats qui ne sont pas pertinents.

Critère d'écart de position

Afin de mesurer l'écart entre deux arcs des réseaux routiers, nous avons utilisé la demi-distance de Haussdorf : elle donne en général une bonne indication de l'écart maximum entre deux arcs, et elle est facile à implémenter et plus rapide en temps de calcul que d'autres distances, par exemple la distance de Fréchet. Cependant, il existe des cas où elle donne une fausse indication, par exemple lorsque les courbes présentent une forte sinuosité ou lorsque les arcs à comparer sont décalés. En Figure 99 nous illustrons un exemple où la demi-distance de Hausdorff d_H est portée sur l'extrémité des arcs. Bien que les deux arcs soient homologues, l'écart de distance nous donne une information non adaptée.

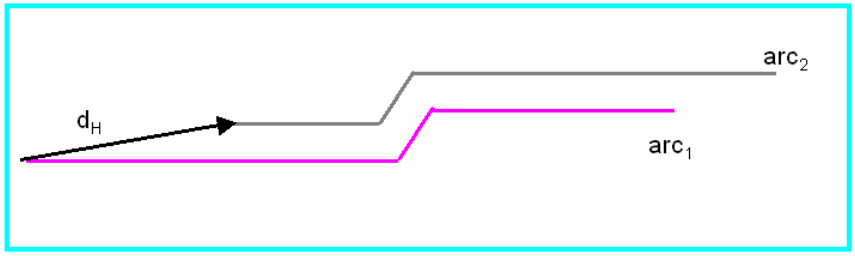


Figure 99. Exemple de demi-distance de Hausdorff entre deux arcs ayant les extrémités décalées

Afin de remédier à ces limites de la demi-distance de Haussdorf, nous proposons à travers la modélisation du critère de ne pas éliminer définitivement un candidat à l'appariement trop éloigné.

Les courbes que nous avons utilisées pour exprimer les connaissances relatives à l'écart de position sont celles qui ont été présentées dans le Tableau 2 du chapitre C et que nous avons rappelées dans cette partie lors de l'expérimentation pour les points remarquables du relief. Par rapport aux expérimentations que nous avons réalisées pour les points remarquables du relief, nous avons utilisé d'autres valeurs pour les seuils T_1 et T_2. Ainsi, T_1 qui représente la précision du jeu de données moins détaillé, est fixé à 30 m et T_2 est le double du seuil T_1, c'est-à-dire 60 m (sachant qu'en milieu urbain nous pouvons utiliser des seuils plus bas qu'en milieu rural).

Critère orientation

L'orientation des arcs a été également comparée à travers le critère d'orientation. Les courbes qui modélisent ce critère sont les mêmes que celles que nous avons présentées dans le Tableau 3 du chapitre C. Nous les rappelons dans le tableau suivant.

Nous remarquons que ce critère est moins important que les autres, puisque l'ignorance a une valeur importante et constante indépendamment de la valeur de l'orientation. Une des raisons est la mesure utilisée pour définir ce critère. En effet, l'orientation des arcs est étudiée localement, c'est-à-dire là où les arcs sont les plus proches, et non pas globalement. Ainsi, par exemple, il suffit que deux lignes soient parallèles au point le plus proche pour que la valeur de l'angle soit égale à 0.

Une autre raison liée à l'importance de ce critère est qu'un arc peut avoir plusieurs arcs candidats qui sont parallèles, sans qu'aucun arc ne soit le vrai homologue. Cela explique pourquoi nous n'attribuons pas une forte masse de croyance à l'hypothèse $appC_i$ lorsque par exemple le candidat C_i a la même orientation que l'arc en cours d'analyse.

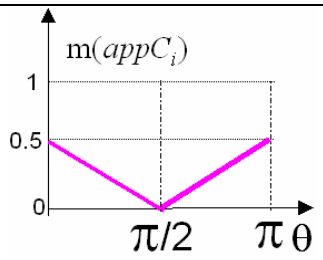
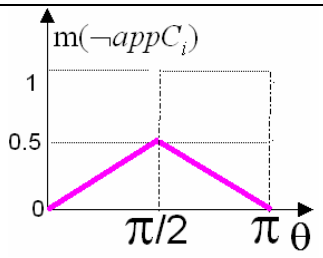
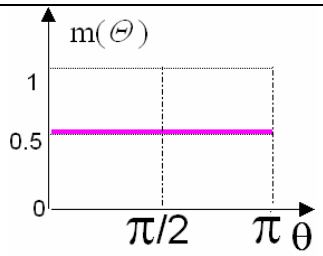
Hypothèse	Critère Orientation
$appC_i$	
$\neg appC_i$	
Θ	

Tableau 19. Représentation des connaissances pour le critère orientation

Critère sémantique

Les courbes que nous avons utilisées pour modéliser le critère sémantique sont les mêmes que celles présentées dans le tableau 4 du chapitre C et que nous avons rappelées dans cette partie lors de l'expérimentation pour les points remarquables du relief. Signalons que les mêmes courbes ont été utilisées pour apparier les points remarquables du relief.

Etant donné que la sémantique des deux jeux de données est très différente et que nous n'avons pas eu à notre disposition de taxonomie de domaine commune à ces deux jeux de données, ou à défaut une pour chaque jeu de données, la mesure qui évalue l'écart sémantique entre deux valeurs de l'attribut « nature/type » a été définie manuellement.

L'intervention d'experts pour calculer cet écart sémantique nous a semblé difficile à mettre en œuvre, puisqu'il est impossible de comparer deux types de classification, par exemple : « routes départementales » et « autres routes prioritaires » sans avoir une bonne connaissance des données.

Afin de calculer la mesure d'écart sémantique, nous nous sommes appuyés sur la taxonomie de domaine réalisée au laboratoire COGIT pour les bases de données de l'IGN BDCARTO et BDTOPO [Abadie et Mustière, 2008].

La taxonomie de domaine utilisée pour les réseaux routiers est composée de deux grandes classes : la classe « voie de communication carrossable » (voir la Figure 100) et la classe « voie de communication non-carrossable » (voir la Figure 102). Afin d'illustrer le niveau de détail de la taxonomie, nous présentons sur la Figure 101 les sous-concepts du concept route.

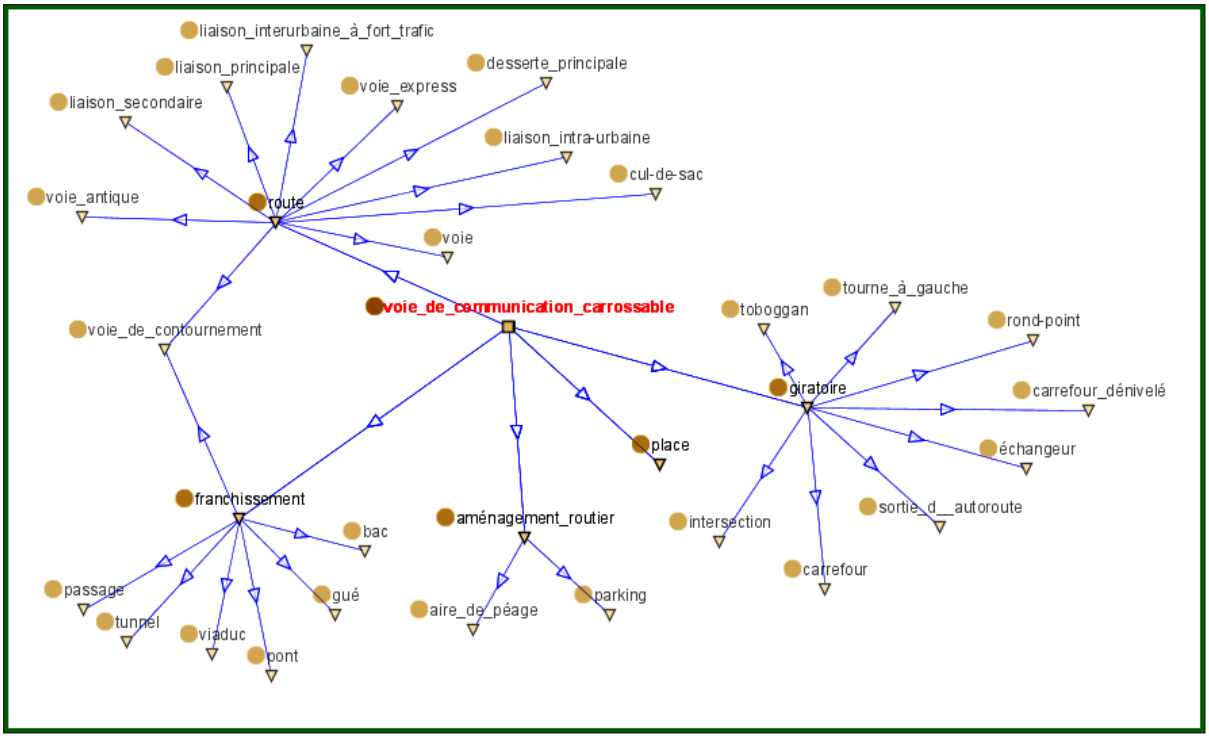


Figure 100. Taxonomie pour la classe « voie de communication carrossable » [Abadie et Mustière, 2008]

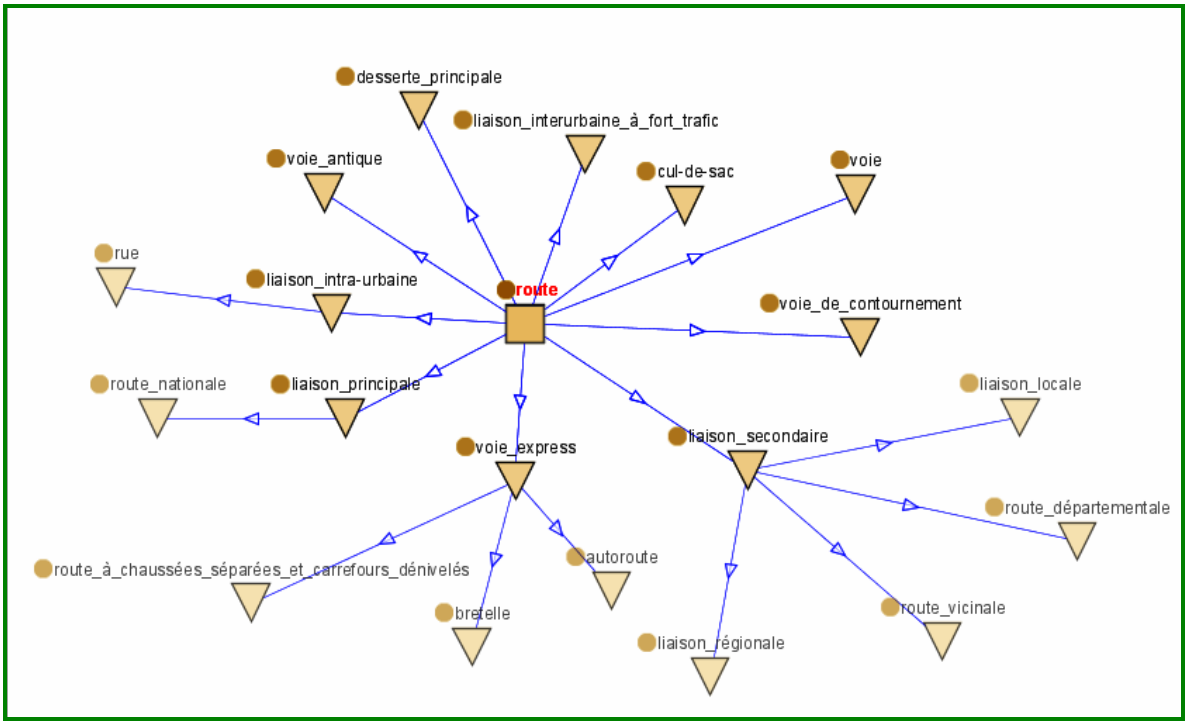


Figure 101. Taxonomie pour la classe « route » [Abadie et Mustière, 2008]

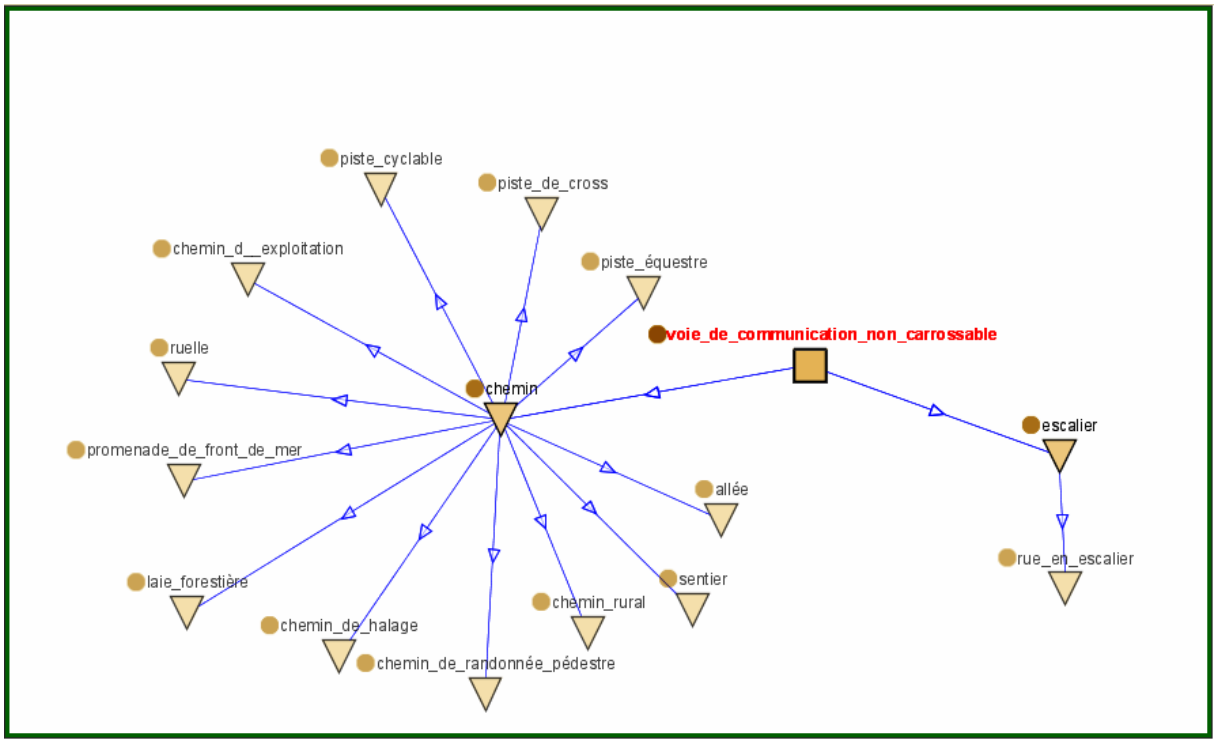


Figure 102. Taxonomie pour la classe « voie de communication non carrossable » [Abadie et Mustière, 2008]

Nous avons également analysé en détail les données issues de MultiNet afin de voir les instances correspondant à un type de classement. L'analyse des données consiste à identifier interactivement la correspondance entre les instances et la valeur de l'attribut FRC. Pour ce faire, nous avons utilisé les attributs qui renseignent sur le numéro et le nom des routes présents dans la base MultiNet. Ainsi, par exemple si le numéro de la route est « D67 » nous savons qu'il s'agit d'une départementale ou si le numéro n'existe pas nous regardons le nom et nous pouvons déterminer s'il s'agit d'une rue, d'une place, d'une avenue, etc. Après cette analyse nous avons pu constater, par exemple, que les routes nationales peuvent avoir comme type soit « route prioritaire moins importantes que l'autoroute », soit « autre route prioritaire », soit encore « route secondaire ».

Les correspondances entre les valeurs de l'attribut FRC et les instances sont présentées dans le Tableau 20.

Une fois que nous avons déterminé les instances pour tous les types de classement définis dans MultiNet, nous avons défini des mesures d'écart sémantique entre tous les concepts des deux jeux de données.

L'écart sémantique entre deux concepts correspond à la valeur moyenne des distances sémantiques calculées entre un concept de la BDCARTO et les concepts composant un type de classement de MultiNet. Signalons que la distance sémantique employée est la distance de [Wu et Palmer, 1994] que nous avons présentée au chapitre A.

Valeur de l'attribut FRC de MultiNet	Instances
0 : autoroute, route prioritaire	autoroutes
1 : route prioritaire moins importante que l'autoroute	routes nationales, routes européennes
2 : autre route prioritaire	routes nationales, routes européennes départementales
3 : route secondaire	routes nationales et départementales
4 : liaison locale	routes départementales
5 : route locale d'importance élevée	routes départementales, rues, places
6 : route locale	routes départementales, rues, avenues, boulevards, places
7 : route locale d'importance moins élevée	rues, places, impasses
8 : autre route	petites rues, impasses, escaliers, sentiers

Tableau 20. Correspondances entre le type de classement et les instances MultiNet

Exemple de calcul de mesure d'écart sémantique

Considérons l'arc A_1 appartenant à MultiNet et l'arc A_2 appartenant à BDCARTO, pour lesquels nous souhaitons déterminer l'écart sémantique entre les types de classement. Supposons que l'arc A_1 soit de type 5 (route locale d'importance élevée) et que l'arc A_2 soit de type « départementale ». Nous savons, en analysant les données d'une manière interactive, que dans le type 5 nous avons des instances représentant des routes départementales, des rues et des places.

La distance sémantique a été appliquée sur les concepts définis dans la taxonomie présentée ci-dessus. Ainsi dans notre cas, trois distances sémantiques relatives aux trois instances se trouvant dans le type de classement 5 sont calculées : $d_{S1} = d_S$ (« route départementale »-« route départementale ») = 0, $d_{S2} = d_S$ (« route départementale »-« rue ») = 0.5 et $d_{S3} = d_S$ (« route départementale »-« place ») = 0,66. Enfin, la mesure illustrant l'écart sémantique correspond à la valeur moyenne des trois distances sémantiques obtenues : $d_{Sfinale} = (d_{S1} + d_{S2} + d_{S3})/3$.

L'écart sémantique obtenu pour cet exemple est de 0,39.

Précisons que pour un arc de la BDCARTO classé dans « autres routes » d'après l'attribut qui renseigne sur le type de classement administratif, nous avons utilisé dans le calcul de la distance sémantique l'attribut vocation. Ce dernier matérialise une hiérarchisation du réseau routier basée sur l'importance de la route pour le trafic routier.

Une autre possibilité pour mesurer l'écart sémantique entre les concepts aurait été de comparer directement l'attribut vocation de la BDCARTO avec l'attribut FRC de MultiNet. Même si les deux attributs sont plus proches au niveau de l'information fournie, le niveau de détail reste assez faible.

Par conséquent, nous avons considéré que la comparaison de l'attribut de la BDCARTO qui renseigne sur le type de classement administratif, de l'attribut vocation et de l'attribut

FRC de MultiNet est plus exploitable, malgré le fait qu'il ne s'agisse pas du même niveau de détail. Un autre avantage de cette comparaison est que nous pouvons nous appuyer sur la taxonomie réalisée au laboratoire COGIT pour le réseau routier.

Etant donnée la différence de classification de l'information sémantique, le seuil S est fixé à 0,7, afin de ne pas éliminer des candidats potentiels.

Critère nom d'objet

Les deux jeux de données possèdent un attribut renseignant sur le numéro de route. Bien que seulement les routes principales, c'est-à-dire les autoroutes, les routes nationales et départementales possèdent un numéro de route, nous utilisons cette information pour définir le critère nom d'objet. Afin de gérer cette incomplétude présente au niveau de cet attribut, nous avons défini un critère discret.

Nous rappelons dans le Tableau 21 la représentation des connaissances pour le critère nom d'objet.

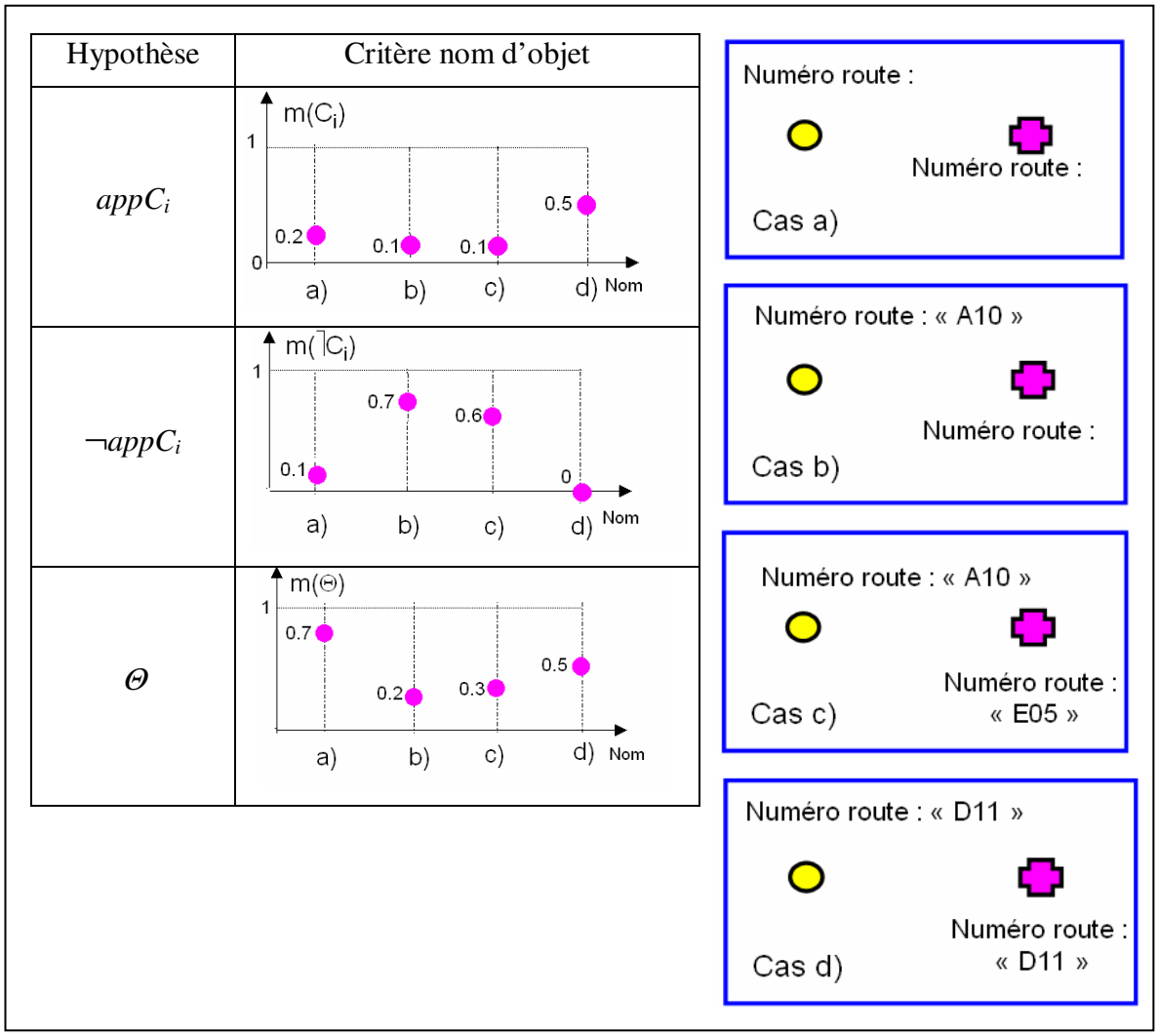


Tableau 21. Représentation des connaissances pour le critère nom d'objet (à gauche) et les quatre cas définis (à droite)

Rappelons que les cas c) et d) correspondent à la situation où les deux arcs à comparer possèdent un nom d'objet. Nous ne pouvons pas dire, par exemple, que la route ayant le numéro « D313 » est l'homologue de la route ayant le numéro « D331 » même si l'écart entre les deux chaînes de caractères est faible. Ainsi, si les numéros sont identiques, nous sommes dans le cas d), c'est-à-dire que nous croyons que les deux arcs sont homologues, et si les numéros sont différents, nous sommes dans le cas c), c'est-à-dire que nous croyons que les deux arcs ne sont pas homologues.

Comme nous l'avons dit précédemment, les autoroutes ont deux notations différentes dans les deux jeux de données : la numérotation française dans la BDCARTO et la numérotation européenne dans MultiNet. En incorporant le tableau de correspondance entre les numérotations dans notre algorithme, nous avons attribué une distance égale à 0 aux autoroutes correspondantes. Par exemple, la distance entre l'autoroute « A131 » et l'autoroute correspondante « E5 » est égale à 0.

Critère voisinage

Afin d'avoir une vue globale et ainsi de corriger d'éventuelles erreurs d'appariement, nous avons défini le critère de voisinage pour les objets appariés et pour les objets non appariés. Ce dernier est instancié à partir des résultats obtenus après une première passe du même processus avec les critères que nous venons de présenter. Le critère de voisinage est basé ici uniquement sur les relations de voisinage topologique. L'idée de base est qu'un arc a plus de chance d'être apparié si ses voisins le sont ou un arc a plus de chance de ne pas être apparié si ses voisins ne le sont pas.

Les résultats que nous présentons sont ceux obtenus avec le critère illustré dans le Tableau 22 et le Tableau 23.

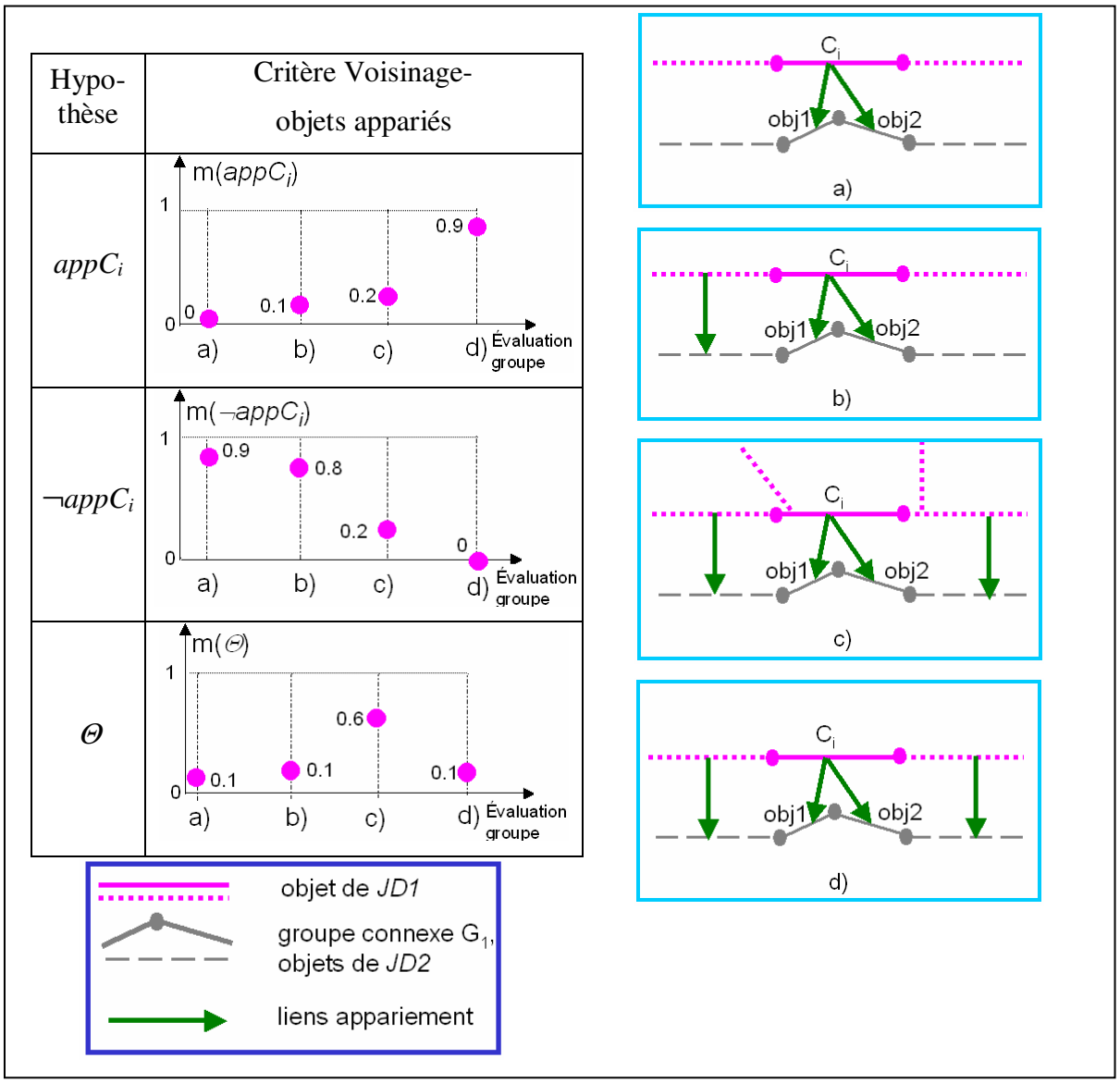


Tableau 22. Représentation des connaissances pour le critère voisinage-objets appariés (à gauche) pour les quatre cas définis à droite (voir la partie C.2.4)

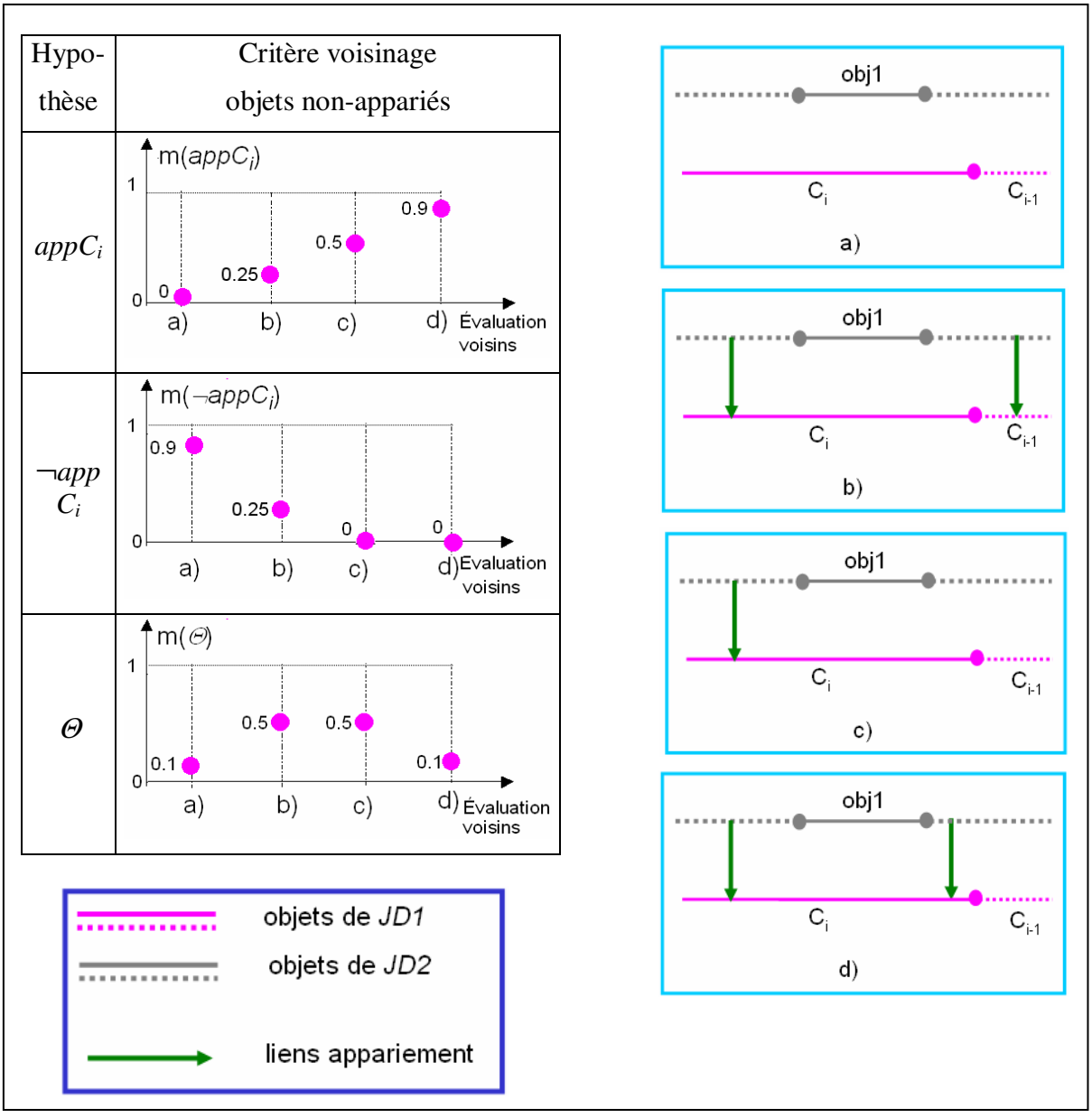


Tableau 23. Représentation des connaissances pour le critère voisinage-objets non-appariés (à gauche) pour les quatre cas définis à droite (voir la partie C.2.4).

D.3.2.2 Résultats des expérimentations : exemples

Nous montrons dans cette partie quelques résultats d'appariement que nous avons obtenus en testant notre approche sur le réseau routier.

Sur les figures suivantes, le jeu de données issu de la BDCARTO est représenté en haut à gauche et le jeu de données issu de MultiNet est représenté en haut à droite. Les deux jeux de données superposés, ainsi que les liens d'appariement sont présentés en bas de chaque figure.

Nous présentons en Figure 103 un exemple typique d'appariement de réseaux. Dans ce cas l'arc A_2 de la BDCARTO est bien apparié avec les arcs E1, F1, G1, H1, et I1. Nous remarquons que notre processus d'appariement n'apparie pas systématiquement à l'objet le plus proche. En effet, étant donné que pour apparier deux jeux de données, nous cherchons pour chaque arc du jeu de données plus détaillé (MultiNet) des arcs homologues dans le jeu de données moins détaillé (BDCARTO), les arcs A1, B1, C1 ou D1 auraient pu choisir

comme arc homologue le candidat le plus proche, c'est-à-dire le candidat A2. Ce cas de figure ne s'est pas produit, car lors de la fusion des critères d'appariement, l'hypothèse NA (non-apparié) a été choisie pour les arcs mentionnés ci-dessus. L'appariement est ici très satisfaisant.

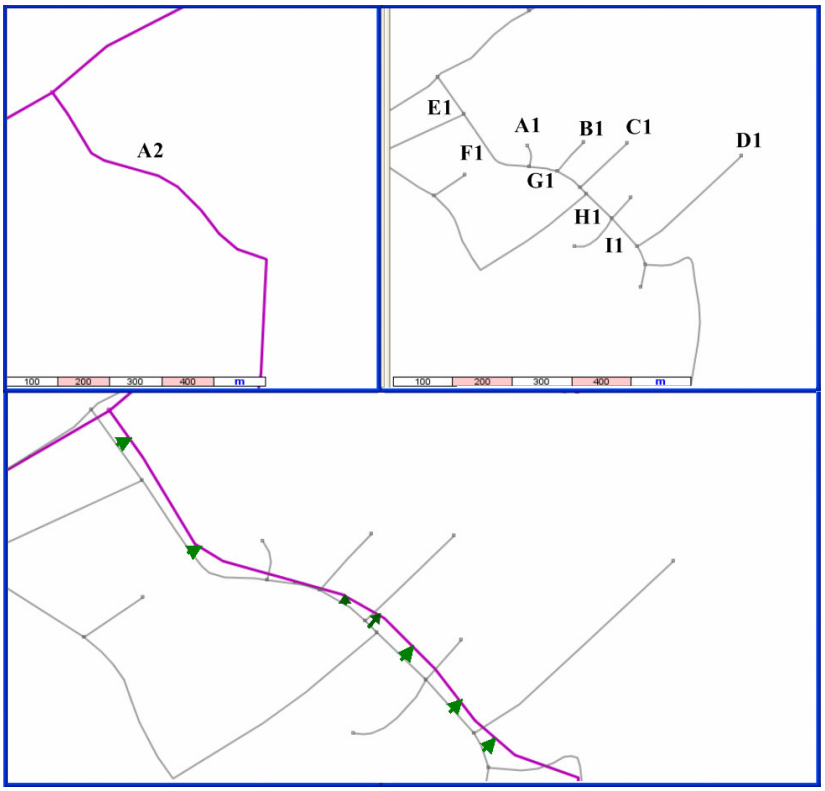


Figure 103. Exemple typique d'appariement d'arcs

Nous illustrons en Figure 104 et Figure 105 l'intérêt d'utiliser l'information relative aux numéros de route à travers le critère nom d'objet. Lorsque le critère nom d'objet n'est pas utilisé dans le processus d'appariement (voir Figure 104), l'arc A2 de la BDCARTO ayant le numéro de route « D81 » est à la fois bien apparié avec les arcs homologues A1, B1, et C1 ayant le numéro de route « D81 » et apparié à tort avec les tronçons de route D1 et E1. Ces derniers composent dans la réalité une place publique et ils ne possèdent pas de numéro de route. En revanche, lorsque nous utilisons le critère nom d'objet cette erreur est corrigée. Cela est possible grâce au fait que nous considérons qu'un arc ayant un numéro de route a très peu de chance d'être apparié avec un arc ne possédant pas de numéro. Cette croyance est justifiée par les spécifications des deux jeux de données qui mentionnent que seules les routes importantes ont un numéro de route.

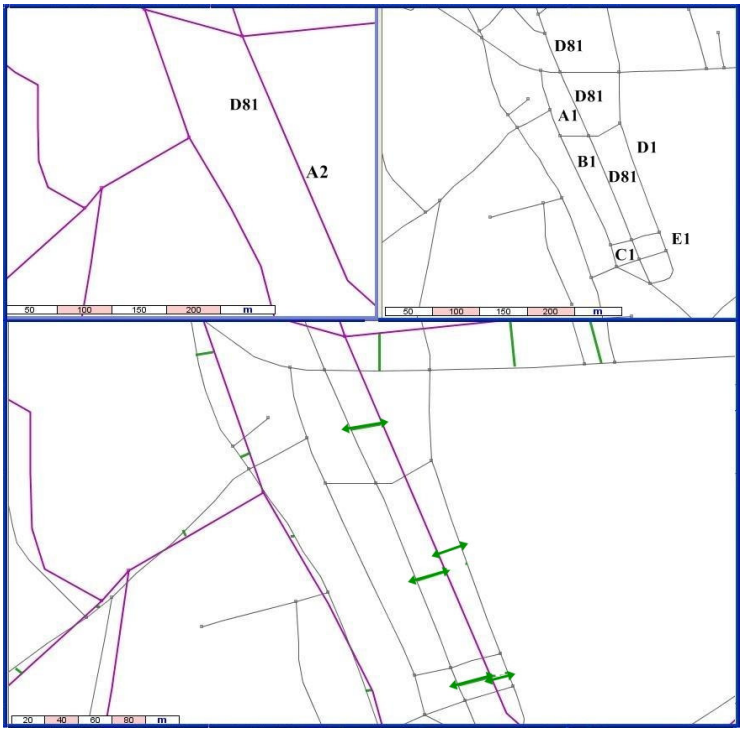


Figure 104. Résultats d'appariement sans le critère nom d'objet

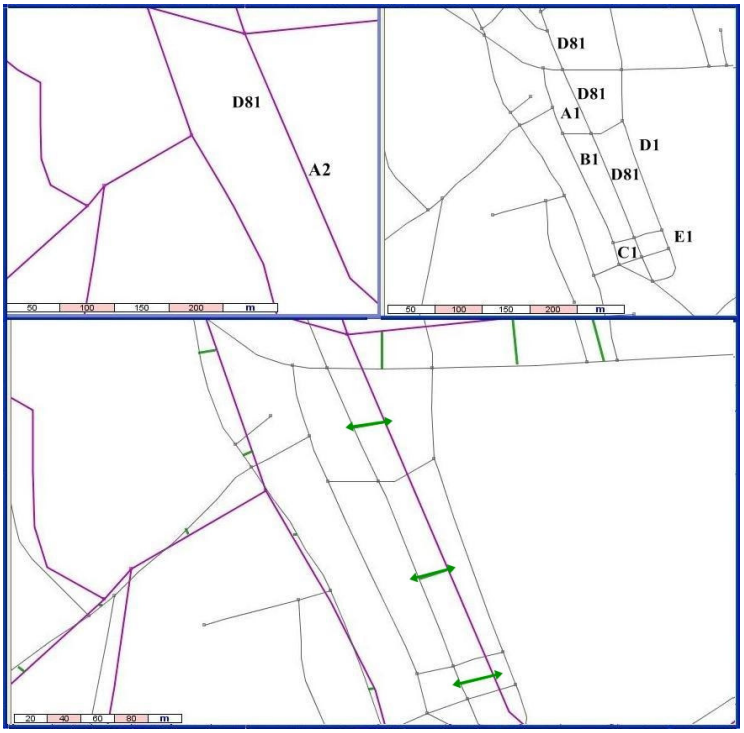


Figure 105. Résultats d'appariement avec le critère nom d'objet

Nous illustrons à travers les figures suivantes l'intérêt d'analyser le contexte géographique des arcs appariés, c'est-à-dire le voisinage, afin d'avoir une analyse globale et d'améliorer les résultats appariement. L'amélioration des résultats peut consister d'une part à apparier des arcs qui n'ont pas été appariés par erreur, mais aussi à corriger des appariements erronés, c'est-à-dire des arcs qui ont été appariés à tort.

Sur la Figure 106 en bas, nous montrons en vert les arcs de MultiNet non appariés après la première passe du processus, mais appariés après la deuxième passe du processus, c'est-à-dire avec le critère de voisinage. Ainsi, une amélioration des résultats est obtenue grâce à la deuxième passe du processus, en appariant des arcs qui n'ont pas été appariés à tort lors de la première passe du processus. Il s'agit de 43 arcs représentant approximativement 1% de la longueur du réseau MultiNet.

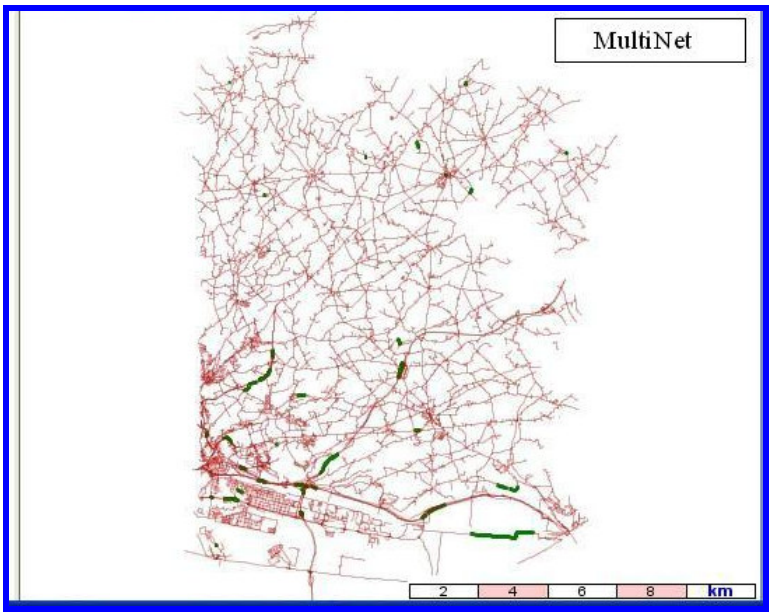


Figure 106. Arcs de MultiNet appariés en utilisant le critère voisinage

La Figure 107 illustre un exemple d'arc qui est apparié après l'introduction du critère du voisinage et non-apparié sinon. Il s'agit de l'arc B1 de MultiNet qui est apparié avec l'arc A2 de la BDCARTO. L'arc B_1 n'a pas été apparié après la première passe du processus en raison de la distance de Hausdorff qui est élevée. Précisons que l'écart entre l'hypothèse non-apparié et l'hypothèse $appA_2$, c'est-à-dire que l'arc B_1 est apparié avec l'arc A2 est très faible. L'appariement est possible après la deuxième passe du processus grâce au fait que les voisins de l'arc B1, c'est-à-dire A1 et C1, sont appariés et que le critère de voisinage a plus de poids dans ce cas-là.

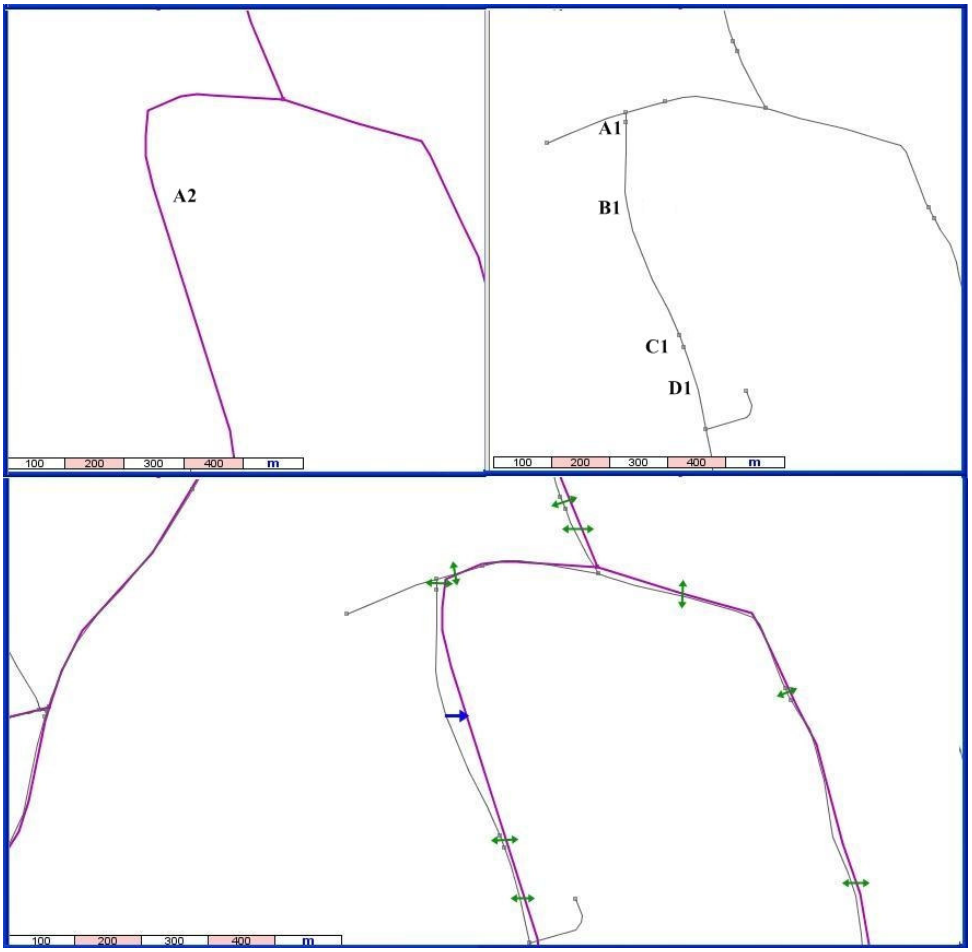


Figure 107. Exemple d'arc de MultiNet apparié après la deuxième passe du processus

Nous illustrons à travers la Figure 108 le deuxième intérêt d'utiliser le critère de voisinage : corriger des appariements erronés. Ainsi, sans le critère de voisinage, les arcs A1, B1, C1, D1 de MultiNet sont appariés à tort avec les arcs A2 et B2 de la BDCARTO. Cette erreur d'appariement est due au fait que les arcs de MultiNet en question sont parallèles aux arcs de BDCARTO, sont de la même importance et ne possèdent pas de numéro de route. De plus nous remarquons qu'en ce qui concerne la localisation des arcs, ils sont proches des arcs de la BDCARTO.

En revanche, lorsque le critère de voisinage est utilisé, cette erreur est corrigée. Nous remarquons en Figure 109 que les arcs A2 et B2 sont appariés à leurs vrais homologues. Ce résultat est possible grâce au fait que les arcs A1, B1, C1, D1 et E1 forment un groupe connexe et que les voisins du groupe n'ont pas été appariés à la première passe du processus aux arcs A2 et B2 ou à leurs voisins.

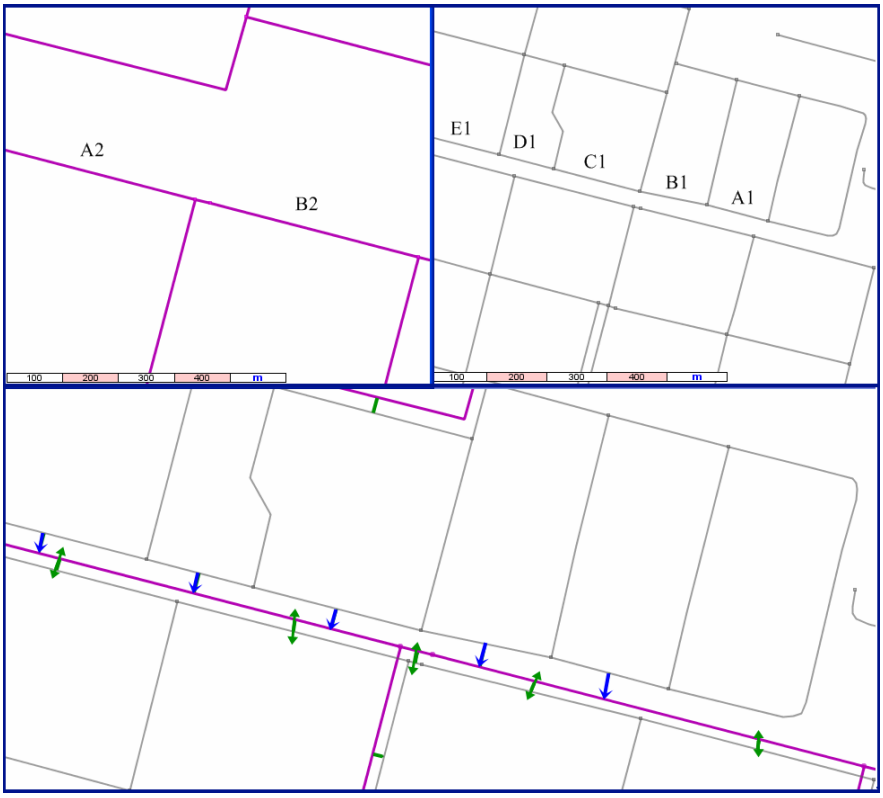


Figure 108. Résultat d'appariement d'arcs obtenu sans le critère voisinage

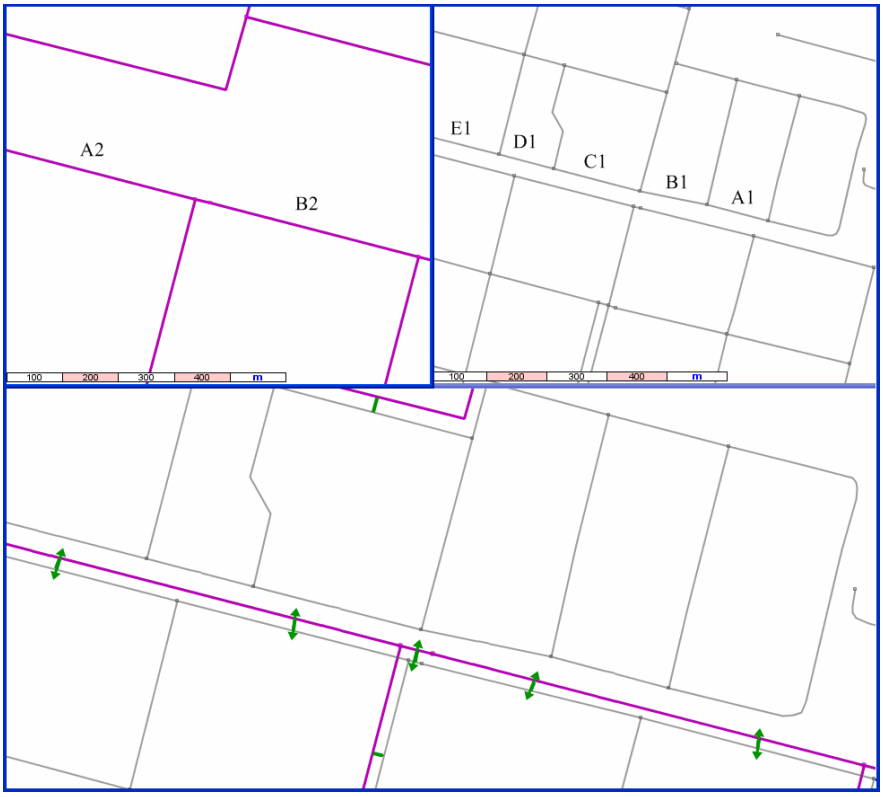


Figure 109. Résultat d'appariement d'arcs obtenu avec le critère voisinage

Nous montrons en Figure 110 un exemple qui illustre l'appariement d'un rond-point. Nous remarquons que notre processus d'appariement ne gère pas bien les ronds-points. Rappelons

que la représentation des ronds-points dans les deux jeux de données est très différente. Dans MultiNet les ronds-points sont représentés par un ensemble d'arcs, tandis que dans la BDCARTO les ronds-points sont représentés par un point. Notre processus apparie les arcs composant un rond-point avec les extrémités des arcs de la BDCARTO. Les propriétés des arcs composant un rond-point ne sont pas très riches. Par exemple, aucune information n'existe qui pourrait spécifier que l'arc fait partie d'un rond-point. Le numéro de route n'est rempli que si le rond-point fait partie d'une route nationale ou départementale. De plus, l'orientation des arcs n'est pas discriminante, c'est-à-dire que l'écart d'orientation diffère d'un rond-point à un autre. Par conséquent, le seul critère qui est discriminant est le critère d'écart de position, puisqu'il a un poids plus important que les autres. Cela signifie qu'en général un arc d'un rond-point est apparié à l'arc le plus proche. Donc, il est possible que des arcs composant un rond-point soient appariés avec le même arc de la BDCARTO, ce qui ne sera pas correct.

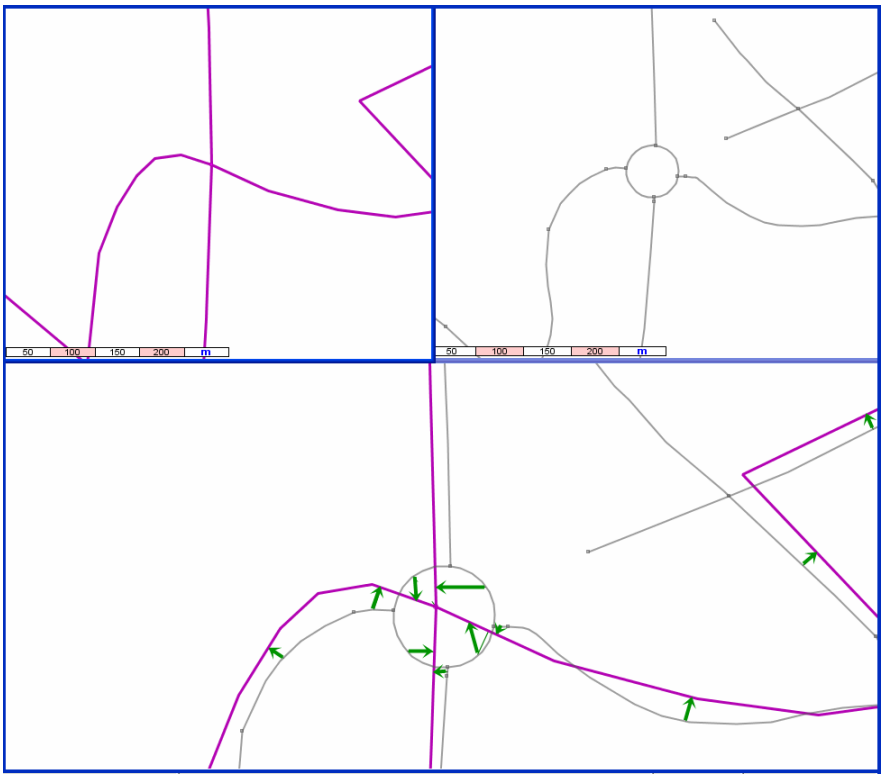


Figure 110. Exemple de résultat d'appariement obtenu dans le cas d'un rond-point

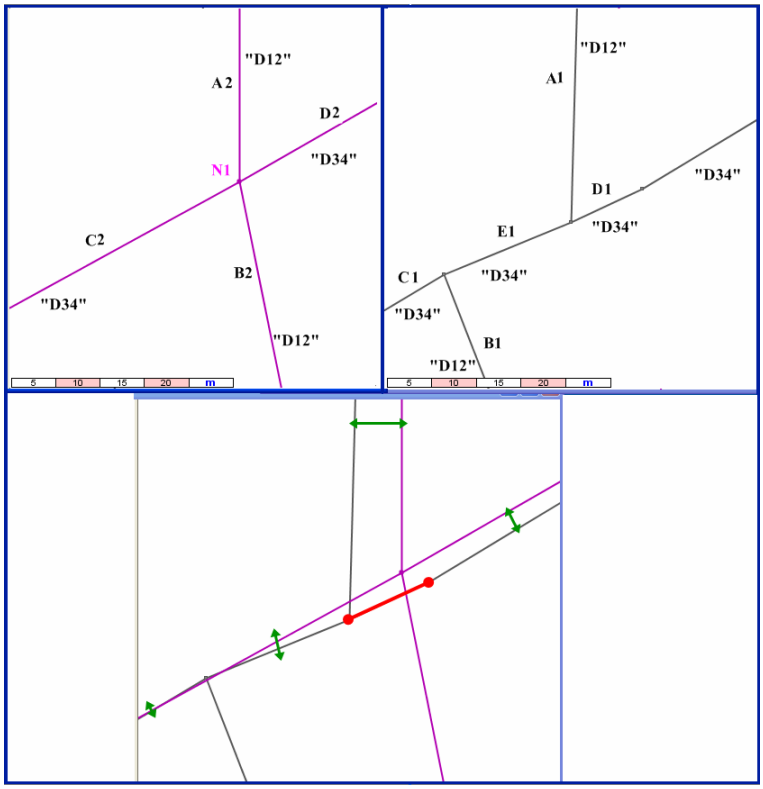


Figure 111. Exemple de conflit total signalé par notre processus

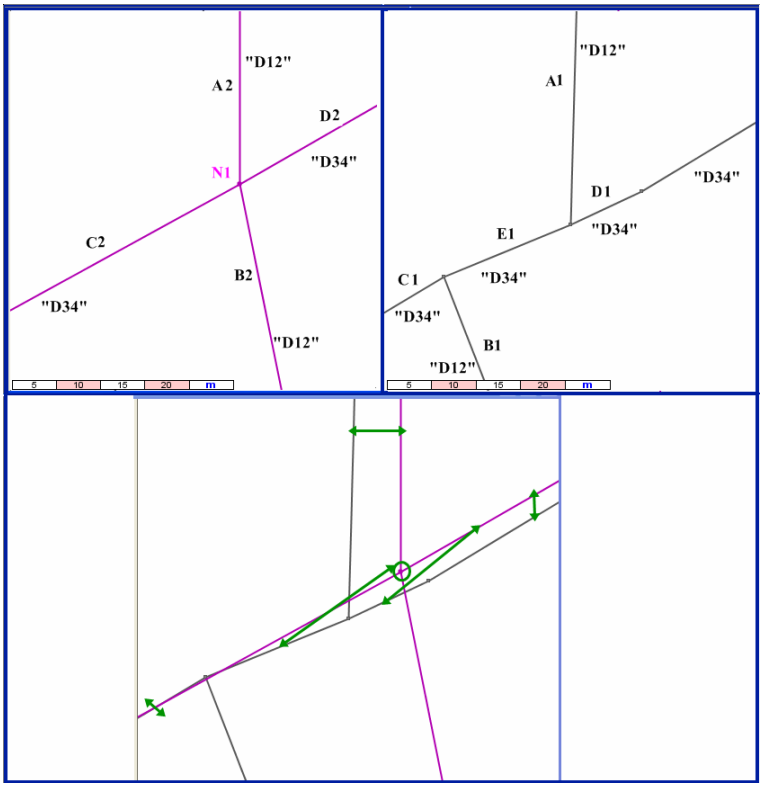


Figure 112. Résultats d'appariement obtenus avec l'approche de [Mustière et Devogele, 2008]

De la même manière que pour les expérimentations réalisées sur les points remarquables du relief, le conflit total est signalé à la fin du processus. Nous donnons à titre d'illustration sur la Figure 111 un exemple de conflit total. Il s'agit de l'arc D1 pour lequel aucune décision n'est prise. La raison du conflit est que l'arc D1, qui a le numéro de route « D34 », a deux candidats dans la BDCARTO parmi d'autres qui ont le même numéro de route. La manière dont nous avons modélisé le critère numéro de route ne devrait pas engendrer de conflit, parce que lorsqu'il s'agit d'un même numéro de route, la masse de croyance est partagée entre l'hypothèse $appC_i$ et l'ignorance. Par contre dans ce cas, les autres critères soutiennent les deux candidats d'une manière sensiblement similaire. Par conséquent, la fusion des critères conduit à un conflit total.

Une autre erreur similaire aux cas précédents, liée à la différence de niveau de détail, est détectée dans notre processus (voir la Figure 111). Il s'agit de l'arc E1 de MultiNet qui est apparié à l'arc C2 de la BDCARTO.

Dans ces configurations, où seule une analyse fine de la topologie nous permet d'apparier correctement, un processus comme celui de [Mustière et Devogele, 2008] est plus efficace, car il est spécialement conçu pour étudier la topologie des réseaux à des niveaux de détail différents (voir Figure 110).

D.3.2.3 Evaluation quantitative et discussion

Nous présentons dans cette partie une évaluation quantitative des résultats d'appariement obtenus pour les réseaux routiers. Signalons que l'évaluation a été réalisée d'une manière interactive. Cette évaluation est exprimée en résultats sur le jeu de données plus détaillé, MultiNet. Ce choix a été fait en raison de la cardinalité des liens d'appariement obtenus, n : 1, c'est-à-dire que n arcs de MultiNet sont appariés avec un arc de la BDCARTO. La cardinalité n : 1 est obtenue après le regroupement des liens 1 : 1 définis par notre processus d'appariement. Les résultats obtenus sont exposés dans le Tableau 24. Nous évaluons les résultats d'une part en nombre d'arcs appariés et en pourcentage de la longueur totale du réseau MultiNet (voir les colonnes 2 et 3), et d'autre part en termes de précision et de rappel (voir les colonnes 4 et 5).

Le nombre d'arcs pour lesquels le processus d'appariement nous a signalé un conflit (voir la quatrième ligne), c'est-à-dire quand la masse associée à l'ensemble vide est supérieure à 0,9, est de onze. Les onze arcs représentent moins de 0,1% de la longueur totale du réseau MultiNet. Le conflit est justifié, les onze cas étant des configurations complexes.

Nous remarquons que 59% de la longueur totale du réseau MultiNet est apparié avec une précision de 96,6%. Cela signifie que parmi tous les appariements trouvés (liens de cardinalité 1 : 1), 3,4% sont faux. Nous avons obtenu un rappel de 95%, ce qui signifie que parmi tous les appariements que nous aurions dû trouver, nous en avons trouvé 95%. Autrement dit, il manque 5% des appariements.

Pour 40% de la longueur du réseau MultiNet, nous n'avons pas trouvé de correspondant dans le réseau de la BDCARTO. La précision obtenue est de 94%, ce qui signifie que parmi tous les appariements 1 : 0 trouvés, 6% sont faux. Un rappel de 93,7% a été obtenu pour les arcs non-appariés, c'est-à-dire que nous avons trouvé moins d'arcs sans homologue qu'on aurait dû.

	Nombre d'objets	Longueur du réseau MultiNet	Précision	Rappel
Appariés	6093 arcs	59%	96,6%	95%
Non-appariés	6621 arcs	40%	94%	93,7%
Conflit	11 arcs	<0,1%	-	-
Total	12725 arcs	100%	95,3%	94,3%

Tableau 24. Evaluation des résultats d'appariement pour les réseaux

En ce qui concerne l'évaluation qualitative des résultats d'appariement obtenus, nous avons remarqué que les erreurs rencontrées sont liées à la spécificité des données, aux connaissances utilisées et à la faiblesse du processus dans certains cas :

1. le critère sémantique est un critère important qui améliore la qualité des résultats lorsque les routes sont importantes : autoroutes, nationales et départementales. L'intérêt de ce critère est qu'il ne permet pas d'apparier par exemple une route départementale avec une rue ou un chemin. Par contre, pour les routes moins importantes, c'est-à-dire les routes qui ne sont ni autoroutes, ni nationales, ni départementales, ce critère n'est plus discriminant, et il est la source d'erreurs d'appariement. A cause de la différence de classification, les distances sémantiques que nous avons déterminées n'illustrent pas d'une manière efficace les différences entre les routes moins importantes. Par exemple, les boulevards et les impasses ont des écarts sémantiques très faibles, bien qu'il s'agisse de deux entités du monde réel très différentes. Dans de tels cas, une solution est de modifier sensiblement la modélisation du critère sémantique en augmentant la masse de croyance attribuée à l'ignorance,
2. d'autres erreurs apparaissent en zone urbaine où les arcs possèdent des attributs semblables, ce qui engendre des écarts très faibles, par conséquent des sur-appariements, c'est-à-dire des arcs qui sont appariés mais qui ne devraient pas l'être. Par exemple, un arc de MultiNet peut avoir plusieurs candidats BDCARTO très proches, ayant la même sémantique et n'ayant pas de numéro de route. Dans ce cas, le seul critère qui va les partager est le critère d'écart de position. L'arc choisi sera celui qui est le plus proche. Dans certains cas, choisir l'arc le plus proche engendre un appariement faux. Le conflit signalé par notre processus est dû dans certains cas au décalage planimétrique entre les réseaux lorsque les arcs possèdent un numéro de route (quatre arcs). Un tel exemple est illustré en Figure 111. Dans d'autres cas le conflit apparaît lorsque les arcs font partie d'un rond-point (sept arcs),
3. Les erreurs dues à la faiblesse du processus concernent les ronds-points, les carrefours complexes, les bretelles, les carrefours dénivelés, les pattes d'oies. Afin d'améliorer notre processus dans ces cas, nous devons introduire l'appariement d'un arc à un nœud et mieux prendre en compte la structure des réseaux. Nous reviendrons plus en détail dans le chapitre E sur les améliorations que nous pouvons apporter au processus.

Comme nous l'avons dit lors de la description de notre processus au chapitre C, chaque lien d'appariement est auto-évalué par le processus comme étant sûr ou incertain en fonction de la différence entre la valeur du premier maximum choisi et le deuxième. Rappelons que si la différence est supérieure ou égale à 0,5, nous considérons le lien comme sûr, dans le cas contraire il est considéré comme incertain. Le nombre de liens sûrs représente 78% des liens trouvés.

Précisons que nous avons constaté une faible corrélation entre les erreurs d'appariement faites par notre processus et l'auto-évaluation calculée à partir des probabilités pignistiques. Cette faible corrélation est spécifique surtout aux appariements jugés incertains. Certains appariements jugés incertains sont justes.

Nous présentons sur la Figure 113 l'histogramme de l'auto-évaluation des liens d'appariement. Nous avons présenté en abscisse les intervalles de valeurs représentant la différence entre le premier et le deuxième maximum et en ordonnée le nombre d'occurrences. Par exemple nous avons presque 2500 liens d'appariement pour lesquels la différence entre le premier et le deuxième maximum est comprise dans l'intervalle [0,9 - 1], ce qui signifie que les candidats choisis se démarquent fortement. Pour presque 500 liens d'appariement, la différence entre le premier et le deuxième maximum est très faible, puisqu'elle est comprise dans l'intervalle]0 – 0,1]. Cela signifie qu'aucun candidat à l'appariement ne se démarque.

Rappelons que la différence entre le premier et le deuxième maximum est égale à 0 lorsqu'un conflit total apparaît.

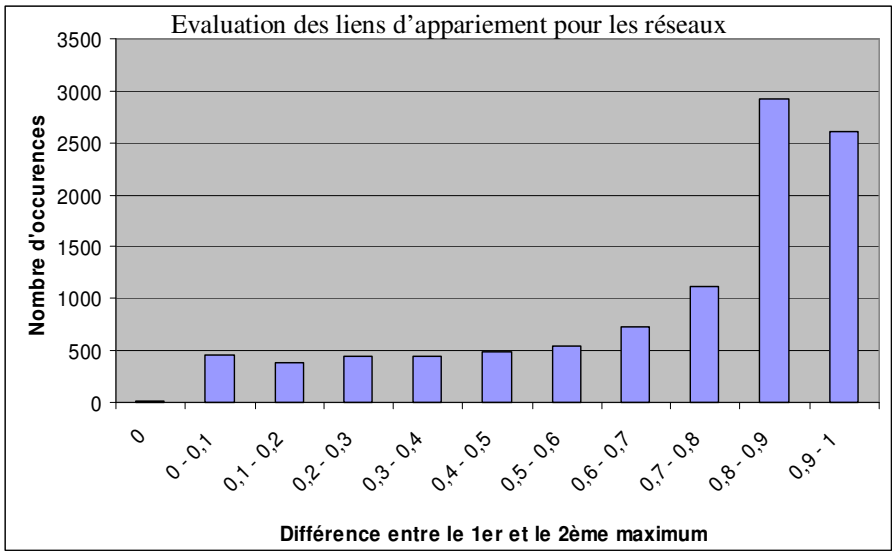


Figure 113. Histogramme de l'auto-évaluation des liens d'appariement pour les réseaux routiers

Précisons que l'auto-évaluation n'a pas été vérifiée interactivement sur toute la zone. Sur un extrait de la zone, nous avons pu remarquer qu'en général les liens d'appariement qui ont une évaluation comprise entre 0,7 et 1 sont vraiment des liens justes. Par contre, il existe de nombreux cas d'appariement pour lesquels l'auto-évaluation est très faible, même dans l'intervalle]0 - 0,2], mais qui sont justes. Cette faible évaluation est due dans certains cas au nombre élevé des candidats à l'appariement. L'auto-évaluation est donc globalement plutôt pessimiste qu'optimiste.

Comparaison avec une autre approche

Nous avons réalisé une étude comparative entre les résultats d'appariement obtenus avec notre approche et ceux obtenus avec l'approche de [Mustière et Devogele, 2008] spécialement conçue pour apparier des réseaux ayant des niveaux de détail différents, mais ne s'appuyant que sur la topologie et la géométrie.

Nous avons remarqué que 87,7% des arcs de MultiNet ont été appariés (ou non-appariés) exactement de la même manière par les deux processus d'appariement. 8,8% des arcs de MultiNet n'ont été appariés qu'avec notre processus, c'est-à-dire que pour ces arcs le processus de [Mustière et Devogele, 2008] n'a pas trouvé d'arc homologue et 3,2% des arcs de MultiNet n'ont été appariés qu'avec le processus de [Mustière et Devogele, 2008], c'est-à-dire que pour ces arcs notre processus a choisi l'hypothèse NA (non-apparié). 0,3% des arcs de MultiNet ont été appariés différemment avec les deux processus, c'est-à-dire que pour un arc de MultiNet un processus a choisi un homologue tandis que l'autre processus a choisi un autre homologue. Parmi ce pourcentage d'arcs appariés différemment, 0,1% des arcs représentent des appariements pour lesquels le processus de [Mustière et Devogele, 2008] a apparié un arc de MultiNet à un nœud de la BDCARTO, tandis que notre processus a apparié le même arc de MultiNet avec un arc de la BDCARTO.

Un exemple d'appariement différent, c'est-à-dire qu'un même arc de MultiNet est apparié avec les deux processus à deux arcs différents de la BDCARTO, est illustré en Figure 114. L'arc B1 de MultiNet est apparié avec notre processus à l'arc D2 de la BDCARTO, tandis que le même arc B1 est apparié avec l'autre processus avec l'arc B2 de la BDCARTO. Dans ce cas, l'appariement de [Mustière et Devogele, 2008] a raison. Signalons que parmi les 0,1 % d'appariement différents, notre processus a raison dans 40% des cas et tort dans les 60% des cas restant.

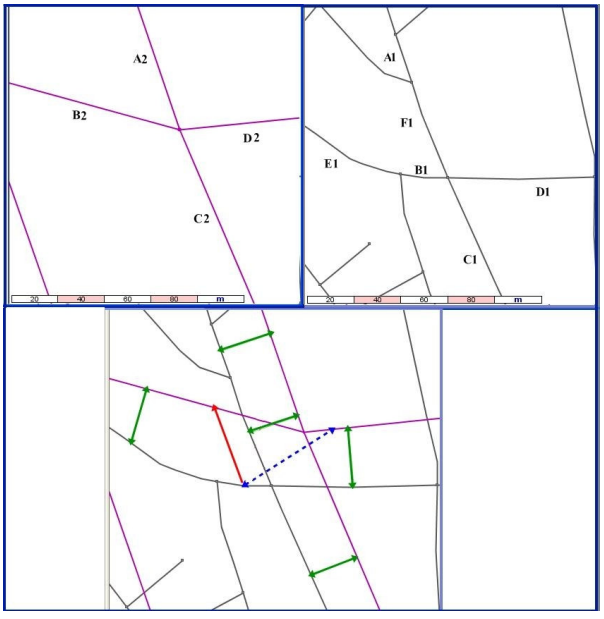


Figure 114. Exemple de résultats différents obtenus avec notre approche et avec l'approche de [Mustière et Devogèle, 2008]

Globalement, les appariements différents concernent surtout les carrefours décalés pour lesquels le décalage est représenté dans le jeu de données plus détaillé à travers un arc et n'est pas du tout représenté dans le jeu de données moins détaillé.

Dans le Tableau 25 nous présentons une étude comparative entre les résultats obtenus avec notre approche et avec l'approche de [Mustière et Devogele, 2008]. Nous remarquons que pour les objets appariés, notre processus fait plus d'erreur que le processus de [Mustière et Devogele, 2008], ayant une précision de 96,6% par rapport à une précision de 98% obtenue avec l'autre processus. Nous avons obtenu un rappel supérieur (95%) au rappel de 90% obtenu avec l'approche de [Mustière et Devogele, 2008]. Cela signifie que nous avons trouvé plus d'appariements que l'autre approche.

En ce qui concerne les objets non-appariés, nous avons remarqué que les résultats pour la précision et le rappel sont inversés, c'est-à-dire que nous avons trouvé une précision (98%) supérieure à celle de l'approche de [Mustière et Devogele, 2008] qui est de 85% et un rappel inférieur au rappel obtenu avec l'autre processus.

En conclusion de cette comparaison, nous pouvons avancer que notre processus est plutôt optimiste, c'est-à-dire qu'il apparie plus d'arcs, tandis que le processus de [Mustière et Devogele, 2008] est plutôt pessimiste. Les erreurs rencontrées dans notre processus concernent donc les ronds-points, les carrefours complexes, les bretelles, les échangeurs, tandis que les erreurs rencontrées dans le processus de [Mustière et Devogele, 2008] concernent les carrefours complexes (échangeurs, bretelles), les routes à chaussées séparées, et les cas où les topologies des réseaux sont imparfaites ou incohérentes.

		Nombre d'objets	Longueur du réseau MultiNet	Précision	Rappel
Notre approche	Appariés	6093 arcs	59%	96,6%	95%
	Non-appariés	6632 arcs	41%	95%	93,7%
	Conflit total	11 arcs	<0,1%	-	-
[Mustière et Devogèle, 2008]	Appariés	5304 arcs	54%	98%	90%
	Non-appariés	7421 arcs	46%	85%	97,5%

Tableau 25. Evaluation des résultats d'appariement pour les réseaux en utilisant notre approche et l'approche de [Mustière et Devogele, 2008]

Sensibilité aux seuils

Les études que nous avons mises en oeuvre pour étudier la sensibilité de notre processus aux seuils sont de la même nature que celles réalisées pour les points remarquables du relief. Nous avons réalisé deux études relatives au critère d'écart de position et au critère sémantique. Les autres critères étant discrets, aucun seuil n'a été défini. L'étude de stabilité consiste à faire varier les seuils pour un seul critère à la fois.

Mentionnons que les résultats sont exprimés en fonction du pourcentage d'arcs appariés et non-appariés selon la valeur du seuil utilisé, mais que nous n'avons pas évalué les résultats.

Etude de la sensibilité du processus aux seuils de sélection des candidats

De la même manière que pour les points remarquables du relief, le processus appliqué aux réseaux routiers dépend du seuil de sélection des candidats.

Etude de la sensibilité du processus aux seuils choisis pour les distances

Afin d'étudier la sensibilité au seuil d'écart de position, nous avons fait varier le seuil T_2 défini dans le critère d'écart de position, en gardant les mêmes seuils pour les autres critères. Notons que le seuil de référence est de T_2 = 60m. Etant donné que le seuil de sélection a été fixé à 100 m, la plage de valeurs possible pour le seuil T_2 est [0 m – 100 m]. Signalons que le seuil T_1 fixé dans les courbes est automatiquement changé, étant égal à $T_2/2$. Nous avons choisi un pas de 10 m. Les résultats obtenus sont illustrés sur la Figure 115.

Nous avons représenté en abscisse le seuil T_2 en mètres, et en ordonnée le pourcentage d'arcs appariés ou non-appariés. Comme nous pouvons le remarquer, pour un seuil T_2 choisi dans l'intervalle [40 m – 70m], le processus est assez peu sensible, des pourcentages sensiblement identiques étant obtenus. Par contre, lorsque les seuils sont choisis autour de 10 m ou de 20 m, la sensibilité est grande. Plus le seuil est petit, plus nous sommes confrontés à des sous-appariements, et plus le seuil est grand, plus nous sommes confrontés à des sur-appariements. Par conséquent, il nous semble pertinent de choisir un seuil compris entre 40m et 60m.

En conclusion, nous pouvons affirmer que choisir le seuil T_2 de l'ordre de grandeur du double de la précision de la base nous semble une solution pertinente.

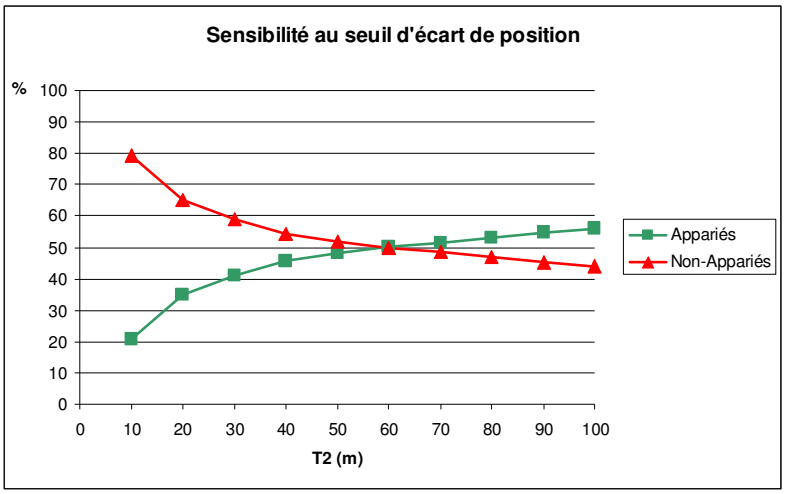


Figure 115. Sensibilité au seuil d'écart de position

Etude de la sensibilité du processus au seuil sémantique

Comme nous l'avons présenté dans le chapitre C, pour le critère sémantique nous avons défini un seul seuil S. Dans nos expérimentations réalisées pour les réseaux routiers, le seuil S de référence a été fixé à 0,7. Nous présentons sur la Figure 116 les résultats que nous avons obtenus en faisant varier le seuil S avec un pas de 0,1. En abscisse, nous avons représenté les seuils possibles entre 0 et 1, et en ordonnée le pourcentage obtenu pour les arcs appariés (courbe verte) et pour les arcs non-appariés (courbe rouge). Nous remarquons que le processus est moins sensible au seuil défini pour le critère sémantique que pour le critère d'écart de position, le pourcentage d'arcs appariés ou non-appariés variant de 10% au plus entre les seuils fixés. Si le seuil est faible nous remarquons que le nombre de sous-appariements est important et qu'à partir d'un seuil de 0,7, les résultats ne changent plus et que nous n'avons pas de cas de sur-appariement. En conclusion, un seuil sémantique de l'ordre de 0,7-0,8 nous semble pertinent pour nos données.

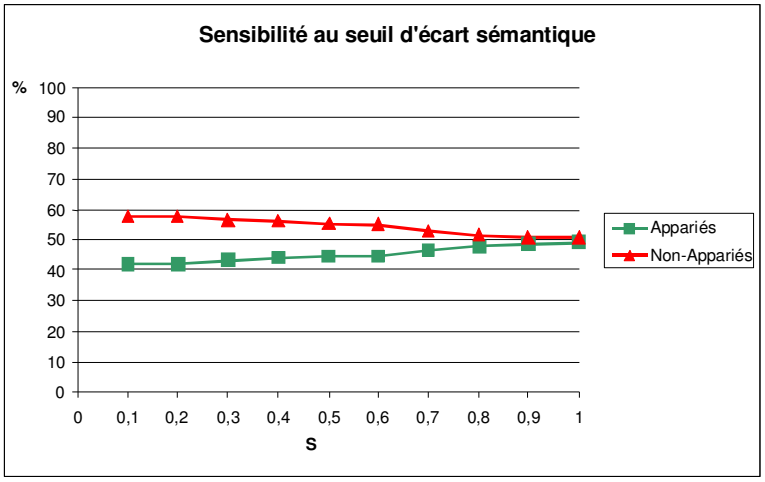


Figure 116. Sensibilité au seuil d'écart sémantique

Le temps de calcul pour le processus appliqué à notre zone d'étude y compris la deuxième passe du processus est de 5'30".

D.3.2.4 Etude de la convergence du processus

Comme nous l'avons exposé au chapitre C, dans le but d'avoir une analyse plus globale, les résultats obtenus après une première passe du processus sont analysés afin de définir le critère voisinage. Ensuite, le processus est relancé avec ce nouveau critère. Ainsi, nous pouvons imaginer que le processus est répété jusqu'au moment où la convergence du processus est obtenue, c'est-à-dire que les résultats ne changent plus.

Dans cette partie nous abordons le problème lié à la convergence de notre processus dans son ensemble. Nous montrons que théoriquement, dans des cas exceptionnels notre processus ne converge pas.

Convergence théorique

Supposons qu'à la fin de la première passe du processus chaque objet du jeu de données moins détaillé trouve son objet homologue ou l'hypothèse NA est choisie. La deuxième passe du processus consiste à analyser les résultats d'appariement pour définir un nouveau critère qui prend en compte l'ensemble des objets, et ensuite nous relançons le processus. Par conséquent, nous nous posons la question de la convergence de notre processus. Nous démontrons à travers un exemple qu'en théorie notre processus ne converge pas dans certains cas extrêmes.

Considérons qu'après une première passe du processus nous obtenons le résultat d'appariement illustré en Figure 117. Notre processus se trouve dans un état 1. Les arcs A_1, D_1 et F_1 du jeu de données plus détaillé sont appariés avec le candidat C_2 du jeu de données moins détaillé, tandis que les arcs B_1 et E_1 sont appariés avec le candidat C_1.

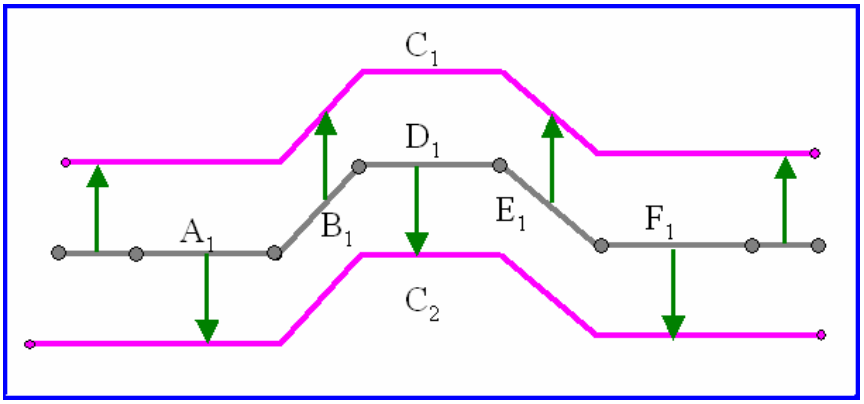


Figure 117. Non-convergence de notre processus : état du processus après une première passe

Ensuite, le processus d'appariement est lancé une deuxième fois. Ainsi pour chaque objet du jeu de données plus détaillé nous cherchons des candidats. Supposons que tous les arcs illustrés, A_1, B_1, D_1, E_1 et F_1 ont deux candidats C_1 et C_2. De plus, supposons que dans le cas idéal, le critère d'écart de position, le critère sémantique, le critère nom d'objet et le critère d'orientation soutiennent les deux candidats exactement de la même manière et que seul le critère du voisinage arrive à les distinguer et qu'après la fusion de critères et des candidats il n'y a pas de conflit total.

Considérons que nous sommes en train d'analyser l'arc B_1 qui après la première phase a été apparié avec le candidat C_1 (voir Figure 117). Après l'analyse de ses voisins, nous remarquons que ses deux voisins A_1 et D_1 sont appariés au candidat C_2. Conformément à la modélisation du critère de voisinage, ce dernier croit que le candidat C_2 est l'arc homologue de l'arc B_1 et ainsi, il accorde à l'hypothèse $appC_2$ une masse de croyance importante. En sachant que seul ce critère arrive à distinguer les deux candidats, l'arc B_1 sera apparié au candidat C_2. De la même manière l'arc E_1 qui après la première passe du processus a été apparié au candidat C_1 sera apparié après l'analyse des résultats avec le candidat C_2 puisque ses deux voisins D_1 et F_1 ont été appariés après la première passe du processus au candidat C_2.

Sous les mêmes hypothèses (le critère d'écart de position, le critère sémantique, le critère nom d'objet et le critère d'orientation soutiennent les deux candidats exactement de la même manière) et en faisant la même analyse des résultats (c'est-à-dire que nous analysons la manière dont les voisins ont été appariés après la première passe du processus), les arcs A_1, D_1 et F_1 qui initialement ont été appariés avec le candidat C_2 seront appariés au candidat C_1. Par exemple, l'arc D_1 a été apparié après la première passe du processus avec le candidat C_2. Etant donné que ses deux voisins, les arcs B_1 et E_1, ont été appariés avec le candidat C_1 (voir Figure 117), le critère de voisinage attribue une masse de croyance importante à l'hypothèse $appC_1$ et une masse de croyance quasi-nulle à l'hypothèse $appC_2$. En sachant que les autres critères soutiennent de la même manière les candidats C_1 et C_2, donc seul le critère de voisinage distingue les deux candidats, l'arc D_1 sera apparié après la fusion des cinq critères avec le candidat C_1.

Les résultats après la deuxième passe du processus sont illustrés sur la Figure 118. Ainsi, notre processus se trouve dans un état 2.

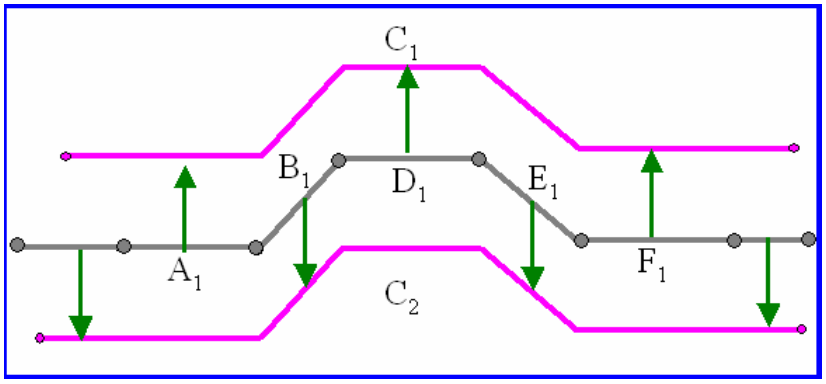


Figure 118. Non-convergence de notre processus : état du processus après une deuxième passe

Si nous continuons notre processus, c'est-à-dire que nous analysons les résultats issus de la deuxième passe pour réinitialiser les masses de croyance pour le critère de voisinage, pour les arcs du jeu de données plus détaillé en question, nous obtenons les résultats illustrés sur la Figure 117, c'est-à-dire que nous retournons à l'état 1. Si, le processus est relancé nous arrivons dans l'état 2, etc. Par conséquent, nous proposons un critère d'arrêt : nous considérons que le système s'arrête lorsqu'il revient à un état qu'il a déjà atteint.

Notre critère est pertinent, d'une part parce que notre processus est déterministe, c'est-à-dire que les résultats sont les mêmes si les données et les connaissances introduites dans le processus ne changent pas, et d'autre part, parce que nous avons un nombre d'états fini. En effet, si nous considérons deux jeux de données à apparier, JD_1 ayant n objets et JD_2 ayant m objets, et si nous cherchons pour chaque objet de JD_1 un homologue dans JD_2, le nombre de toutes les combinaisons d'appariement possibles est de m^n. Dans la pratique, le nombre de combinaisons possibles est considérablement réduit grâce au cadre de discernement que nous définissons au début du processus d'appariement.

Convergence empirique

Nous avons donc démontré à travers un exemple extrême la non-convergence théorique du processus. Nous avons également étudié empiriquement la convergence de notre processus sur deux extraits de notre zone d'étude choisis aléatoirement. Le premier extrait contient 1510 objets, tandis que le deuxième extrait contient 1176 objets. Nous avons observé que pour le premier extrait, à partir de la quatrième passe du processus, la convergence est atteinte, c'est-à-dire que les résultats d'appariement sont les mêmes que ceux obtenus après la troisième passe du processus. Par contre, pour le deuxième extrait, nous avons remarqué qu'après la troisième passe du processus les résultats d'appariements alternent entre les résultats de la deuxième passe et les résultats de la troisième passe. La non-convergence est due à deux objets qui sont en conflit à la deuxième passe, appariés à la troisième passe et à nouveau en conflit à la passe suivante, etc.

Dans le cas où le processus ne converge pas, une solution est d'arrêter le processus et de signaler les résultats qui changent d'une passe à l'autre puis de les analyser interactivement.

D.3.2.5 Bilan des expérimentations sur les réseaux routiers

Nous avons présenté dans cette partie les expérimentations que nous avons mises en oeuvre sur les jeux de données représentant les réseaux routiers. Les jeux de données utilisés

sont sensiblement différents en contenu et ils ont des niveaux de détail différents. De plus, ils présentent une forte différence au niveau de l'information attributaire.

Nous avons obtenu une précision moyenne de 95% et un rappel moyen de 93%, le taux d'erreur, c'est-à-dire les objets mal appariés, étant de 4 %. Cependant, il reste quelques cas d'appariement que le processus d'appariement n'arrive pas à gérer d'une manière efficace. Dans le chapitre E, nous présentons plus en détail les limites du processus, ainsi que les perspectives que nous proposons pour y remédier.

D.4 Conclusion

Dans ce chapitre, nous avons d'abord présenté l'environnement de notre travail, c'est-à-dire la plate-forme GeOxygene, et notre modèle conceptuel de données.

Ensuite, nous avons étudié la faisabilité de notre approche basée sur la théorie des fonctions de croyance à travers deux expérimentations. Elles ont été mises en œuvre sur des données ponctuelles représentant les points remarquables du relief et sur des données linéaires représentant les réseaux routiers. Les jeux de tests utilisés pour les deux expérimentations présentent des niveaux de détail différents.

En fonction des connaissances spécifiques à chaque type de données, nous avons utilisé plusieurs critères. Pour les points remarquables du relief, nous avons défini trois critères : le critère d'écart de distance, le critère sémantique et le critère toponymique, tandis que pour les réseaux nous avons utilisé cinq critères : le critère d'écart de position, le critère sémantique, le critère d'orientation, le critère nom d'objet et le critère de voisinage. Précisons que les critères d'écart de position et le critère sémantique sont similaires dans les deux expérimentations, aucune adaptation n'étant faite.

L'évaluation des résultats a été réalisée d'une manière empirique et a été présentée en termes de précision et de rappel. Pour les deux expérimentations menées, une précision et un rappel supérieurs à 90% ont été obtenus.

Nous avons également montré que notre processus permet une auto-évaluation des résultats d'appariement en se basant sur la différence entre le premier maximum choisi et le deuxième. Nous avons constaté que globalement, l'auto-évaluation des résultats est pertinente.

Les résultats obtenus avec notre approche ont été comparés avec les résultats obtenus avec d'autres approches d'appariement. Ainsi, pour les points remarquables du relief la comparaison a été réalisée avec deux approches spécifiques aux données ponctuelles, tandis que pour les réseaux routiers la comparaison a été réalisée avec une approche spécialement conçue pour les réseaux.

Nous avons remarqué que d'une part les résultats que nous avons obtenus sont similaires ou meilleurs, et d'autre part notre processus d'appariement présente une plus grande généricité, étant adapté à différents types de données et à des jeux de données ayant des niveaux de détail différents ou le même niveau de détail.

CHAPITRE E
Vers la conception d'un système générique d'appariement de données géographiques

E Vers la conception d'un système générique d'appariement de données géographiques

E.1 Introduction

L'objectif de ce chapitre est de présenter une proposition de système générique d'appariement de données géographiques qui permettrait à l'utilisateur d'adapter notre processus aux données qu'il souhaite apparier. L'adaptabilité de notre approche serait réalisée à travers une interface d'aide à la paramétrisation du processus, permettant par exemple le chargement des données et la spécification des données à traiter et des informations pertinentes pour l'appariement. Ainsi, l'utilisateur serait conseillé sur les paramètres pertinents qu'il peut utiliser.

Par ailleurs, des améliorations à notre processus d'appariement sont proposées.

Nous présentons ci-après les étapes à parcourir afin d'apparier deux jeux de données. Elles sont illustrées sur la Figure 119.

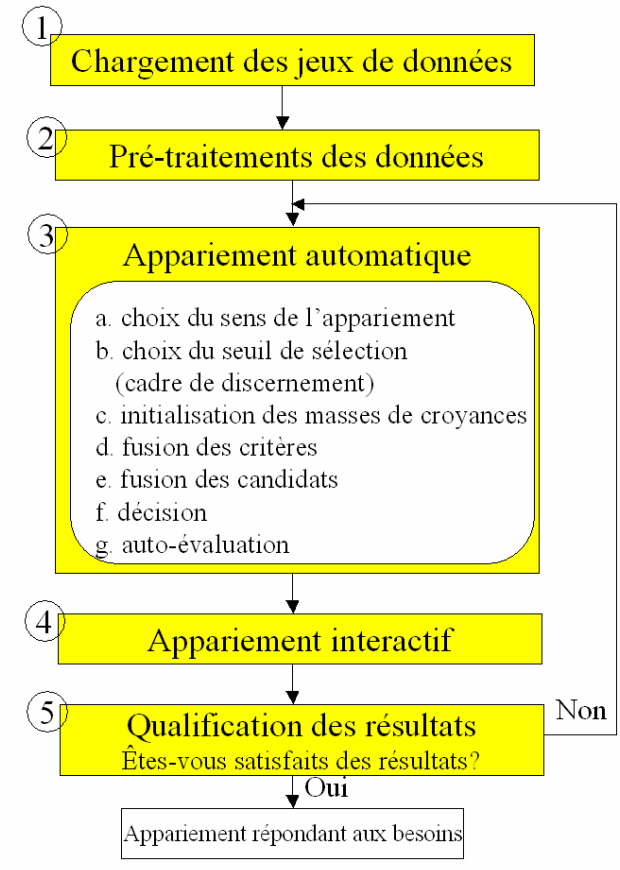


Figure 119. Les étapes du processus d'appariement de données géographiques

Supposons qu'un utilisateur souhaite apparier deux jeux de données géographiques. Afin d'y parvenir, il doit tout d'abord exprimer ses besoins et ensuite paramétrer le processus

d'appariement. Pour faciliter l'étape de traduction des besoins de l'utilisateur et pour aider l'utilisateur à paramétrer correctement le processus, nous lui proposons une interface. Précisons qu'afin de mettre en œuvre un processus complet d'appariement, une bonne connaissance des données géographiques est nécessaire.

L'interface générale du processus d'appariement est illustrée en Figure 120. Elle est composée de trois options : appariement de données, comparaison des appariements et stockage des liens.

L'option appariement de données consisterait à lancer le processus d'appariement. Ainsi, une fois cette option choisie, une nouvelle interface s'ouvrirait qui permettrait à l'utilisateur de choisir ses jeux de données. Nous y reviendrons dans le paragraphe E2 ci-après.

La comparaison des appariements permettrait à l'utilisateur de charger des résultats d'appariement et de les comparer. Ainsi, l'utilisateur pourrait utiliser la comparaison automatique afin de déterminer la meilleure méthode d'appariement ou d'évaluer les résultats. On pourrait supposer que les résultats identiques sont justes et que l'utilisateur évalue manuellement uniquement ceux qui sont différents ou jugés incertains.

L'intérêt de l'option stockage des liens serait pour l'utilisateur de pouvoir stocker les résultats d'appariement dans une table afin de les réutiliser dans une autre application ou de les comparer avec une autre méthode d'appariement.

La comparaison des appariements et le stockage des liens ne seront pas développés.

Figure 120. Interface générale d'appariement de données géographiques

E.2 Description du processus d'appariement de données (proposition)

Une fois que l'utilisateur aurait choisi l'option appariement de données, une nouvelle interface s'ouvrirait qui lui permettrait de choisir les deux jeux de données à apparier (voir la Figure 121). Précisons que dans ce mémoire de thèse nous nous sommes intéressés uniquement à l'appariement de deux jeux de données.

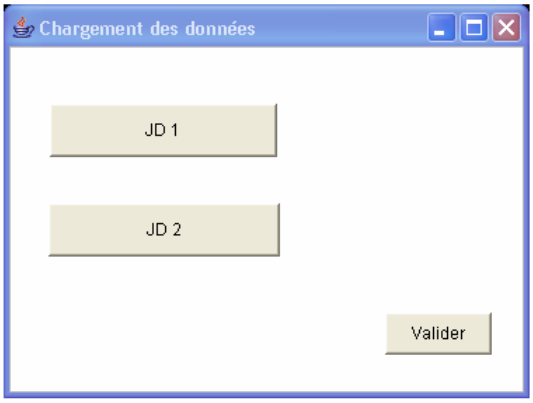


Figure 121. Interface pour le chargement des jeux de données.

Une fois les jeux de données chargés, l'utilisateur saisirait les spécifications des données à apparier. La saisie des spécifications est une étape importante, dont dépendent les résultats d'appariement.

La saisie des spécifications pourrait être réalisée à travers l'interface illustrée en Figure 122.

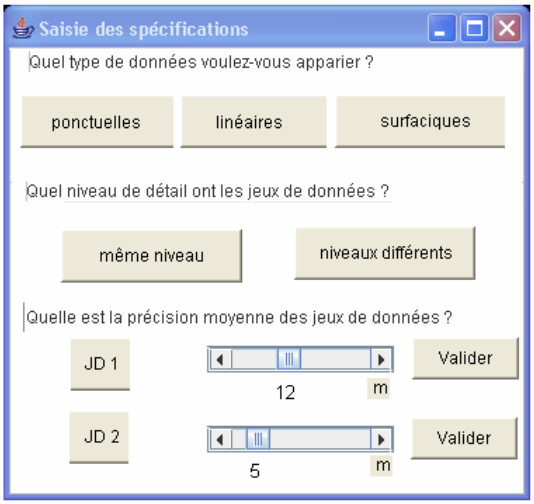


Figure 122. Interface pour la saisie des spécifications

Cette interface permettrait de saisir :

- le type des données à apparier, c'est-à-dire des données ponctuelles, linéaires ou surfaciques.

  L'intérêt de cette spécification est qu'en fonction du type de données nous proposons à l'utilisateur des mesures adéquates. Par exemple, s'il a des données linéaires nous lui proposerons la distance de Hausdorff plutôt que la distance euclidienne, etc. Un deuxième intérêt consiste à pré-traiter les données. Par exemple si l'utilisateur souhaite apparier des points et des surfaces représentant la même réalité, nous définissons un nouveau jeu de données à partir des centroïdes des surfaces,

- la spécificité des jeux de données, c'est-à-dire si les jeux de données ont le même niveau de détail (recouvrement total), ou des niveaux de détail différents (inclusion) ou des contenus sensiblement différents et des niveaux de détails différents.

Cette spécification est importante, puisqu'elle peut influencer l'appariement (à savoir si nous devons gérer des appariements de type 1 : 1 ou n : m),

- la précision de chaque base de données.

  A partir d'un curseur, l'utilisateur choisit la précision des jeux de données. La précision est un point important parce qu'elle permet de définir les seuils spécifiques au critère d'écart de position.

E.3 Pré-traitement des données

L'intérêt de l'étape de pré-traitement de données est de préparer les données à apparier afin de faciliter l'appariement proprement-dit. Cette étape dépend du type de données à apparier. Si les jeux de données ont des types géométriques différents, nous proposons à l'utilisateur un traitement qui permet de les rendre du même type. Par exemple, s'il s'agit d'un jeu de données de type linéaire et d'un jeu de données de type surfacique, un traitement de squelettisation pourrait lui être proposé afin de transformer les surfaces en lignes. Si les jeux de données sont de type linéaire, afin que notre approche d'appariement puisse être appliquée, une structure de réseau est nécessaire. D'autres traitements peuvent être imaginés, tels que l'élimination des nœuds avec seulement 2 arcs incidents, la fusion des nœuds très proches, ou le découpage d'un réseau en certains points.

Notre approche a été testée sur des jeux de données ponctuels et linéaires, mais elle est également adaptable aux jeux de données surfaciques. Par ailleurs, elle a été testée sur des jeux de données ayant des niveaux de détail différents, mais elle peut être aussi utilisée pour apparier des jeux de données ayant le même niveau de détail. Les résultats d'appariement seront en principe améliorés puisque les jeux de données se ressemblent plus.

E.4 Appariement automatique

L'étape d'appariement automatique est celle que nous avons proposée au chapitre C. Tout au long de cette étape, plusieurs choix sont à faire. Cela peut parfois sembler fastidieux à l'utilisateur mais cette étape permettrait une plus grande faculté d'adaptation aux données à apparier. Nous considérons que la flexibilité représente un grand avantage de notre approche d'appariement.

E.4.1 Choix du sens de l'appariement

Le premier choix à faire est de déterminer le jeu de données à apparier, c'est-à-dire le jeu de données pour lequel nous cherchons des homologues dans l'autre jeu de données. Pour une meilleure compréhension, nous appelons le jeu de données pour lequel nous cherchons des homologues, « jeu de référence » et l'autre jeu de données « jeu de comparaison ». Précisons que les termes de « référence » et de « comparaison » ne sont pas liés à la qualité des jeux de données.

Cette étape est nécessaire en raison de la cardinalité des liens d'appariement définie par notre approche. Nous la considérons comme un point faible de notre approche qui réduit sa généricité. En effet, notre approche permet uniquement de définir des liens d'appariement de cardinalité 1 : 1, puisqu'afin de prendre la décision finale, nous avons choisi la probabilité pignistique. Comme nous l'avons déjà dit, cette dernière favorise les hypothèses simples, c'est-à-dire un seul candidat à l'appariement ou l'hypothèse NA (non-apparié).

Si les jeux de données ont le même niveau de détail, le choix de la base de référence et de la base de comparaison n'est pas primordiale, puisqu'en général l'appariement est de

cardinalité 1 : 0 ou 1 : 1. Signalons que dans ce cas, notre approche fonctionne dans les deux sens sans avoir d'importantes répercussions sur les résultats. Une analyse des résultats d'appariement pourrait être réalisée afin de voir si un objet de la base de comparaison a été choisi par deux objets différents de la base de référence. Cette situation peut arriver puisque dans notre approche chaque objet de la base de référence choisit un candidat. Ensuite, une vérification de tels cas d'appariement est nécessaire afin de décider s'il s'agit d'un appariement juste ou faux. Nous y reviendrons dans l'étape d'auto-évaluation du processus.

Si l'utilisateur le souhaite, il peut apparier les jeux de données dans les deux sens, et comparer les résultats grâce à la fonction de comparaison de l'interface générale.

En revanche, si les jeux de données ne présentent pas le même niveau de détail, le choix de la base de référence et celle de comparaison est primordial. Pour pouvoir choisir la base de référence, l'utilisateur doit connaître le type de correspondances entre les objets des jeux de données. Si les types de correspondances recherchés sont de 1 : 0 ou 1 : 1, alors le jeu de référence est celui qui est moins détaillé. Par contre, si le type de correspondance est de 1 : n ou n : m, ce qui est le plus possible, le jeu de référence doit être celui qui est le plus détaillé. Notre approche, telle que nous l'avons décrite au chapitre C permet de gérer des correspondances de type 1 : 0, 1 :1 et 1 : n. Le type 1 : n est obtenu par regroupement des appariements 1 : 1.

Nous présentons sur la Figure 123 un exemple de correspondance 1 : n. Notre processus définit les liens d'appariement de cardinalité 1 : 1 (A_1, A_2), (B_1, A_2), (C_1, A_2) et (D_1, A_2). Après le regroupement des liens, l'arc A_2 du jeu de données moins détaillé est apparié avec quatre arcs, donc le type de correspondance 1 : n (avec n = 4).

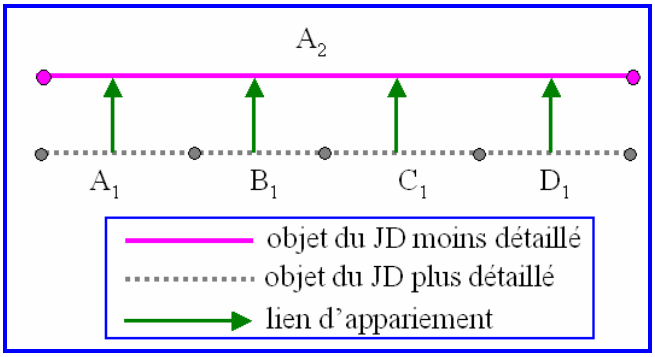


Figure 123. Exemple de type de correspondance 1 : n

Discussion sur le type de correspondance n : m

Notre approche ne permet pas de définir directement des liens d'appariement n : m. La définition des liens d'appariement n : m représente un grand défi pour l'appariement de données géographiques. Ce type de correspondance est surtout rencontré lorsque les données à apparier sont linéaires ou surfaciques. Prenons par exemple les réseaux routiers. Les découpages différents de deux jeux de données, les carrefours complexes ainsi que les routes à chaussées séparées, peuvent engendrer des liens d'appariement de cardinalité n : m. Dans la pratique, il existe peu de cas où les correspondances entre les objets sont du type n : m, lorsque les entités du monde réel sont représentées par des lignes. En revanche, dans les jeux de données où les entités sont représentées par des objets surfaciques, les correspondances du type n : m sont beaucoup plus fréquentes.

Une extension de l'approche qui permettra de définir des liens n : m est absolument nécessaire pour les jeux de données surfaciques et elle peut être aussi envisageable pour les réseaux ou pour les jeux de données ponctuels. A notre avis, pour les données linéaires, il serait nécessaire d'étudier d'abord l'impact de cette extension sur les autres cas 1 : n afin de voir si elle ne dégrade pas les résultats d'appariement. Si c'est le cas, nous considérons qu'une solution immédiate consiste à redécouper les réseaux (projeter chaque nœud d'un jeu de données sur l'autre jeu de données). Par conséquent uniquement des correspondances 1 : n existeront après découpage.

Afin de mieux gérer les cas n : m nous proposons deux solutions :

- si on veut rester dans le cadre de la théorie des fonctions de croyance, une solution peut être apportée au niveau de la décision, c'est-à-dire qu'au lieu d'utiliser la probabilité pignistique qui privilégie les hypothèses simples, nous pouvons utiliser la fonction de crédibilité ou de plausibilité qui privilégie les hypothèses combinées. Dans ce cas, le OU des hypothèses composées est transformé en ET. Par exemple, si pour un objet *obj1* l'hypothèse $app(C_1, C_2, C_3)$ est choisie, signifiant que l'objet est apparié soit avec le candidat C_1, soit C_2, soit C_3, nous pouvons considérer que l'objet *obj1* est apparié avec les candidats C_1, C_2 et C_3. De cette manière, notre approche définit des correspondances du type 1 : n directement. Par conséquent, un deuxième appariement est nécessaire dans l'autre sens, c'est-à-dire qu'il faut changer le jeu de données de référence. A la fin de cette étape nous obtenons des liens 1 : m. En réalisant un regroupement nous obtenons les liens n : m,
- une autre solution est de garder le choix de l'utilisation de la probabilité pignistique et de réaliser deux appariements de données, c'est-à-dire dans les deux sens. Par contre après chaque appariement un groupement de liens d'appariement doit être réalisé pour définir des liens 1 : n et ensuite un troisième pour définir les liens n : m.

La première solution présente l'inconvénient que des sur-appariements peuvent apparaître plus facilement, d'une part en raison de la transformation du OU en ET, et d'autre part en raison de la modélisation prudente des connaissances que nous avons proposée. Elle a l'avantage de définir directement des appariements du type 1 : n. L'inconvénient de la deuxième solution est que plusieurs étapes de regroupements doivent être mises en œuvre.

E.4.2 Choix du seuil de sélection

L'interface adoptée à cette étape est montrée en Figure 124.

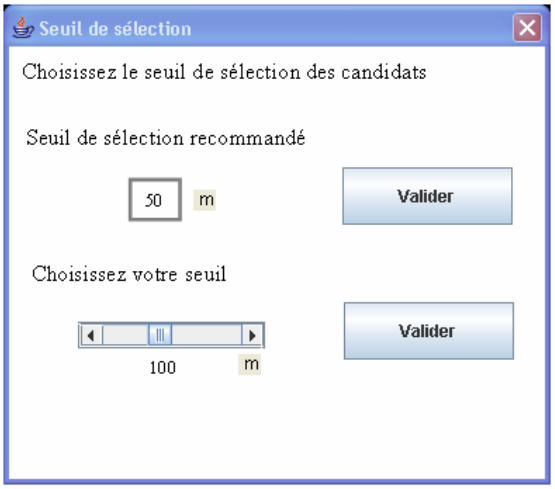


Figure 124. Interface pour la sélection des candidats à l'appariement

Afin de choisir le seuil de sélection qui nous permet de définir le cadre de discernement, l'utilisateur aurait le choix entre choisir les seuils que nous lui recommandons et déterminer ses propres seuils à travers un curseur. Il doit décider également s'il souhaite un seul seuil pour tous les objets du jeu de données de référence ou plusieurs seuils en fonction de la nature des objets. Cette solution peut être utile lorsque le recouvrement des deux jeux de données est très faible ou lorsque les deux jeux de données représentent des entités hétérogènes, comme par exemple les points remarquables du relief.

Il se peut que le recouvrement de deux jeux de données soit important dans certains zones et très faible dans d'autres. Dans ce cas un seuil de sélection large adapté à la zone du faible recouvrement ne sera pas adapté à la zone du recouvrement important en raison de l'augmentation du nombre de candidats à l'appariement. Un nombre important de candidats fait croître le temps de calcul. Afin d'y remédier nous pourrions proposer un seuil large accompagné d'un filtrage des candidats en fonction de leur nature. Cette solution permettrait d'une part à tous les objets du jeu de données d'avoir au minimum un candidat et d'autre part, de ne pas apparier un immeuble avec une église, par exemple. En revanche cette solution de choisir les candidats en fonction de leur nature serait pertinente uniquement si la sémantique des deux jeux de données n'est pas trop hétérogène.

Nous avons vu au chapitre D que pour nos expérimentations nous avons utilisé un seuil empirique. Cependant, ces seuils de sélection pourraient être déterminés automatiquement par un processus d'apprentissage automatique soit à partir des données, soit à partir des résultats d'appariement déjà obtenus. Cette perspective nous semble pertinente, d'autant plus que l'apprentissage automatique a fait ses preuves dans le domaine de l'information géographique [Sester, 1998 ; Plazanet *et al.*, 1998 ; Hubert et Ruas, 2003 ; Mustière, 2001 ; Sheeren, 2005 ; Taillandier, 2007].

E.4.3 Initialisation des masses de croyance

Cette étape représente le cœur de notre processus. Elle consiste d'une part à aider l'utilisateur à choisir des critères d'appariement pertinents en fonction des données à apparier et d'autre part à modéliser les connaissances pour les critères choisis.

Définition des critères

A partir de la saisie des spécifications des données, nous seront capables de proposer à l'utilisateur plusieurs critères mais aussi lui laisser le choix d'en ajouter d'autres en fonction des données et en fonction des connaissances qu'il possède. L'interface relative à la définition des critères est illustrée sur la Figure 125.

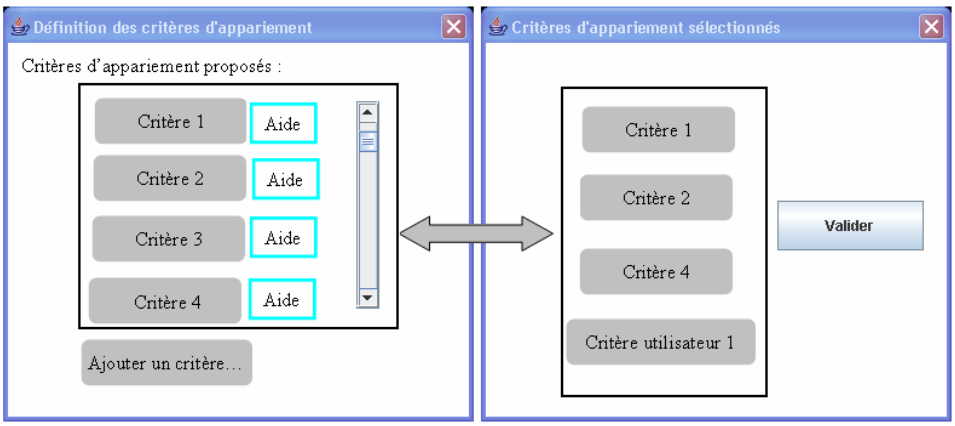


Figure 125. Interface pour définir des critères d'appariement pertinents

A chaque bouton désignant un des critères proposé, une aide est associée. Celle-ci détaille les connaissances nécessaires pour définir le critère et également les principes de base de ce critère. A titre d'illustration, nous montrons sur la Figure 126 un exemple d'aide pour le critère d'écart de position.

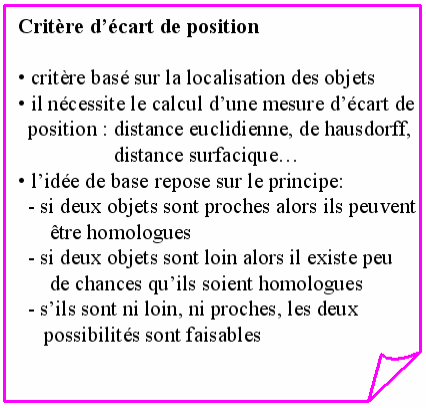


Figure 126. Exemple d'aide pour le critère d'écart de position

Discussion sur les critères d'appariement proposés

Dans le chapitre C nous avons présenté quelques critères d'appariement typiques, tels que le critère d'écart de distance, le critère sémantique, le critère orientation, etc. Ensuite, dans la partie expérimentation présentée dans le chapitre D, nous avons étudié la pertinence de ces critères en fonction des données à apparier.

Comme nous l'avons vu, la pertinence des critères utilisés a été montrée puisqu'on obtient des taux de précision et de rappel supérieurs à 90%. Cependant, nous avons vu que des améliorations sont nécessaires et que l'ajout d'autres critères peut être efficace.

Pour les points remarquables du relief nous avons distingué deux problèmes principaux que notre approche n'arrive pas à gérer :

- un premier problème qui peut apparaître est lié au fait qu'un candidat soit choisi par deux objets de référence (voir la Figure 127). Ce cas ne pose pas de problème si nous cherchons des appariements du type 1 : n, il est tout à fait normal que cela se produise. Par contre, si nous cherchons des appariements du type 1 : 1, cela pose problème. Ce cas de figure est possible puisque dans notre processus, lorsque nous cherchons à apparier un objet *obj1,* nous n'analysons pas la manière dont ses voisins ont été appariés, nous nous intéressons uniquement aux candidats à l'appariement. Précisons que ce cas de figure ne s'est pas produit dans nos expérimentations, mais nous considérons qu'il mérite d'être approfondi.

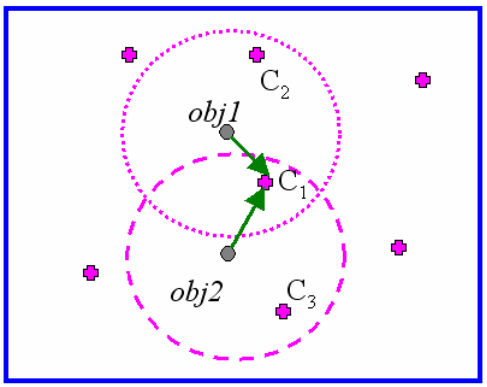


Figure 127. Exemple d'appariement où un candidat a été choisi par deux objets

Nous proposons deux solutions :

1. La première consiste à exploiter la valeur de la probabilité pignistique qui a permis de choisir le candidat. Par exemple, si nous avons deux liens d'appariement (*obj1*, C_1) et (*obj2*, C_1), chaque lien ayant un poids donné par la probabilité pignistique, nous pouvons choisir le lien qui a le poids le plus important.
2. La deuxième solution consiste à faire une analyse globale des objets. Par exemple, [Samal *et al.*, 2004] proposent une approche d'appariement qui analyse le contexte spatial des objets. Cette idée peut être introduite comme un nouveau critère d'appariement. Ce critère s'appuie sur le contexte spatial entre les objets d'un même jeu de données et sur un vecteur d'erreur systématique entre les objets des deux jeux de données. Considérons le cas de figure illustré en Figure 128. Supposons qu'après un premier appariement nous obtenions les couples d'objets homologues entourés par un cercle vert. Nous remarquons que le candidat C_5 a été choisi à la fois par l'objet *obj1* et par l'objet *obj2*. En haut de la figure, nous montrons le contexte spatial de l'objet *obj1* (à droite) et de l'objet candidat C_5 (à gauche).

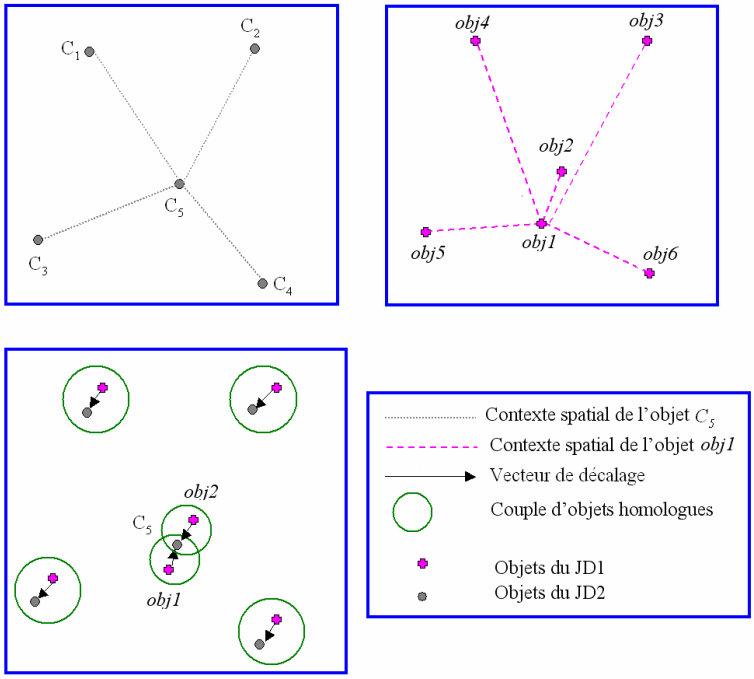


Figure 128. Nouveau critère : analyse du contexte spatial et du vecteur d'erreur systématique

Le principe de ce critère est que pour tous les objets faisant partie du contexte spatial de *obj1* nous pourrions analyser le vecteur de décalage formé avec les objets composant le contexte spatial de l'objet candidat C_5, ainsi que la manière dont les objets ont été appariés. Le vecteur de décalage, illustré par une flèche noire, relie deux objets homologues appartenant à des contextes spatiaux différents. Nous pouvons remarquer par exemple que pour tous les couples d'objets homologues, sauf pour le couple (*obj1*, C_5), le vecteur de décalage est différent. Par conséquent, d'après le contexte spatial, nous pouvons affirmer qu'il y a plus de chances que le candidat C_5 soit l'objet homologue de l'objet *obj2* que de l'objet *obj1*.

- Un deuxième problème est lié à la possibilité que dans un jeu de données plusieurs entités du monde réel soient représentées par un seul objet, et que dans l'autre jeu de données toutes les entités soient représentées. Ce cas n'est pas géré par notre approche si nous cherchons des correspondances du type 1 : 1 et si nous considérons comme jeu de référence celui qui est le moins détaillé. Une solution peut être d'apparier les jeux de données dans les deux sens, et de faire un post-traitement afin d'éliminer les liens d'appariement parasites.

En ce qui concerne les réseaux routiers, nous avons vu dans la partie D.3 que notre approche ne gère pas bien les ronds points, les pattes d'oies et les carrefours. Afin d'y remédier, nous proposons d'une part d'apparier un arc à un nœud, et d'autre part d'introduire un nouveau critère qui s'intéresse uniquement à ces structures. Pour ce faire, nous pouvons d'abord détecter ces structures [Grosso, 2004 ; Sheeren, 2005]. Une fois qu'elles sont détectées, une sélection des candidats (points et arcs) est faite, puis le nouveau critère les analyse et se prononce en leur faveur/défaveur, exactement comme les autres.

Modélisation des critères d'appariement

Une fois que l'utilisateur a choisi les critères qu'il utilisera dans le processus d'appariement, il peut passer à leur modélisation afin d'initialiser les masses de croyance. Comme nous l'avons vu au chapitre C, la modélisation des connaissances pour chaque critère consiste à déterminer les seuils utilisés et la forme de la fonction définissant les masses de croyance.

L'utilisateur pourrait choisir la modélisation et les seuils concernés que nous lui proposons, mais également fixer ses seuils, modifier la forme des courbes et même modéliser les connaissances pour définir ses propres critères d'appariement.

Les seuils que nous lui proposons sont fonction de la saisie des spécifications. Dans nos expérimentations, les seuils ont été déterminés empiriquement à partir des connaissances issues des données (précision des jeux de données), des spécifications ou des experts. Nous avons également vu que notre processus est peu sensible aux seuils sélectionnés, donc des seuils approximatifs sont suffisants. Cette faible sensibilité est possible grâce à notre modélisation prudente, à la fusion de plusieurs critères d'appariement et au fait que nous n'éliminons jamais de candidat et qu'il y a peu de candidats.

Cependant, nous pouvons améliorer cet aspect en utilisant des seuils déterminés grâce à des méthodes basées sur la fouille de données [Sheeren, 2005] ou en utilisant les courbes ROC[8] [Fawcett, 2004 ; Champion, 2007]. Les seuils ainsi déterminés seront un compromis entre la qualité des résultats et le nombre d'appariements trouvés, c'est-à-dire un compromis entre la précision et le rappel.

[8] ROC = Receiver Operating Characteristic

Un exemple d'interface pour la détermination des seuils pour le critère d'écart de position est illustré en Figure 129.

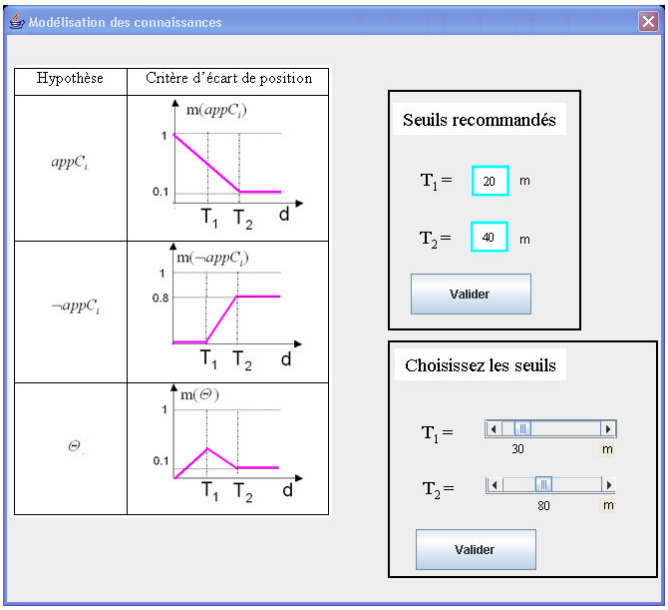


Figure 129. Exemple d'interface pour la détermination des seuils pour le critère d'écart de position

Conseils pour la modélisation des connaissances

Comme nous l'avons vu dans le chapitre C, la forme des courbes qui définissent les masses de croyance est différente d'un critère à l'autre. Les critères d'appariement n'ont pas la même importance, des poids différents étant attribués. Ceci est dû au fait que chaque critère modélise des connaissances différentes qui peuvent être plus ou moins parfaites.

Cette flexibilité représente un avantage important de notre approche : nous pouvons exprimer dans le cadre d'un même processus des connaissances différentes (liées à la localisation, à la sémantique, ou à la forme des objets géographiques) et des connaissances ayant des niveaux de précision différents (par exemple, les connaissances utilisées pour le critère basé sur la localisation des objets géographiques sont plus fiables que celles utilisées pour le critère nom d'objet). Nous croyons que cette étape de modélisation ne pose pas de vrai problème.

Par conséquent, la question que nous nous posons est comment pouvons-nous améliorer cette étape, en la rendant plus générique et plus facile pour les utilisateurs ? Nous essayons d'apporter une réponse à cette question.

Comment définir les courbes ?

Afin de définir la forme des courbes pour chaque critère nous nous sommes basés sur les idées suivantes :

- la condition m ($appC_i$) +m ($\neg appC_i$) + m(Θ) =1 doit toujours être respectée,
- lorsque nous avons des bonnes connaissances, nous définissons d'abord la forme des courbes pour les hypothèses $appC_i$ (l'objet homologue est le candidat C_i) et $\neg appC_i$ (l'objet homologue n'est pas le candidat C_i), et le complément est attribué à l'ignorance,

- lorsque nous n'avons pas de connaissances précises, nous définissons d'abord la forme de la courbe pour l'ignorance, puis le complément est reparti entre les hypothèses $appC_i$ et $\neg appC_i$. Dans ce cas, nous considérons que l'ignorance détermine le poids du critère,
- nous proposons que la masse de croyance m($appC_i$) ne soit jamais nulle afin de ne pas éliminer définitivement le candidat C_i,
- les courbes pour les hypothèses $appC_i$ et $\neg appC_i$ doivent être opposées et la courbe pour l'ignorance doit suivre la variation (croissante ou décroissante) soit de l'hypothèse $appC_i$ soit de l'hypothèse $\neg appC_i$,
- s'il arrive que pour un attribut d'un objet géographique (à part ceux pour lesquels nous avons l'information que les attributs peuvent ne pas avoir une valeur remplie) la valeur d'un attribut n'est pas utilisée nous proposons d'attribuer à l'ignorance la valeur maximale, c'est-à-dire 1.

Nous donnons des pistes pour déterminer les courbes relatives à quelques types de connaissances que nous pouvons utiliser pour définir un critère.

1. Critère basé sur la localisation des objets géographiques

A notre avis, le critère basé sur la localisation des objets géographiques est le critère le plus important.

- si l'écart de position est faible, alors la masse de croyance associée à l'hypothèse $appC_i$ doit être importante et la masse associée à l'hypothèse $\neg appC_i$ doit être faible,
- si l'écart de position est important, alors la masse de croyance associée à l'hypothèse $appC_i$ doit être très faible mais non nulle, et la masse associée à l'hypothèse $\neg appC_i$ doit être importante. Dans les deux cas le complément est attribué à l'ignorance,
- enfin, si l'écart n'est ni faible ni important, la valeur de l'ignorance est plus importante que les masses de croyance attribuées aux hypothèses $appC_i$ et $\neg appC_i$.

2. Critère basé sur la sémantique

Le critère basé sur la sémantique des objets géographiques exprime des conditions nécessaires mais non suffisantes. Les principes de ce critère sont :

- si l'écart sémantique est nul, alors les hypothèses importantes sont $appC_i$ et l'ignorance. Plus l'écart augmente, moins les hypothèses deviennent possibles, ce qui se traduit par une courbe descendante. La courbe pour l'hypothèse $\neg appC_i$ est le complément des deux autres,
- à partir d'un certain écart sémantique considéré comme important, l'hypothèse principale est $\neg appC_i$. Nous fixons la courbe pour cette hypothèse et le complément est partagé entre les courbes pour l'hypothèse $appC_i$ et l'ignorance,
- les conditions nécessaires mais non suffisantes peuvent être exprimées par des courbes relativement similaires pour l'hypothèse $appC_i$ et l'ignorance.

3. Critère basé sur la toponymie

Ce critère a un poids important si les toponymes se ressemblent, à plus forte raison s'ils sont identiques. Les principes de ce critère sont :

- si l'écart toponymique est nul, l'hypothèse la plus importante est $appC_i$. Plus l'écart augmente, plus l'hypothèse $\neg appC_i$ devient possible. L'ignorance peut être le complément des deux,
- à partir d'un certain écart sémantique considéré comme important, l'hypothèse $\neg appC_i$ devient possible, mais pour modéliser les éventuelles imperfections, nous utilisons l'ignorance. Nous proposons donc de fixer la courbe pour l'ignorance et le complément est partagé entre les courbes pour les hypothèses $appC_i$ et $\neg appC_i$.

4. Critère basé sur les noms d'objets géographiques

Ce type de critère discret est à utiliser lorsqu'il existe des attributs qui n'ont pas toujours de valeur remplie, et que le fait d'avoir une valeur ou pas est lié à l'importance des objets. Nous devons déterminer dès le début de la modélisation des connaissances si nous devons accepter ou pas un écart entre les noms. Par exemple pour les numéros de route, la route « N111 » n'est pas du tout la même que la route « N11 » bien que l'écart entre les deux noms soit très faible. Les principes de ce critère sont les suivants :

- si deux objets n'ont pas de nom, nous proposons une masse de croyance forte pour l'ignorance, mais pas égale à l'unité puisque nous savons qu'ils ont la même importance. Par conséquent, le complément est partagé entre les hypothèses $appC_i$ et $\neg appC_i$,
- si parmi les deux objets à comparer, il y en a un qui possède un nom et l'autre pas, l'hypothèse la plus crédible est $\neg appC_i$, en raison de l'importance des objets. Le complément est partagé entre les hypothèses $appC_i$ et l'ignorance,
- si les deux objets ont bien un nom, qu'ils sont différents, c'est-à-dire que l'écart n'est pas nul, l'hypothèse la plus crédible est $\neg appC_i$. En fonction de la décision prise relative à l'acceptation ou pas d'un écart, la valeur de la masse de croyance attribuée à cette hypothèse peut être plus élevée ou plus faible. Dans ce cas, la variation de la masse de croyance se fera en parallèle avec l'ignorance, c'est-à-dire qu'une diminution de la masse attribuée à l'hypothèse $\neg appC_i$ entraîne une augmentation de la masse attribuée à l'ignorance, la masse attribuée à l'hypothèse $appC_i$ restant constante,
- si les deux objets ont le même nom les hypothèses importantes sont $appC_i$ et l'ignorance. Il s'agit dans ce cas d'une condition nécessaire mais non suffisante.

5. Critère basé sur la topologie

Etant donné que ce critère est initialisé à partir de résultats d'appariement déjà obtenus, que son but est de corriger d'éventuelles erreurs d'appariement et qu'ensuite le processus est relancé avec tous les critères y compris ce critère, nous proposons qu'il soit plus rigide que les autres. Les erreurs à corriger peuvent être des objets qui ont été appariés mais à tort ou des objets qui n'ont pas été appariés mais qui devaient l'être. Les caractéristiques que nous proposons pour ce critère sont les suivantes :

- si une condition n'est pas satisfaite, alors nous proposons d'éliminer le candidat, c'est-à-dire m($appC_i$) =0. La masse attribuée à l'hypothèse ¬$appC_i$ doit être élevée,
- si une condition sur le voisinage est satisfaite, alors l'hypothèse la plus crédible est $appC_i$, donc la masse de croyance sera importante. Nous proposons d'éliminer complètement l'hypothèse ¬$appC_i$. De la même manière que dans le cas précédent, nous pouvons attribuer une masse de croyance non nulle à l'ignorance,
- par contre, si une condition sur le voisinage n'est pas complètement satisfaite, alors l'ignorance a un poids très important et le complément de la masse de croyance doit être partagé entre les hypothèses $appC_i$ *et* ¬$appC_i$,
- en ce qui concerne les conditions nécessaires sur le voisinage, nous considérons que celles-ci dépendent des niveaux de détail des deux jeux de données à apparier. Il nous semble difficile de donner des conseils génériques.

6. Critère basé sur les informations géométriques implicites : la forme, l'orientation, l'angle entre les objets géographiques

Nous considérons qu'un critère basé sur des informations implicites issues de la géométrie des objets géographiques est sensiblement dépendant du niveau de détail des jeux de données. Par conséquent il aurait un poids moins important que les autres. Pour cela, nous proposons d'exploiter l'avantage offert par l'ignorance. Parmi les principes de base, nous mentionnons les suivants :

- afin de rendre le critère indépendant du niveau de détail, nous proposons d'attribuer à l'ignorance une masse de croyance constante, indépendante de la valeur de la mesure utilisée,
- les hypothèses $appC_i$ et ¬$appC_i$ auront donc des courbes opposées dont la somme des deux est le complément de la courbe fixée pour l'ignorance. Pour une mesure faible, signifiant une forte ressemblance entre deux objets, l'hypothèse $appC_i$ sera plus importante que l'hypothèse ¬$appC_i$, tandis que pour une mesure élevée, signifiant une très faible ressemblance, l'hypothèse ¬$appC_i$ sera plus élevée que $appC_i$.

7. Critère basé sur le contexte spatial

Nous pensons que le principe de ce critère devrait être semblable au critère de voisinage et surtout en ce qui concerne le poids de ce critère, puisqu'il est rajouté à la deuxième passe du processus. Nous pensons que ce critère pourrait être continu, comme les critères d'écart de position, sémantique, etc. Pour un contexte spatial donné, nous proposons donc d'analyser les vecteurs de décalages, chacun défini par une distance et un angle. Parmi les principes de base, nous exposons les suivants :

- si le vecteur de décalage du couple d'objets en cours d'analyse est très semblable aux autres vecteurs de décalage du contexte spatial, alors l'hypothèse $appC_i$ est la plus crédible. L'ignorance peut être très faible mais plus élevée que l'hypothèse ¬$appC_i$,
- si le vecteur de décalage du couple d'objets est très différent des autres vecteurs, la masse attribuée à l'hypothèse $appC_i$ doit être élevée. L'ignorance peut être très

faible mais plus élevée que l'hypothèse $\neg appC_i$. Cette dernière peut même avoir une masse de croyance nulle.

- enfin, si le vecteur de décalage du couple d'objets n'est ni très semblable aux autres vecteurs ni très différent, alors l'ignorance joue un rôle très important, et la masse de croyance associée est élevée. Le complément est partagé entre les hypothèses $appC_i$ et $\neg appC_i$.

Ce critère serait surtout adapté au cas où un des deux jeux de données à apparier aurait été mal recalé.

La modélisation des connaissances est-elle générique ?

Lorsque nous parlons de la généricité de la modélisation des connaissances, nous faisons référence à deux aspects qui peuvent la rendre générique ou qui peuvent affaiblir sa généricité, à savoir : le type de données à apparier et le niveau de détail.

Nous pensons que la forme des courbes pour les critères continus ou les distributions ponctuelles des masses de croyance pour les critères discrets est indépendante du type de géométries des objets à apparier, point, ligne ou surface. De ce point de vue, les seuls changements qui peuvent intervenir sont les mesures utilisées, par exemple la distance euclidienne est plutôt employée pour les objets ponctuels que pour les objets linéaires.

Par ailleurs, nous considérons que la modélisation des connaissances ne dépend pas du niveau de détail, sauf pour le critère de voisinage. Cette indépendance est due au fait que notre modélisation est prudente, c'est-à-dire qu'elle permet en général d'apparier des objets pour lesquels le degré de ressemblance est moyen. Par conséquent, nous pensons que notre modélisation est adaptée aux jeux de données ayant des niveaux de détail différents. A plus forte raison nous croyons qu'elle est utilisable pour apparier des jeux de données ayant le même niveau de détail, et que les résultats d'appariement seront plus satisfaisants, puisque les objets homologues ont un degré de ressemblance plus élevé.

Le critère basé sur le voisinage est sensiblement dépendant du niveau de détail. Par rapport au critère que nous avons proposé dans le chapitre C, nous considérons que les modifications nécessaires concerneraient surtout la condition exprimée à travers le cas c), c'est-à-dire quand plusieurs voisins du candidat C_i sont appariés à plusieurs voisins du groupe connexe. Par conséquent, dans ce cas de figure, nous proposerions une ignorance moins importante et une masse de croyance plus importante attribuée à l'hypothèse $appC_i$.

La modélisation des connaissances pour les critères qui s'appuient sur l'information attributaire n'est pas dépendante des données utilisées. En revanche, les seuils utilisés pour le critère qui exploite la localisation des objets géographiques peuvent être dépendants des niveaux de détails utilisés. Les résultats d'appariement peuvent sensiblement varier.

E.4.4 Fusion des critères et des candidats

Bien que nous ayons proposé dans le chapitre C une fusion des critères basée sur un opérateur conjonctif, d'autres opérateurs peuvent être proposés à l'utilisateur.

Une fois la fusion des critères réalisée, l'opérateur doit mettre en œuvre la fusion des candidats. Lors de cette étape, il n'a plus d'autres opérateurs à choisir puisque nous considérons que l'opérateur doit être le même dans les deux étapes de fusion. En revanche, après la fusion des candidats il peut, s'il le souhaite, analyser le conflit global obtenu et ensuite choisir une distribution du conflit ou pas. En fonction de son choix, des opérateurs lui sont proposés. Par exemple, s'il ne souhaite pas redistribuer le conflit, une normalisation par

défaut est faite. Dans le cas contraire, d'autres opérateurs peuvent lui être proposés [Royère, 2002 ; Colot, 2000].

E.4.5 Décision

Afin de prendre une décision finale, nous avons utilisé la probabilité pignistique. Cependant, si l'utilisateur souhaite définir des liens d'appariement autres que ceux du type 1 : 1, nous pourrions lui proposer d'autres solutions, telles que la fonction de crédibilité et la fonction de plausibilité. Dans ce cas, l'utilisateur devrait être averti de l'impact que ces deux méthodes peuvent avoir sur le résultat d'appariement. L'utilisation de la fonction de crédibilité peut rendre des résultats pessimistes, c'est-à-dire des sous-appariements, tandis que la fonction de plausibilité peut produire des résultats optimistes, c'est-à-dire des sur-appariements.

Un exemple d'interface relative à l'étape de décision finale est illustré sur la Figure 130.

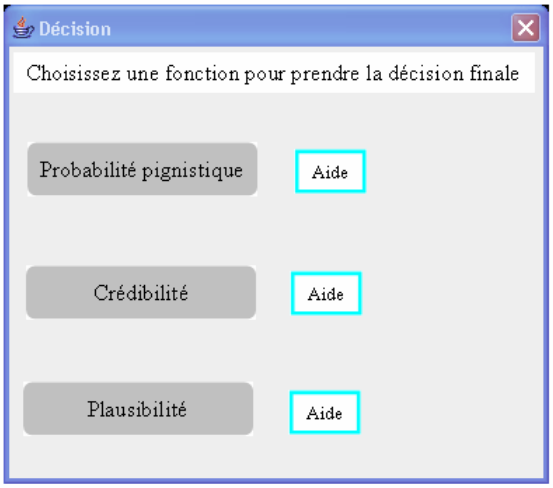


Figure 130. Exemple d'interface pour l'étape de décision finale

E.4.6 Auto-évaluation

Un des avantages de notre processus est qu'il permet une auto-évaluation. Nous pouvons imaginer deux types d'auto-évaluation :

1. la détection des objets de référence qui ont choisi le même candidat, sous l'hypothèse que ce cas ne doit pas arriver. Si c'est le cas, l'utilisateur aurait la possibilité de faire une analyse, en choisissant une des propositions que nous avons présentées ci-dessus. Il s'agit de l'analyse de la probabilité pignistique qui a comme conséquence de garder l'appariement pour lequel la probabilité pignistique est la plus élevée et d'ajouter le critère basé sur le contexte spatial.
2. la deuxième auto-évaluation consisterait à évaluer les appariements comme sûrs/ incertains en utilisant la probabilité pignistique, la fonction de plausibilité [Appriou, 1999], etc. L'utilisateur pourrait choisir le seuil à partir duquel un appariement est jugé comme sûr ou incertain.

E.5 Appariement interactif

Afin d'évaluer les résultats d'appariement, l'utilisateur doit disposer d'une interface lui permettant de visualiser et gérer les cas de conflit total, de visualiser et valider l'auto-

évaluation, d'évaluer les résultats par comparaison de deux méthodes d'appariement différentes ou d'évaluer les résultats en analysant si des représentations considérées par le processus d'appariement comme homologues, équivalentes ou incohérentes [Sheeren *et al.*, 2004]. A notre avis, une option permettant à l'utilisateur de corriger d'éventuelles erreurs d'appariement doit être prévue.

Un exemple d'interface dédiée à l'appariement interactif est illustré sur la Figure 131.

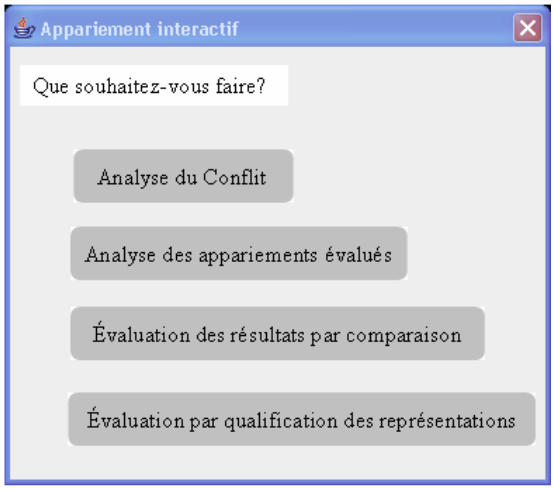


Figure 131. Interface pour l'appariement interactif

Les types d'évaluation que nous pourrions proposer sont les suivants :

- visualisation et analyse du conflit : l'utilisateur pourrait visualiser les cas de conflit un par un, puis après analyse du contexte du conflit il pourrait apparier l'objet qui est en conflit manuellement,
- visualisation et analyse des appariements évalués : s'il le souhaite, l'utilisateur pourrait visualiser les appariements auto-évalués afin de valider ou pas l'auto-évaluation du processus. Cette étape nous semble importante, puisqu'il arrive des cas où un appariement est auto-évalué comme sûr mais est faux et à l'inverse un appariement peut être évalué comme incertain mais est juste,
- évaluation des résultats en comparant les résultats de deux méthodes d'appariement : afin de faciliter l'étape d'évaluation qui dans certains cas peut devenir fastidieuse, l'utilisateur pourrait comparer les résultats d'appariement issus de deux méthodes d'appariement. Nous pourrions considérer que les appariements trouvés par les deux méthodes sont justes, et ainsi les valider sans faire une évaluation interactive. Par contre les appariements différents, (par exemple un objet a été apparié par une méthode mais pas une autre ou un objet a été apparié avec des objets différents par les deux méthodes, etc.), seront évalués manuellement afin de les valider.

  Deux méthodes différentes d'appariement peuvent être un appariement basé sur la fusion de quatre critères d'appariement et un autre réalisé en utilisant cinq critères.

- enfin, une autre solution pour faciliter l'étape d'évaluation est celle proposé par David Sheeren dans sa thèse [Sheeren, 2005]. Elle consiste à analyser les couples d'objets appariés afin de qualifier chaque représentation composant le couple d'équivalente ou d'incohérente. Selon [Sheeren, 2005] une représentation est dite équivalente si la représentation est conforme aux spécifications de la base de données et elle est appelée

incohérente dans le cas contraire. Un exemple d'interface issue de [Sheeren, 2005] est illustré en Figure 132.

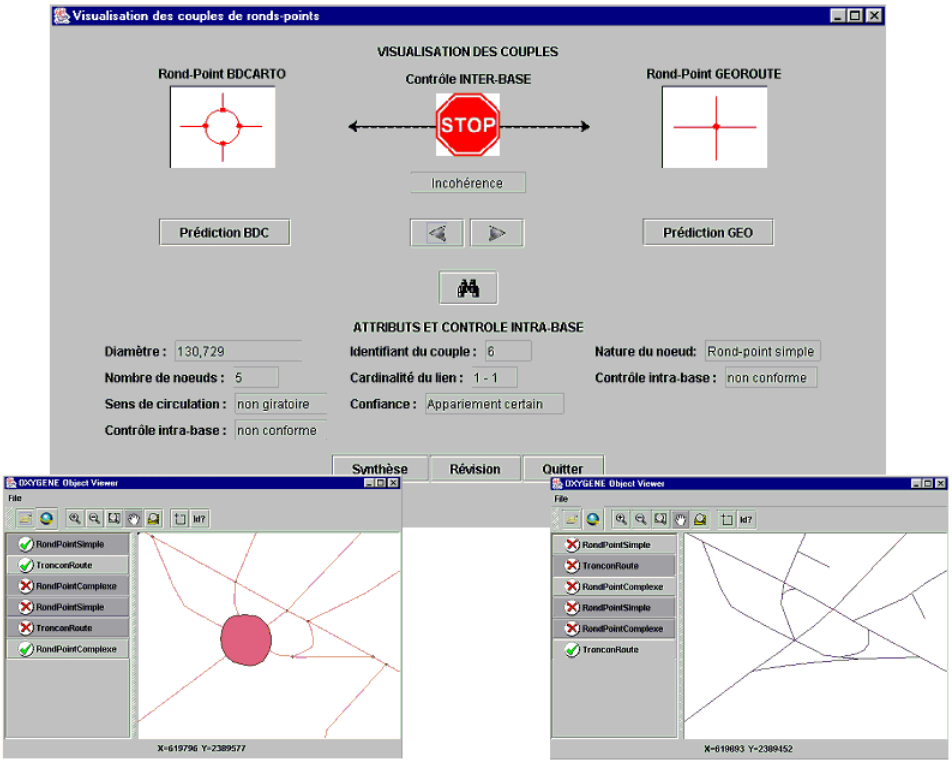


Figure 132. Interface de contrôle des couples d'objets appariés, d'après [Sheeren, 2005]

E.6 Qualification des résultats d'appariement

Enfin, la dernière étape consiste à qualifier le processus d'appariement en fonction de la qualité des résultats obtenus. Cette qualification pourrait s'appuyer sur les mesure de rappel et de précision (voir le paragraphe D.2.3). Si l'utilisateur n'est pas satisfait des résultats obtenus, il pourrait relancer un nouveau processus d'appariement en utilisant d'autres critères d'appariement ou en changeant des paramètres. Dans le cas contraire, il quitte après avoir fait une sauvegarde des résultats.

E.7 Conclusion

Nous avons présenté dans ce chapitre des pistes vers un système générique d'appariement de données géographiques qui s'appuie sur des interfaces, permettant une adaptabilité aux données à apparier. Le processus d'appariement commencerait par le chargement des données à apparier et se terminerait par la qualification finale des résultats d'appariement. Nous rappelons que les interfaces que nous avons proposées dans ce chapitre ne sont ni implémentées, ni testées.

Nous avons montré dans ce chapitre que le processus d'appariement que nous avons proposé au chapitre C possède la capacité à s'adapter aux types de données, aux niveaux de détail des jeux de données ainsi qu'à de nouvelles connaissances.

Nous avons également proposé, par chaque étape d'appariement exposée, des améliorations au processus ainsi que des conseils relatifs à la détermination des paramètres et à la modélisation des connaissances.

Conclusion et perspectives

Conclusion et perspectives

Rappel des objectifs

Dans le contexte de l'intégration de bases de données géographiques, nous avons abordé la problématique de l'appariement de données géographiques. Notre objectif était de proposer un processus d'appariement de données flexible et adaptable à différentes données géographiques, qui permette de prendre en compte les imperfections présentes dans les données géographiques.

Approche d'appariement proposée

Afin d'atteindre notre objectif nous avons proposé une approche d'appariement de données géographiques guidée par des connaissances, c'est-à-dire qui les représente explicitement. Les connaissances sont issues des données elles-mêmes, des spécifications des bases de données géographiques et des experts.

Nous avons vu au chapitre A que l'appariement de données n'est pas immédiat en raison de la complexité des données géographiques représentées dans les bases de données géographiques, et que celui-ci s'appuie en général sur différents critères d'appariement. Nous avons vu également que les connaissances qui permettent de définir les critères d'appariement ne sont pas parfaites, pouvant présenter des imperfections. Aux chapitres A et B nous avons présenté différentes taxonomies de l'imperfection et nous avons choisi d'utiliser la taxonomie illustrée en Figure 41 [Smets, 1997 ; Bouchon-Meunier, 1995]. D'après cette taxonomie l'imperfection fait appel à trois concepts : imprécision, incertitude et incomplétude.

Afin de représenter explicitement les connaissances dont nous disposons, ainsi que les éventuelles imperfections qu'elles peuvent présenter, nous avons basé notre approche d'appariement de données géographiques sur la théorie des fonctions de croyance. Les avantages que nous avons présentés dans le chapitre B, au paragraphe B.3.1 nous semblent justifier l'utilisation de la théorie des fonctions de croyance par rapport à nos besoins.

Nous rappelons en Figure 133 le processus d'appariement que nous avons proposé (voir le Chapitre C). Il est composé de cinq étapes principales : la sélection des candidats, l'initialisation des masses de croyance, la fusion des critères d'appariement pour chaque candidat, la fusion des candidats à l'appariement et la décision.

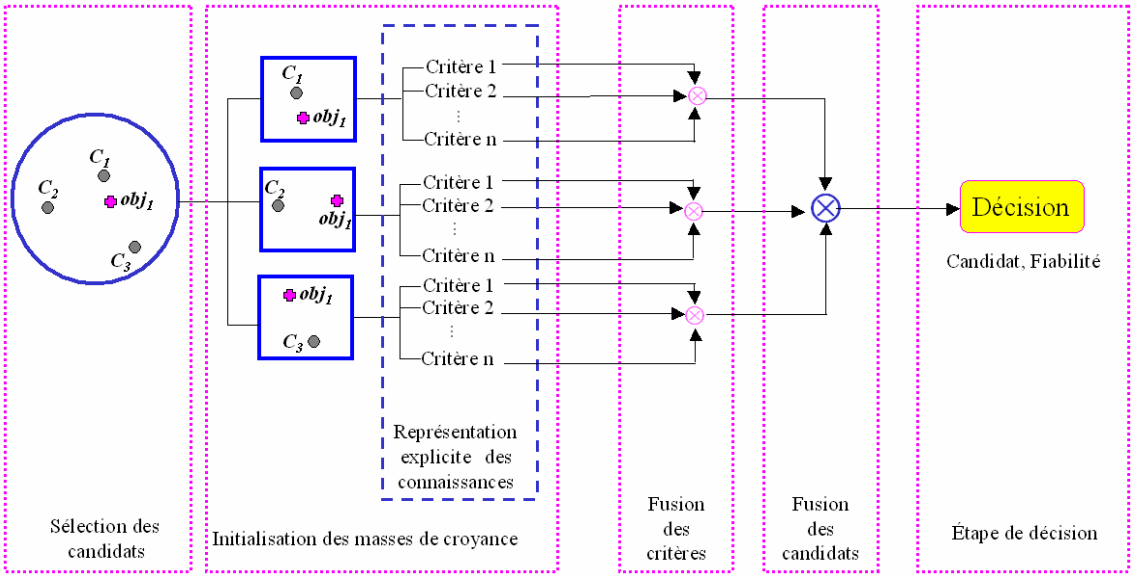


Figure 133. Processus d'appariement de données géographiques proposé

Avantages du processus d'appariement proposé

Si plusieurs approches d'appariement existent dans la littérature, nous estimons que notre approche présente les avantages suivants :

1. Processus d'appariement applicable à des données géographiques variées.

   Un premier avantage de notre approche réside dans le fait que notre processus d'appariement peut être appliqué d'une part aux trois types de représentation (points, lignes ou surfaces) et aux différents thèmes (routier, hydrographique, bâtiments, occupation du sol, etc.) et d'autre part aux jeux de données ayant le même niveau de détail ou des niveaux de détail différents. Cela est possible grâce au fait que les critères d'appariement sont définis indépendamment les uns des autres en s'appuyant sur différentes caractéristiques des données géographiques et qu'ils sont fusionnés en même temps grâce à l'opérateur de Dempster [Dempster, 1967].

2. Processus d'appariement adaptable, voire évolutif.

   L'approche proposée ne s'arrête pas uniquement à quelques critères d'appariement comme ceux que nous avons proposés dans le chapitre C. Elle nous permet d'une part d'ajouter un critère d'appariement sans avoir à retoucher les critères définis précédemment, et d'autre part pour chaque critère nous pouvons définir les fonctions de croyance adaptées. Comme nous l'avons vu au chapitre E (cf. E.4.3), il est possible d'en rajouter autant que l'on souhaite en fonction des caractéristiques et de la complexité des données géographiques à apparier.

3. Représentation explicite des connaissances et de leurs imperfections.

   Un troisième avantage de notre approche est la représentation explicite des connaissances et de leurs imperfections utilisées dans le processus d'appariement pour définir les critères d'appariement. Dans le cadre de cette thèse nous nous sommes limités à trois types de connaissance : celles issues des données géographiques qui permettent de calculer les mesures de distance entre les objets géographiques des deux jeux de données, celles issues des spécifications des bases de données géographiques qui sont utilisées pour définir les

seuils, et enfin les connaissances issues des experts qui nous permettent de définir les masses de croyance.

De plus, l'approche est flexible car elle permet, dans le cadre du même formalisme, de représenter des connaissances provenant de sources différentes ayant des niveaux de fiabilité différents (géométrie, relations spatiales, sémantique) ou d'exprimer des conditions nécessaires mais non suffisantes (par exemple : « deux objets sont homologues s'ils ont la même nature mais les objets qui ont la même nature ne sont pas tous homologues » ou « si les voisins d'un objet géographique O_1 sont appariés, alors l'objet O_1 est aussi apparié »).

Nous avons également modélisé les trois types d'imperfection présents dans les connaissances. Ainsi, l'imprécision est représentée formellement par les masses de croyance attribuées à une proposition, c'est-à-dire une union d'hypothèses simples, l'incertitude est représentée explicitement par les masses de croyance partielles dans différentes hypothèses, à condition que la somme soit égale à 1. Enfin l'incomplétude est représentée explicitement par l'ignorance, c'est-à-dire qu'une masse de croyance est attribuée à toutes les hypothèses composant le cadre de discernement.

Afin d'initialiser les masses de croyance, nous avons utilisé le modèle d'Appriou [Appriou, 1991] où un jeu de masses est qualifié par trois éléments focaux ($appC_i$, $\neg appC_i$, Θ). Par conséquent, nous avons pu exprimer explicitement la connaissance complète ($appC_i$), la connaissance partielle ($\neg appC_i$) et l'ignorance (Θ).

4. Représentation du doute et de l'incomplétude.

   A la différence des méthodes statistiques traditionnelles, notre méthode d'appariement permet, grâce à l'ignorance Θ, d'utiliser des connaissances partielles, mêmes incomplètes et de ne pas tirer une conclusion lorsqu'un doute existe.

5. Pondération de l'appariement en fonction de critères et de distances.

   Afin d'apparier les données géographiques, plusieurs critères d'appariement peuvent être utilisés. Les critères d'appariement sont basés sur différentes propriétés des données géographiques telles que la géométrie, l'information attributaire, l'information sémantique, les relations spatiales, les informations implicites (la forme, l'orientation, l'angle, etc.). Nous avons vu au chapitre A que les critères d'appariement sont habituellement basés sur des règles définies à partir de mesures issues de la comparaison de différentes propriétés des données, et qu'ils sont appliqués soit les uns après les autres (enchaînement des critères) soit en parallèle (combinaison des critères). Cette dernière solution s'appuie sur des pondérations en fonction uniquement de l'importance du critère d'appariement.

   Dans notre approche, la définition d'un critère d'appariement consiste à initialiser les masses de croyance attribuées aux trois hypothèses $appC_i$, $\neg appC_i$, Θ. Pour ce faire, nous utilisons deux pondérations :

   - une pondération des critères d'appariement qui nous permet d'affaiblir les critères en fonction de leur importance et du type de connaissance utilisée. Par exemple, si deux objets ont le même toponyme et la même nature, les deux critères toponymique et sémantique attribuent des masses de croyance importantes à l'hypothèse que les deux objets sont homologues, $appC_i$. Par contre, étant données les connaissances que ces deux critères modélisent et du fait que le critère sémantique exprime des conditions nécessaires mais non suffisantes, la masse de croyance attribuée à l'hypothèse que les objets sont homologues n'ont pas la même

valeur malgré le fait que les distances toponymique et sémantique sont égales à 0. Par conséquent, la masse de croyance attribuée par le critère toponymique à l'hypothèse $appC_i$ peut être égale à 1 tandis que la masse de croyance attribuée par le critère sémantique à la même hypothèse peut être égale à 0,5,

- une pondération pour la valeur d'une distance qui donne la masse de croyance. Comme nous l'avons dit dans le chapitre A, l'appariement de données est basé sur la notion de ressemblance. Si deux objets se ressemblent fortement (par exemple l'écart de position est très faible), nous croyons qu'ils sont homologues, donc nous attribuons à l'hypothèse $appC_i$ une masse de croyance importante. Par contre, si les objets sont très éloignés (écart de position élevé), nous croyons que les deux objets ne sont pas homologues et par conséquent la masse de croyance attribuée à l'hypothèse $appC_i$ est très faible.

Modèle mis en œuvre et implémenté

Notre approche d'appariement de données a été mise en œuvre sur la plate-forme GeOxygene du laboratoire COGIT (voir le chapitre D.1.3.2). Afin de valider notre approche, nous avons réalisé deux études. La première consiste à apparier deux jeux de données représentant les points remarquables du relief, ayant des niveaux de détail différents et issus des bases de données BDCARTO et BDTOPO de l'IGN (cf. chapitre D.2). La deuxième étude consiste à apparier deux jeux de données représentant les réseaux routiers, ayant des niveaux de détail différents et issus de la base de données BDCARTO de l'IGN et de la base de données MultiNet de TeleAtlas. Les résultats d'appariement ont été évalués d'une manière interactive en termes de précision et de rappel. Ils ont été également comparés aux résultats d'appariement obtenus en utilisant d'autres méthodes d'appariement de données.

Les études que nous avons réalisées nous ont permis d'une part de valider notre approche, et d'autre part de voir les avantages de l'approche proposée ainsi que ses limites. Ces dernières ont été identifiées au chapitre D et quelques pistes d'amélioration ont été proposées au chapitre E.

Nous avons pu également constater que les critères d'appariement et plus particulièrement la forme des courbes permettant l'initialisation des masses de croyance que nous avons illustrées à titre d'exemple au chapitre C ne dépendent pas du type de données. Par exemple, le même critère d'écart de position ou sémantique a été utilisé à la fois pour les points remarquables du relief et pour les réseaux. Les seules caractéristiques qui changent en fonction du type des données utilisées sont la manière dont nous déterminons les valeurs des distances, des paramètres et des seuils.

A travers des exemples de résultats obtenus, nous avons montré l'importance de la fusion de plusieurs critères d'appariement même si les connaissances utilisées ne sont pas parfaites.

Même si notre approche a été testée seulement sur des données géographiques ponctuelles et linéaires et pour des jeux de données ayant des niveaux de détails différents, nous pensons qu'elle peut être utilisée également pour des données surfaciques et pour des jeux de données ayant le même niveau de détail. De plus, nous estimons que l'appariement des jeux de données ayant le même niveau de détail peut donner une meilleure satisfaction au niveau des résultats obtenus en raison de la ressemblance plus importante des jeux de données.

Perspectives

Nous avons vu que notre approche a néanmoins des limites liées à la cardinalité, au seuillage ou à l'ambiguïté de l'interprétation. Des pistes d'amélioration ont été proposées plus

en détail au chapitre E. Nous les rappelons brièvement ci-dessous et nous en ajoutons d'autres :

- vers une meilleure solution pour gérer le type de correspondance n : m.

  Comme nous l'avons vu au chapitre C, notre approche ne permet pas de définir directement des liens d'appariement du type n : m. Afin d'améliorer cette limite nous proposons deux solutions. La première consiste à réaliser deux appariements de données, en changeant le jeu de données de référence et ensuite de regrouper les liens d'appariement obtenus. La deuxième solution est d'utiliser à la place de la fonction pignistique soit la fonction de crédibilité, soit la fonction de plausibilité, afin de prendre la décision finale et de transformer le OU des hypothèses composées en ET.

- vers une meilleure initialisation des masses de croyance.

  Afin d'améliorer l'étape d'initialisation des masses de croyance, nous pouvons imaginer une approche multi-résolution. Il s'agit de définir un nombre de n critères au départ. Puis nous pouvons comparer les résultats obtenus avec l'ensemble de toutes les combinaisons possibles, c'est-à-dire 2 critères, 3 critères,… n critères. Ainsi, les masses de croyance fixées peuvent être affinées et nous pouvons faire une analyse statistique sur l'importance des critères.

- vers de meilleures distances utilisées pour définir les critères d'appariement.

  Dans les expérimentations réalisées, nous avons fait des choix pour les mesures utilisées. Une mesure est plus ou moins pertinente en fonction des propriétés utilisées. Nous considérons que la mesure qui permet de comparer la sémantique des objets géographiques doit être améliorée. Si pour les points remarquables du relief cette mesure arrive à donner une information pertinente relative à la ressemblance de la nature des objets, pour les réseaux routiers la mesure sémantique n'est pas toujours fiable et elle nécessite une bonne connaissance des données à apparier (les classifications utilisées dans les deux bases de données). Afin d'améliorer cette mesure, une piste d'amélioration est d'abord de mettre en correspondance les classes des deux bases de données, même imparfaitement, et de réaliser une étape d'analyse de ces résultats pour calculer des distances sémantiques. Une solution pour la mise en correspondance des classes est de définir une ontologie ou une taxonomie pour chaque base de données et ensuite de les aligner (apparier) entre elles, ou d'utiliser une ontologie de référence. Ainsi, les distances sémantiques pourront être calculées au niveau de l'ontologie commune. En rapport à ce sujet, une thèse est en cours au laboratoire COGIT [Abadie et Mustière, 2008].

- vers l'automatisation des seuils utilisés pour définir les critères d'appariement.

  Dans nos tests expérimentaux, les seuils de sélection des candidats et ceux utilisés dans les critères d'appariement ont été déterminés d'une manière empirique. Afin d'automatiser cet aspect, nous pouvons envisager d'utiliser des méthodes de fouille de données [Sheeren, 2005] ou les courbes ROC [Champion, 2007].

- vers une meilleure auto-évaluation des résultats d'appariement.

  Nous avons vu dans le chapitre D qu'en fonction des données utilisées il existe une bonne corrélation (pour les points remarquables du relief) ou une moins bonne corrélation (pour les réseaux routiers) entre l'auto-évaluation des résultats que nous avons proposée et l'évaluation interactive. Rappelons que l'auto-évaluation que nous avons proposée est basée sur l'écart entre le premier et le deuxième maximum de la probabilité pignistique. Nous avons remarqué qu'elle est dépendante du nombre de

candidats à l'appariement. Si aucun candidat ne se détache, plus le nombre de candidats est élevé, plus l'écart entre le premier et le deuxième maximum est faible. Par conséquent, l'appariement sera qualifié d'incertain.

Il nous semble nécessaire d'améliorer l'auto-évaluation du processus. Pour ce faire, nous pouvons envisager soit d'utiliser une mesure calculée à partir de la fonction de plausibilité [Appriou, 1999], soit d'utiliser un intervalle d'incertitude basé sur les fonctions de crédibilité et de plausibilité.

- vers une meilleure convergence du processus.

  Nous avons montré dans le chapitre D.3.2.4 que dans des cas exceptionnels, notre processus qui analyse itérativement les résultats d'appariement d'une étape pour définir le critère de voisinage à l'étape suivante ne converge pas. Etant donné que la convergence ou la non-convergence du processus se remarque après trois ou quatre passes du processus, nous considérons qu'arrêter le processus après trois ou quatre passes peut être une solution envisageable. Afin d'atteindre la convergence, nous pouvons aussi envisager d'affaiblir le critère de voisinage à chaque passe jusqu'au moment de la convergence. Une autre piste d'amélioration peut être de réaliser quelques passes du processus, et si la convergence n'est pas atteinte, de mettre des avertissements sur les résultats qui changent, puis de les vérifier d'une manière interactive.

Pour conclure

L'appariement de données géographiques est un outil qui répond à de nombreux besoins des producteurs de données et des utilisateurs. L'automatisation du processus d'appariement leur permet d'atteindre plus facilement le but final, qu'il s'agisse d'intégrer des bases de données géographiques, d'étudier la qualité des données ou de les mettre à jour. Nous considérons que l'approche d'appariement de données géographiques que nous avons proposée dans cette thèse constitue une avancée pour l'automatisation du processus.

Bibliographie

[Abadie *et al.*, 2006] Abadie N., Gesbert N. et Mustière S. *Création d'une ontologie à partir des spécifications textuelles pour l'intégration des bases de données géographiques*. In : *Actes de Ingénierie des Connaissances*, Nantes, 2006

[Abadie *et al.*, 2007] Abadie, N., Olteanu, A-M. et Mustière, S. Comparaison de la nature d'objets géographiques. In : *Actes de Ingénierie des connaissances*, *la journée OGHS «Ontologies et Gestion de l' Hétérogénéité Sémantique»*, Grenoble, 3 juillet 2007

[Abadie et Mustière, 2008] Abadie, N. et Mustière, S. Création d'une taxonomie géographique à partir des spécifications de bases de données. In: *Actes de SAGEO'08*, Montepellier, 2008

[Ahlqvist *et al.*, 2003] Ahlqvist, O., Keukelaar, J. et Oukrir, K. *Rough and fuzzy geographical data integration*, *International Journal of Geographical Information Science*, vol.17, n°3, 2003, p. 223-234

[Akyürek et Kıvanç, 2006] Akyürek, Z. et Kıvanç, O. *A fuzzy-based tool for spatial reasoning: A case study on soil erosion hazard prediction.* In : *Proceedings of the 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, Lisbonne, 7-9 juillet 2006, p. 719-729

[Alt et Godau, 1995] Alt, H. et Godau, M. *Computing the Fréchet distance between two polygonal curves*. *International Journal of Computational Geometry and Applications*, 1995, vol.5, n°1/2, p. 75-91

[Anders et al., 2007] Anders, K.-H., Sester, M. et Bobrich J. *Incremental update in an MRDB.* In : *Proceedings of the 20th International Cartographical Conference*, Moscou, 5-9 août 2007

[Anders et Bildirici, 2004] Anders, K.H. et Bildirici, I.Ö. *MRDB approach to handle and visualise multiple DLM'S in a consistent*. In : *Proceedings of the 20th ISPRS Congress*, 12-23 juillet, Istanbul, 2004

[Appriou 1991] Appriou, A. *Probabilités et incertitudes en fusion de données multi-senseurs*. *Revue Scientifique de Technique de la Défense*, 1991, n°11, p. 27-40

[Appriou, 1999] Appriou, A. *Multisensor signal processing in the framework of the theory of evidence*. *Application of Mathematical Signal Processing Techniques to Mission Systems*, 1999, vol. 5, n°1, p. 5–31

[Arkin *et al*., 1991] Arkin, E.M., Chew, L.P., Huttenlocher, D.P., Kedem, K., Kedem, K. et Mitchel, J.S.B. *An efficient computable metric for comparing polygonal shapes*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1991, vol. 13, n° 3, p. 209-216

[Badard, 1998] Badard, T. *Extraction des mises à jour dans les BDG – De l'utilisation de méthodes d'appariement*. *Revue Internationale de géomatique*, 1998, n°8 (1-2), p. 121–147

[Badard, 2000] Badard, T. *Propagation des mises à jour dans les bases de données géographiques multi-représentations par analyse des changements géographiques.* Thèse de Doctorat, Université Marne-la-Vallée, 2000

[Badard et Lemarié, 2002] Badard, T. et Lemarié, C. Associer des données : l'appariement In : Ruas, A (éd.), *Généralisation et représentation multiple*. Paris : Lavoisier, 2002, p. 163-183

[Badard et Braun, 2003] Badard, T. et Braun, A. *OXYGENE : an open framework for the deployment of geographic web services.* In : *Proceedings of the 18th ICC,* 2003, p. 994-1003

[Badard et Braun, 2004] Badard, T. et Braun, A. *OXYGENE : une plate-forme inter-opérable pour le déploiement de services Web géographiques*. *Bulletin d'Information scientifique et technique de l'IGN*, n° 74, 2004, p. 113-120

[Bard *et al.,* 2003] Bard, S., Bouchon-Meunier, B., Ruas, A. et Detyniecki, M. *Gestion des connaissances imprécises pour évaluer la généralisation cartographique*. In : *Actes de Logique Floue et Applications (LFA)*, 2003, Tours

[Bard, 2004] Bard S., *Méthode d'évaluation de la qualité de données géographiques généralisées - Application aux données urbaines*. Thèse de doctorat, Université Paris 6, 2004

[Barnett, 1981] Barnett, J. A. Computational methods for a mathematical theory of evidence, In : *Proceedings of IJCAI*, 1981, p. 868-875

[BDCARTO, 2004] *Spécifications de contenu de la BDCARTO*, version 2.5., IGN, 2004

[BDTOPO, 2001] *Spécifications de contenu de la BDTOPO Pays*, version 1.1., IGN, 2001

[BDTopoPays, 2002] *Spécifications de contenu de la BDTOPO Pays*, version 1.2., IGN, 2002

[BDCARTO-BDTOPO, 2005] Duchêne, C., Grosso, E., Mustière, S. et Touya, G. *Automatisation de l'appariement de la BDCARTO avec la BDTOPO Pays et de la généralisation de la BDCARTO à partir de la BDTOPO Pays, en vue de l'intégration de ces bases*. Étude interne du laboratoire COGIT, IGN, 2005

[Beeri *et al.,* 2004] Beeri, C., Kanza, Y., Safra, E. et Sagiv, Y. *Object Fusion in Geographic Information Systems*. In : *Proceedings of the 30th VLDB Conference*, Toronto, 2004

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

[Benferhat *et al.,* 2003] Benferhat, S., Dubois, D., Kaci, S. et Prade, H. Fusion de données numériques et d'informations symboliques. In : Bloch, I. et Cholvy, L. (éd.), *Technique et sciences informatiques*. Lavoisier, 2003, p. 1035-1064

[Besl et McKay, 1992] Besl, P.J. et McKay, N.D. *A method for registration of 3D shapes*. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 1992, vol. 14, p. 239–254.

[Blasby *et al.*, 2004] Blasby, D., Davis, M., Kim, D. et Ramsey, P. *GIS conflation using open source tools*. http://www.jump-project.org/assets/JUMP_Conflation_Whitepaper.pdf

[Bloch, 1996] Bloch, I. *Some aspect of Dempster-Shafer evidence theory for classification of multi-modality medical images taking partial volume effect into account*. *Pattern Recognition Letters*, 1996, n°17, p. 905-919

[Bonin et Rousseaux, 2005] Bonin, O. et Rousseaux, F. *Digital terrain model computation from contour lines: How to derive quality information from artifact analysis*. *GeoInformatica*, 2005, vol. 9, n° 3, p. 253-268

[Bouchon-Meunier, 1995] Bouchon-Meunier, B. *La logique floue et ses applications*. Paris : Addison-Wesley France, 1995, p. 257

[Bouziani, 2003] Bouziani, M. *Définition d'une méthode d'extraction des mises à jour de l'information spatiale dans un réseau routier en milieu urbain*. In : *Proceedings of the 2nd FIG Regional Conference,* Marrakech, 2-5 décembre 2003

[Bouziani et Pouliot, 2008] Bouziani, M. et Poliot, J. *Optimisation de la mise à jour de bases de données géospatiales*. *Revue Internationale de géomatique*, 2008, n°18 (1), p. 113-137

[Brown, 1998] Brown, D.G. *Classification and boundary vagueness in mapping presettlement forest types*. *IJGIS*, vol. 12, n° 2, 1998

[Bruns et Egenhofer, 1996] Bruns, H.T. et Egenhofer, M. *Similarity of spatial scenes*. In : *Proceedings of the 7th International Symposium on Spatial Data Handling,* Delft, août 1996. Londres : Taylor & Francis, p. 173-184

[Buttenfield, 1991] Buttenfield, B. A rule for describing line feature geometry. In : Buttenfield, B. et McMaster, R. (éd.), *Map generalization : making rules for knowledge representation*. Harlow (Grande-Bretagne) : Longman Scientific and Technical, 1991, p. 150-171

[Capelle, 2003] Capelle, A.S. *Segmentation d'images IRM multi-échos tridimensionnelles pour la détection des tumeurs cérébrales par la théorie de l'évidence*. Thèse de Doctorat, Université de Poitiers, 2003

[Champion, 2007] Champion, N. *2D Building Change Detection from High Resolution Aerial Images and Correlation Digital Surface Models*. In : *Proceedings of ISPRS Conference Photogrammetric Image Analysis (PIA)*, Munich, septembre 2007, vol. 36, p. 197-202

[Chen *et al.*, 2006] Chen, C.C., Knoblock, C.A. et Shahabi, C. *Automatically Conflating Road vector Data with Orthoimagery*. *GeoInformatica*, 2006, vol. 10, n°4, p. 495-530

[Chen *et al.*, 2008] Chen, C.C., Knoblock, C.A. et Shahabi, C. *Automatically and accurately conflating raster maps with orthoimagery*. *GeoInformatica*, 2008, vol. 12, n°3, p. 377-409

[Chitroub, 2004] Chitroub, S. *Combinaison de classifieurs : une approche pour l'amélioration de la classification d'images multisources / multidates de télédétection*. *Télédétection*, 2004, vol. 4, n°3, p. 289-301

[Chrisman, 1982] Chrisman, N.R. *A theory of cartography error and its measurement in digital database*. In : *Proceedings of AutoCarto*, 1982, p. 159-168

[Clementini *et al.*, 1993] Clementini, E., Di Felice, P. et van Oosterom, P. *A small set of formal topological relationships suitable for end-user interaction*. In : Abel, D. et Ooi, B.C. (éd.), *Proceedings of the 3rd International Symposium of Advances in spatial databases*, Singapour, 1993. Springer, p. 277-295

[Clodoveu *et al.*, 2007] Clodoveu, A. D. et Fonseca F. *Assessing the Certainty of Locations Produced by an Address Geocoding System*. *GeoInformatica,* 2007, vol. 11, n°1, p. 103-129

[Cohen, 1985] Cohen, P.R. *Heuristic reasoning about uncertainty: an artificial intelligence approach*. Boston : Pitman Advanced Publishing, 1985

[Cohen et Guibas, 1997] Cohen, S.D. et Guibas, L.J. *Partial matching of planar polylines under similarity transformations*. In : *Proceeding of the 8th Annual ACMSIAM Symposium on Discrete Algorithms*, 1997, p. 777-786

[Cohen *et al.*, 1987] Cohen, M.S., Laskey, K.B. et Ulvila, J.W. *The management of uncertainty in intelligence data*: *A self-reconciling evidential database.* Falls Church (Etats-Unis) : Decision Science Consortium, Inc., juin 1987

[Cohen *et al.,* 2003] Cohen, W.W., Ravikumar P. et Fienberg, S.E. *A Comparison of String Distance Metrics for Name-Matching Tasks*. In : *Proceedings of the IJCAI*, Acapulco, Mexique, 9-10 août 2003, p. 73-78

[Cohn et Gotts, 1994] Cohn, A. et Gotts, N. *Spatial regions with undetermined boundaries*. In : *Proceedings of Gaithesburg Workshop on GIS*, ACM, 1994

[Colot, 2000] Colot, O. *Systèmes de perception d'informations incertaines - Application au diagnostic médical*, Mémoire de HDR, Université de Rouen, 2000

[Comber *et al.*, 2004] Comber A., Ficher, P. et Wadsworth, R. *Integrating land cover data with different ontologies : identifying change from inconsistency*. *IJGIS*, 2004, vol.18, n°7, p. 691-708

[Comber *et al.*, 2005a] Comber, A., Wadsworth, R. et Fisher, P.F. Méthodes de raisonnement pour manipuler l'information incertaine en occupation de sol. In : *Qualité de l'information géographique*. Paris : Hermès et Lavoisier, 2005, p. 153-167

[Comber *et al.*, 2005b] Comber, A., Wadsworth, R. et Fisher, P.F. Nature de l'incertitude pour les données spatiales. In : *Qualité de l'information géographique*. Paris : Hermès et Lavoisier, 2005, p. 49-64

[Comber, 2007] Comber, A., Fisher, P.F. et Brown, A. *Uncertainty, vagueness and indiscernibility : the impact of spatial scale in relation to the landscape elements*. In : *Proceedings of Spatial Data Quality, Enschede,* juin 2007

[Corgne, 2004] Corgne, S. *Modélisation prédictive de l'occupation des sols en contexte agricole intensif : Application à la couverture hivernale des sols en Bretagne.* Thèse de doctorat, Université de Rennes 2, 2004

[Cowell, 1999] Cowell, R. G. *Parameter estimation from incomplete data for Bayesian networks*. In : D. Heckerman and J. Whittaker (éd.), *Proceedings of the 7th International Workshop on Artificial Intelligence and Statistics*, San Francisco, 1999, p. 193-196

[Dassonville *et al.*, 2002] Dassonville, L., Vauglin, F., Jakobsson, A. et Luzet, C. Quality Management, Data Quality and users, Metadata for Geographical Information. In : Shi, W., Fisher, P.F. et Goodchild, M.F (éd.), *Spatial Data Quality*. Taylor & Francis, 2002, p. 13

[David, 1988] David, B. *Map is better than graph to store topology*. Poster présenté à Euro-Carto Seven, Enschede (Pays-Bas), septembre 1988

[David et Fasquel, 1997] David, B. et Fasquel, P. *Qualité d'une bases de données géographiques : concepts et terminologie*. *Bulletin d'information de l'IGN*, 1997, n° 67

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

[Denœux, 1995] Denoeux, T. *A k-nearest neighbour classification rule based on Dempster-Shafer theory*. *IEEE Transactions on Systems, Man and Cybernetics*, 1995, vol. 25, n° 5, p. 804–813

[Denœux et Bjanger, 2000] Denoeux, T. et Bjanger, M.S. *Induction of decision trees from partially classified data using belief functions*. In : *Proceedings of SMC*, Nashville, 2000, p. 2923–2928

[Denœux et Masson, 2003] Denoeux, T. et Masson, M. Clustering of proximity data using belief functions. In : Bouchon-Meunier, B., Foulloy, L., et Yager, R.R. (éd.), *Intelligent Systems for Information Processing: From representation to applications*. Amsterdam : Elsevier, 2003, p. 291-302

[Denœux, 2006] Denoeux, T. *The cautious rule of combination for belief functions and some extensions*. In : *Proceedings of the 9th International Conference on Information Fusion (FUSION'2006)*, Florence, 10-14 juillet 2006

[Devillers, 2004] Devillers, R. *Conception d'un système multidimensionnel d'information sur la qualité des données géospatiales*. Thèse de Doctorat, Université Marne-la-vallée, décembre 2004

[Devillers et Jeansoulin, 2005] Devillers, R. et Jeansoulin R. *Qualité de l'information géographique*. Paris : Hermès et Lavoisier, 2005

[Devogele *et al.,* 1998] Devogele, T., Parent, C. et Spaccapietra, S. *On spatial database integration*. *International Journal of Geographical Information Science*, 1998, n°12(4), p. 335-352

[Devogele, 1997] Devogele, T. *Processus d'intégration et d'appariement de bases de données Géographiques*, *Applications à une base de données routières multi-échelles*, Thèse de Doctorat, Université de Versailles, 1997

[Devogele *et al.*, 2002] Devogele, T., Badard, T. et Libourel, T. La problématique de la représentation multiple. In : Ruas, A. (éd.), *Généralisation et représentation multiple*. Paris : Lavoisier, 2002, 390 p.

[Diaz *et al.*, 2007] Diaz, J., Rifqi, M. et Bouchon-Meunier, B. *QCM évidentiels pour le diagnostic des apprenants*. In : *Actes des Rencontres francophones sur la Logique Floue et ses Applications*, Nîmes, 22-23 novembre 2007, p. 149-156

[Dilo *et al.*, 2007] Dilo A., de By, R.A et Stein, A. *A system of types and operators for handling vague spatial objects. IJGIS,* 2007, vol. 21, n° 3-4, p. 397-426

[Do *et al.*, 2002] Do, H., Melnik, S. et Rahm, E. *Comparison of schema matching evaluations.* In : *Proceedings of the 2nd International Workshop on Web Databases (German Informatics Society),* 2002

[Doytsher *et al.*, 2001] Doytsher, Y., Filin, S. et Ezra, E. *Transformation of datasets in a linear-based map conflation framework. Surveying and Land Information Systems,* 2001, vol. 61, n°3, p. 165-175

[Dubois et Prade, 1985] Dubois, D., et Prade, H. *Théorie des possibilités : applications à la représentation des connaissances en informatique*. Paris : Masson, 1985

[Dubois et Prade, 1988] Dubois, D. et Prade, H. *Representation and combination of uncertainty with belief functions and possibility measures. Computer Intelligence,* 1988*,* vol. 4, p. 244-264

[Dubois et Prade, 1994] Dubois, D. et Prade, H. *Fusion d'informations imprécises. Traitement du signal*, 1994, vol. 11, n°6

[Duchêne et Cambier, 2003] Duchêne, C. et Cambier, C. *Généralisation cartographique avec des agents qui voient et communiquent*. In : *Actes des 9èmes Journées Francophones sur les Systèmes Multi-Agents (JFSMA'03)*, Hammamet (Tunisie), 2003, p. 13

[Duchêne, 2004] Duchêne, C. *Généralisation cartographique par agents communicants : le modèle CARTACOM - Application aux données topographiques en zone rurale-*. Thèse de doctorat, Université Paris 6 Pierre et Marie Curie, 2004

[Duckham et Worboys, 2005] Duckham, M. et Worboys, M. *An algebraic approach to automated geospatial information fusion*. *International Journal of Geographical Information Science*, 2005, n°19(5), p. 537-557

[Duckham *et al.,* 2001] Duckham, M., Mason, K., Stell, J. et Worboys, M. *A Formal Approach to Imperfection in Geographic Information*. *Computers, Environment and Urban Systems*, 2001, vol. 25, p. 89-103

[Dunkars, 2003] Dunkars, M. *Matching of Datasets*. In : *Proceedings of the 9th Scandinavian Research Conference on Geographical Information Science*. Espoo (Finland), 4-6 juin 2003, p. 67-78

[Dupouey *et al.*, 2007] Depouey J-L., Bachacou J., Cosserat R., Aberdam S., Vallauri D., Chappart G., et Corvisier de Villèle, M-A. Vers la réalisation d'une carte géoréférencée des forêts anciennes de France, In : *La revue du comité français de cartographie : Le monde des cartes, 2007*, n° 191, p. 85-98

[Dupuis, 2000] Dupuis, O. *Fusion entre les données ultrasonores et les images de radioscopie à haute résolution : Application au contrôle de cordon de soudure*. Thèse de Doctorat, Institut National des Sciences Appliquées de Lyon, 2000

[Dutton, 1992] Dutton, G. *Handling positional uncertainty in spatial databases*. In : *Proceedings of the Spatial Data Handling Symposium,* Charleston, août 1992, vol. 2, p. 460-469.

[Egenhofer, 1989] Egenhofer, M. J. *A formal definition of binary topological relationships*. In : *Proceedings of the 3rd International Conference on Foundations of Data Organization and Algorithms,* 1989. Springer-Verlag, vol. LNCS 367, p. 457–472

[Egenhofer et Herring, 1990] Egenhofer, M.J. et Herring, J.R. *A mathematical framework for the definition of topological relationships*. In : Brassel, K. et Kishimoto, H. (éd.), *Proceedings of the 4th International Symposium on Spatial Data Handling*, Zurich, 1990, p. 803-813

[Egenhofer et Franzosa, 1991] Egenhofer, M.J. et Franzosa, R. *Point-Set Topological Spatial Relations*. *International Journal of Geographical Information Systems*, 1991, vol. 5, n°2, p. 161-174

[Egenhofer *et al.*, 1994] Egenhofer M.J., Clementini E. et Di Felice P. Evaluating inconsistencies among multiple representations, In Proceedings of the 6th International Symposium on Spatial Data Handling, 1994, p. 901-920

[El Najjar, 2003] El Najjar, M.E. *Localisation dynamique d'un véhicule sur une carte routière numérique pour l'assistance à la conduite*. Thèse de doctorat, Université de Compiègne, 2003

[Euzenat et Shvaiko, 2007] Euzenat, J. et Shvaiko, P. *Ontology matching*. Springer, 2007, p. 333

[Erwig et Schneider, 1997] Erwig, M. et Schneider, M. *Vague regions*. In : *Proceedings of the 5th International Symposium Spatial Databases (SSD'97)*, Lecture Notes in Computer Science 1262. Berlin : Springer, 1997, p. 298-320

[Faux et Luthon, 2007] Faux, F. et Luthon, F. *Etude de différentes règles de fusion d'information couleur appliquées à la détection d'un visage en temps réel*. In : *Actes des Rencontres francophones sur la Logique Floue et ses Applications*, Nîmes, 22-23 novembre 2007, p. 9-16

[Fawcett, 2004] Fawcett, T. *Roc graphs: Notes and practical considerations for researchers*. Technical report, HP Laboratories, Etats-Unis, 2004

[Fisher, 2003] Fisher, P.F. *Models of uncertainty in spatial data. Geographical Information System*, vol. 1, 2ème édition, 2003, p. 191-203

[Fisher *et al.*, 2005] Fisher, P.F., Comber, A. et Wadsworth, R. Nature de l'incertitude pour les données spatiales. In : *Qualité de l'information géographique*. Paris : Hermès et Lavoisier, 2005, p. 49-64

[Fonseca *et al.*, 2002] Fonseca, F., Egenhofer, M. et Agouris, P. *Using ontologies for integrated geographic information systems. Transactions in Geographic Information Systems*, 2002, vol. 6, n°3

[Fonte et Lodwick, 2004] Fonte, C. et Lodwick, W. *Areas of fuzzy geographical entities. IJGIS*, 2004, vol.18(2), p.127-150

[Foody, 2006] Foody, G.M. *The Evaluation and comparison of thematic maps derived from remote sensing*. In : *Proceedings of the 7th International Symposium on Spatial Accuracy Assessement in Natural Resources and Environmental Sciences*, Lisbonne, 7-9 juillet 2006, p. 18-31

[Fritz et See, 2004] Fritz, S. et See, L. *Comparison of land cover maps using fuzzy agreement. IJGIS*, 2004, vol.19, n°1, p. 787-807

[Gabay, 1994] Gabay, Y. et Doytsher Y. Automatic adjustment of line maps. In : *Proceedings of the GIS/LIS Conference,* Phoneix, 25-27 October 1994, p. 333-341

[Gascône, 1997] Gascône, L. *Eléments de la logique floue.* Paris : Hermès, 1997, p. 251

[Gesbert, 2005] Gesbert, N. *Formalisation des spécifications de bases de données géographiques en vue de leur intégration*. Thèse de Doctorat, Université de Marne-la-Vallée, 2005

[Gomboši *et al.,* 2003] Gomboši, M., Žalik, B. et Krivograd, S. *Comparing two sets of polygons. International Journal of Geographical Information Science*, 2003, vol. 17 (5), p. 431-443

[Goodchild, 1995] Goodchild, M.F. *Sharing imperfect data*. In H.J. Onsrud and G. Rushton, editors, *Sharing Geographic Information*. New Brunswick, NJ: Rutgers University Press, 1995, p. 413–425

[Goodchild, 1991] Goodchild, M. F. *Issues of quality and uncertainty. Advances In Cartography*, edited by Muller J.C., London, Elsevier, 1991, p. 113-139

[Goodchild, 2005] Goodchild, M.F. *GIS and modeling overview.* In : Maguire, D.J., Batty, M. et Goodchild, M.F. (éd.), *GIS, Spatial Analysis, and Modeling*. Redlands, CA: ESRI Press, 2005, p. 1–18,

[Hunter et Goodchild, 1993] Hunter G.J. et Goodchild, M.F. *Managing uncertainty in spatial databases: putting theory into practice. Proceedings, URISA,* Atlanta, July 25–29, 1993, p : 1–14.

[Gösseln et Sester, 2003] Gösseln, G.V. et Sester, M. *Semantic and geometric Integration of geoscientific Data Sets with ATKIS - Applied to Geo-Objects from Geology and Soil Science*. In : *Proceedings of the ISPRS Commission IV Joint Workshop Challenges in Geospatial Analysis, Integration and Visualization II*, Stuttgart, 8-9 september 2003, p. 111-116

[Grosso, 2004] Grosso, E. *Etude des carrefours d'un réseau routier.* Rapport de stage, Université Paris 1, 37 p.

[Gruber, 1993] Gruber, T.R. Toward principles for the design of ontologies used for knowledge sharing. In : Guarino, N. et Poli, R. (éd.), *Formal ontology in conceptual analysis and knowledge representation*. Dordrecht : Kluwer academic, 1993

[Guptill et Morrison, 1995] Guptill, S.C. et Morrison, J.L. *Elements of data spatial quality*, Elvesier, 1995

[Hagen *et al.,* 2005] Hagen Z.A., Straatman, B. et Uljee, I., *Further developments of a fuzzy set map comparison approach. IJGIS*, août 2005, vol. 19, n°7, p. 769-785

[Hamming, 1950] Hamming, R. *Error detecting and error correcting codes*, *Technical Report 2*. *Bell System Technical Journal*, 1950

[Hansen, 2003] Hansen, H.S. *A Fuzzy Logic Approach to Urban Land-use Mapping*. In : *Proceedings of ScanGIS*, 2003

[Haunert, 2005] Haunert, J.H, *Link based Conflation of Geographic Datasets*. In : *Proceedings of the 8th ICA Workshop on Generalisation and Multiple Representation*, la Corogne, 7-8 juillet 2005

[Hazarika et Cohn, 2001] Hazarika, S.M. et Cohn, A.G. *A taxonomy of spatio-temporal vagueness: An alternative egg-yolk interpretation*. In : *Proceedings of the COSIT/FOIS Workshop on Spatial Vagueness, Uncertainty and Granularity*, Maine (Etats-Unis), 2001

[Heuvelink, 1998] Heuvelink, G.B.M. *Error propagation in Environmental Modeling with GIS*. Londres : Taylor et Francis, 1998

[Hirst et St Onge, 1998] Hirst, G. et St Onge, D. *Lexical chains as representations of context for the detection and correction of malapropisms*. In : Fellbaum, C. (éd.), *WordNet: An electronic lexical database*. Cambridge (Etats-Unis) : The MIT Press, 1998

[Hubert et Ruas, 2003] Hubert, F. et Ruas, A. *A method based on samples to capture user needs for generalisation*. In : *Proceedings of the 5th Workshop on Progress in automated map Generalisation*, 2003

[Hunter, 1998] Hunter, G.J. *Managing Uncertainty in GIS*. *NCGIA Core Curriculum in GIScience*, février 1998

[Hunter et Goodchild, 1997] Hunter, G.J. et Goodchild, M.F. *Modelling uncertainty of slope and aspect estimates derived from spatial databases*. *Geographical Analysis*, 1997, vol. 29, p. 529-537

[Kieler, 2007] Kieler, B. *A geometry-driven approach for the semantic integration of geodata sets*. In : *Proceedings of the XXIII International Cartographic Conference*, Moscou, 4-10 août 2007

[Kilpelaïnen, 2000] Kilpelaïnen, T. *Maintenance of Multiple Representation Databases of Topographic Data*. *The Cartographic Journal*, 2007, vol. 37, n°2, p. 101-107

[Lauriette-Rougegrez, 2006] Lauriette-Rougegrez, S. *Théorie de l'évidence et logique floue pour la détermination du stade de réalisation d'un scenario*. In : *Actes des 12èmes rencontres francophones sur la Logique Floue et ses Applications*, Toulouse, 19-20 octobre 2006, p. 405-412

[Lefevre *et al.*, 2000a] Lefevre, E., Colot, O., Vannorenberghe, P. et De Brucq., D. *Contribution des mesures d'information à la modélisation crédibiliste de connaissances. Traitement de signal,* 2000, vol. 17, n°2, p. 87-97

[Lefevre *et al.*, 2000b] Lefevre, E., Colot, O., Vannoorenberghe, P. et De Brucqv D. A generic framework for resolving the conflict in the combination of belief structures. In : *Proceeding of 3rd International Conference on Information Fusion*, Paris, 2000

[Lefevre *et al.*, 2002] Lefevre, E., Colot, O., et Vannorenberghe, P. *Belief function combination and conflict management*. *Information Fusion*, 2002, vol. 3, n°2, p. 149-162

[Le Hégarat-Mascle *et al.*, 2003] Le Hégarat-Mascle, S., Richard, D. et Ottlé, C. *Multi-scale data fusion using Dempster-Shafer evidence theory*. *Integrated Computer-Aided Engineering*, 2003, vol. 10, n°1, p. 9-22

[Lemarié, 1997] Lemarié, C. *Etude de l'appariement géométrique Cadastre/BDTOPO – Version 1.1* –IGN, COGIT, Rapport technique DT/97-0019, 1997

[Lemarié et Bucaille, 1998] Lemarié, C., et Bucaille, O. Spécifications d'un module générique d'appariement de données géographiques, In : *Proceedings of the Reconnaissance des Formes et Intelligence Artificielle,* Clermont-Ferrant, janvier 1998, p 397-406

[Lemeret *et al.*, 2004] Lemeret Y., Lefevre, E. et Jolly, D. *Simulator of obstacle detection and tracking*. In : *Proceedings of the 5th EUROSIM Congress on Modelling and Simulation*, 2004

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

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

[Mackaness et Ruas, 2007] Mackaness, W. et Ruas, A. Evaluation in the Map Generalisation Process. In : Mackaness, W., Ruas, A. et Sarjakoski, T. (éd), *The Generalisation of Geographic Information: Models and Applications*. Elsevier, 2007, p. 89-111

[Madhavan *et al.*, 2001] Madhavan, J., Bernstein, P.A. et Rahm, E. *Generic schema matching with cupid*. In : *Proceedings of the 27th International Conference on Very Large Databases*, Rome, 11-14 septembre 2001, p. 49-58

[Martin, 2005] Martin, A. *Fusion de classifieurs pour la classification d'images sonar*. In : *RNTI Extraction des connaissances : Etat et perspectives*, novembre 2005, p. 259-268

[Mascret et Devogele, 2006] Mascret, A. et Devogele, T. *Intégration de données Terre/Mer basée sur une approche paysagère*. In : *Actes du Colloque International de Géomatique et d'Analyse Spatiale (SAGEO)*, *Strasbourg*, *septembre 2006*

[Masson, 2005] Masson, M.-H. *Apports de la théorie des possibilités et des fonctions de croyance à l'analyse de données imprécises*, Mémoire de HDR, Université de Compiègne, 2005

[Mathevet *et al.*, 1999] Mathevet, S., Trassoudaine, L., Checchin, P. et Auzon, J. *Combinaison de segmentations en regions. Traitement du signal*, 1999, vol.16, n°2, p. 93-104

[Mauclair et Pinquier, 2004] Mauclair J. et Pinquier J. *Fusion de paramètres en classification Parole/Musique*. In : *Actes des XXVèmes Journées d'Etude sur la Parole*, Fès (Maroc), 19-21 avril 2004, p. 353-356

[Maussang *et al.*, 2008] Maussang, F., Rombaut, M., Chanussot, J., Hétet, A. et Amate, M. *Fusion of Local Statistical Parameters for Buried Underwater Mine Detection in Sonar Imaging. EURASIP Journal on Advances in Signal Processing*, 2008

[McMaster, 1983] McMaster, R.B. *Mathematical measures for the evaluation of simplified lines on maps*. Thèse de Doctorat, Université du Kansas, Etats-Unis, 1983

[McMaster et Shea, 1992] McMaster, R.B. et Shea, E. *Generalisation in digital Cartography.* Washington : Association of American Geographers, 1992

[McMaster, 1986] McMaster, R. *A statistical Analysis of Mathematical Measures for Linear Simplification, The American Cartographer*, 1986, vol. 23

[Mellouli *et al.*, 1987] Mellouli, K., Shafer, G., et Shenoy, P. Qualitative Markov networks. In: *Uncertainty in Knowledge-Based Systems*, Springer-Verlag, 1987, p. 69-74

[Mercier, 2006] Mercier, D. *Fusion d'informations pour la reconnaissance automatique d'adresses postales dans le cadre de la théorie des fonctions de croyance*. Thèse de Doctorat, Université de Technologie de Compiègne, 2006

[Mikhail et Ackerman, 1976] Mikhail, E. M. et Ackerman, F. *Observations and Least Squares*. *IEP*-A Dun-Donnelley Publisher, New York, 1976

[Minami, 2000] Minami, M. *Using ArcMap*. ESRI press, 2000, p. 544

[Min *et al.*, 2007] Min, D., Zhilin, L. et Xiaoyong, C. *Extended Hausdorff distance for spatial objects in GIS*. *International Journal of Geographical Information*, 2007, vol. 21, n°4, p. 459-475

[Mitropoulos *et al.*, 2005] Mitropoulos, V., Xydia, A., Nakos, B. et Vescoukis, V. *The use of epsilon convex area for attributing bends along a cartographic line*. In : *Proceedings of the International Cartographic Conference*, la Corogne, 2005

[Mostafavi, 2006] Mostafavi, M.A. *Semantic similarity assessment in support of spatial data integration*. In : *Proceedings of the 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, Lisbonne, 7-9 juillet 2006, p. 685-693

[MultiNet, 2004] Spécifications de contenu de la base MultiNet, version 1.0, TeleAtlas, 2004

[Mustière, 2001] Mustière, S. *Apprentissage supervisé pour la généralisation cartographique*. Thèse de Doctorat, Université Paris 6, 2001

[Mustière, 2006] Mustière, S. *Results of Experiments on Automated Matching of Networks at Different Scale*, In : *Proceedings of the ISPRS Workshop, Multiple representation and interoperability of spatial data*, Hanovre (Allemagne), 22-24 février 2006, p. 92-100

[Mustière *et al.*, 2007] Mustière, S., Abadie, N. et Laurens, F. *Appariement de schémas de BD géographiques à l'aide d'ontologies déduites des spécifications*. In : *Actes de la Conférence EGC*, Namur, janvier 2007, p. 22-27

[Mustière et Devogele, 2008] Mustière, S. et Devogele, T. *Matching networks with different levels of detail*. *GeoInformatica*, à paraître en 2008

[Mustière et van Smaalen, 2007] Mustière, S. et van Smaalen, J. Database Requirements for Generalisation and Multiple Representations. In : Mackaness, W., Ruas, A. et Sarjakoski, T. (éd), *The Generalisation of Geographic Information: Models and Applications*. Elsevier, 2007, p. 113-136

[Niskanen, 1989] Niskanen, V.A. *Introduction to imprecise reasoning, uncertainty, decision making and knowledge engineering*. *Publications of the Society for Artificial Intelligence in Finland*, 1989, vol. 1

[Olteanu, 2007] Olteanu, A.M. *Matching geographical data using the Theory of Evidence*. In : *Proceedings of the 20th International Cartographical Conference*, Moscou, 5-9 août 2007

[Olteanu-Raimond et Mustière, 2008] Olteanu-Raimond, A.M. et Mustière, S. *Data matching - a matter of belief*. In : *Proceedings of Spatial Data Handling*, Montpellier, 23-25 juillet 2008

[Omrani *et al.*, 2007] Omrani, H., Ion-Boussier, L. et Trigano, P. *A New Approach for Impacts Assessment of Urban Mobility*. *WSEAS Transactions on Information Science and Applications*, 2007, vol. 4, n°3, p. 439-444

[Parent et Spaccapietra, 1996] Parent, C. et Spaccapietra, S. *Intégration de bases de données : panorama des problèmes et des approches. Ingénierie des systèmes d'information*, 1996, vol. 4, n° 3, p. 333-358

[Patwardham, 2003] Patwardham, S. *Incorporating Dictionary and Corpus Information in a Measure of Semantic Relatedness*. Thèse de Doctorat, Université du Minnesota, Etats-Unis, 2003

[Pawlak, 1982] Pawlak, Z. *Rough sets. International Journal of Computer and Information Sciences,* 1982, vol, 11, p. 341-356

[Pawlak, 1991] Pawlak, Z. *Rough Sets: Theoretical Aspects of Reasoning about Data*. Kluwer, 1991, p. 1-252

[Pham, 2005] Pham, T.T. *Fusion de l'information géographique hiérarchisée*. Thèse de Doctorat, Université de Provence, 2005

[Plazanet, 1996] Plazanet, C. *Enrichissement des bases de données géographiques : analyse de la géométrie des objets linéaires pour la généralisation cartographique (application aux routes)*. Thèse de Doctorat, Université de Marne-la-Vallée, 1996

[Plazanet *et al.*, 1998] Plazanet, C., Martini Bigolin, N. et Ruas, A. *Experiments with Learning Techniques for Spatial Model Enrichment and Line Generalization*. *GeoInformatica,* 2003, vol. 2, n° 4, p. 315-333

[Pontius et Cheuk, 2006] Pontius, R.G. et Cheuk, M.L. *A generalized cross tabulation matrice to compare soft-classified maps at multiple resolution*. *IJGIS*, janvier 2006, vol. 20, n°1, p. 1-30

[Poroseva *et al.*, 2006] Poroseva, S.V., Letschert, J. et Yousuff Hussaini, M. *Application of evidence theory to quantify uncertainty in hurricane path*. In : *Proceedings of the American Meteorological Society 86th Annual Meeting*, Atlanta, 29 janvier-2 février 2006

[Rahm et Bernstein, 2001] Rahm, E. et Bernstein, P.A. *On Matching Schemas Automatically*. Microsoft Research Technical Report MSR-TR-2001-17, 2001

[Ramasso *et al.*, 2007] Ramasso, E., Rombaut, M. et Pellerin D. *L'ordonnanceur crédibiliste pour la reconnaissance de séquences d'états dans le cadre du Modèle des Croyances Transférables* In : *Actes des Rencontres francophones sur la Logique Floue et ses Applications*, 22-23 novembre 2007, Nîmes, p. 141-148

[Ramsey, 2004] Ramsey, P. *The jump project and direction*, 2004 [référence du 15 avril 2008]. http://www.jump-project.org/assets/JUMP_Project_and_Direction.pdf

[Resnik, 1995] Resnik, P. *Using information content to evaluate semantic similarity in a taxonomy*. In : *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montréal, 1995

[Rodriguez et Egenhofer, 2004] Rodriguez, A. et Egenhofer, M.J. *Comparing geospatial entity classes: an asymmetric and context similarity measure*. *IJGIS*, 2004, vol. 18, n°3, p. 229-256

[Rohmer, 2007] Rohmer, J. *Application de la théorie des fonctions de croyance pour la synthèse des courbes sismiques probabilistes paramétriques imprécises*. In : *Actes des Rencontres francophones sur la Logique Floue et ses Applications*, Nîmes, 22-23 novembre 2007, p. 99-106

[Rombaut, 1998] Rombaut, M. *Decision in multi-obstacle matching process using Dempster-Shafer's theory*. In : *Proceedings of Advances in Vehicle Control and Safety*, Amiens, 1-3 juillet 1998

[Rombaut et Zhu, 2002] Rombaut, M. et Zhu, Y.M. *Study of Dempster-Shafer theory for image segmentation applications*. *Image and vision computing*, 2002, vol. 20, n°1, p.15-23

[Rottensteiner *et al.*, 2005] Rottensteiner, F., Trinder, J., Clode, S. et Kubik, K. *Using the Dempster-Shafer method for the fusion of LIDAR data and multi-spectral images for building detection. Information Fusion,* 2005, vol. 6, n°4, p. 283-300

[Royère, 2002] Royère, C. *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 de Doctorat, Université de Compiègne, 2002

[Ruas, 2000] Ruas, A. The role of meso objects for generalisation, In: *Proceedings of 9th* International Symposium on Spatial Data Handling, Pékin, 2000

[Ruas, 2001] Ruas A. *Automatic generalization project: Learning process from interactive generalization*. *OEEPE Official Publication*, 2001, n°39, p. 98

[Ruas, 2002] Ruas, A. *Pourquoi associer les représentations des données géographiques ?* In : Ruas, A (éd.), *Généralisation et représentation multiple*. Paris : Lavoisier, 2002, p. 390

[Ruas, 1999] Ruas, A. *Modèle de généralisation de données géographiques à base de contraintes et d'autonomie*. Thèse de Doctorat, Université Marne-la-vallée, 1999

[Saalfeld, 1988] Saalfeld, A. *Conflation: Automated map compilation*. *International Journal of Geographic Information systems*, 1988, vol. 2, n°3, p. 217-228

[Saber Naceur *et al.*, 2000] Saber Naceur, M., Belhadj, Z. et Boussema, M.R. *Fusion de données satellitales basée sur la théorie de Dempster-Shafer pour la cartographie et l'occupation du sol en milieu semi-aride*. *Bulletin de la Société française de photogrammétrie et de télédétection*, 2000, n°158, p. 3-11

[Safra *et al.*, 2006] Safra, E., Kanza, Y., Sagiv, Y. et Doytsher, Y. *Efficient Integration of Road Maps*. In : *Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems*, Arlington (Etats-Unis), 10-11 novembre 2006. ACM Press, p. 59-66

[Samal *et al.*, 2004] Samal, A., Seth, S. et Cueto, K. *A feature-based approach to conflation of geospatial sources*. *IJGIS*, juillet-août 2004, vol. 18, n°5, p. 459-489

[Schneider, 1999] Schneider, M. *Uncertainty Management for Spatial Data in Databases: Fuzzy Spatial Data Types*. In : *Proceedings of the 6th International Symposium On Advances in Spatial Databases*, Londres, 20-23 juillet 1999. Springer-Verlag, p. 330-351

[Schneider, 2001] Schneider, M. *A Design of Topological Predicates for Complex Crisp and Fuzzy Regions*. *Lecture Notes in Computer Science*, 2001, vol. 2224, p. 149-162

[Sester, 1998] Sester, M. *Interpretation of Spatial Data Bases using Machine Learning Techniques*. In : *Proceedings of the 8th International Symposium on Spatial Data Handling*, 2003

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

[Shafer, 1982] Shafer, G. *Belief Functions and Parametric Models. Journal of the Royal Statistical Society*, 1982, p. 322-352

[Shafer, 1987] Shafer, G. Belief functions and possibility measures. In : *The Analysis of Fuzzy Information*, *Vol. 1: Mathematics and Logic*, 1987, p. 51-84

[Shannon, 1948] Shannon, C.E. *A Mathematical Theory of Communication*. Bell Syst. Techn. Vol.

[Sheeren *et al.*, 2004] Shereen, D., Mustiere, S. et Zucker, J.-D. *How to Integrate Heterogeneous Spatial Databases in a Consistent Way?* In : *Proceedings of the Conference on Advanced Databases and Information Systems (ADBIS),* Budapest, septembre 2004, p. 364-378

[Sheeren, 2005] Sheeren, D. *Méthodologie d'évaluation de la cohérence inter-représentations pour l'intégration de bases de données spatiales*. Thèse de Doctorat, Université Paris 6, 2005

[Sheeren *et al.*, 2008] Sheeren, D., Mustiere, S. et Zucker, J.-D. *A data mining approach for assessing consistency between multiple representations in spatial databases. IFGIS,* à paraître en 2008

[Shi *et al.*, 2002] Shi, W. et Guo, W. Topological relationships between spatial objects with uncertainty. In : Shi, W., Fisher, P.F. et Goodchild, M. (éd.), *Spatial Data Quality*. New-York : Taylor & Francis, 2002, p. 50-61

[Smets, 1998] Smets, P. The Transferable Belief Model for Quantified Belief Representation.Handbook of Defeasible Reasoning and Uncertainty Management Systems. In : Gabbay, D. et Smets, P. (éd.), Vol. 1 : *Quantified Representation of Uncertainty & Imprecision*, Kluwer Academic Publishers, Doordrecht, 1998, p. 267-301

[Smets, 1988] Smets, P. Belief Functions. In : Smets, P., Mamdani, A., Dubois, D. et Prade, H. (éd.), *Non Standard Logics for Automated Reasoning*. Londres : Academic Press, 1988, p. 253-286

[Smets, 1990] Smets, P. *The Combination of Evidence in the Transferable Belief Model*. *IEEE Trans.*, 1990, p. 447-458

[Smets, 1992] Smets, P. The nature of the unnormalized beliefs encountered in the transferable belief model. In : *Proceedings of the 8th Conference on Uncertainty in Artificial Intelligence*, San Mateo (Etats-Unis), 1992, p. 292-297

[Smets, 1993] Smets, P. *Belief functions: The disjonctive rule of combination and the generalized bayesian theorem*. *International Journal of Approximate Reasoning*, 1993, vol. 9, p. 1-35

[Smets et Kennes, 1994] Smets, P. et Kennes, R. *The Transferable Belief Model. Artificial Intelligence,* 1994, vol. 66, n°2, p. 191-234

[Smets, 1997] Smets, P. Imperfect information: Imprecision - Uncertainty. In : Motro, A. et Smets, P. (éd), *Uncertainty Management in Information Systems: From Needs to Solutions*, Kluwer Academic Publishers, 1997, p. 225-254

[Smets et Kruse, 1997] Smets, P. et Kruse, R. The Transferable Belief Model for Belief Representation. In : Motro, A. et Smets, P. (éd.), *Uncertainty Management in information systems: from needs to solutions*. Kluwer Academic Publishers, Boston, (1997), p. 343-368

[Smith et Varzi, 1997] Smith, B. et Varzi, A.C. *Fiat and bona fide boundaries.* In : *Proceedings of COSIT-97*, Springer-Verlag, 1997, p. 103-119

[Stigmar, 2005] Stigmar, H. Matching Route Data and Topographic Data in a Real-time Environment. In : *Proceedings of 10 th ScanGIS Congress*, Stockholm, 2005

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

[Tøssebro, 2002] Tøssebro, E. *Representing uncertainty in spatial and spatiotemporal databases*. Norwegian University of Science and Technology, Department of Computer and Information Science, 2002

[Taillandier, 2007] Taillandier, P. *Automatic Knowledge Revision of a Generalisation System*. In : *Proceedings of the Workshop on Generalisation and Multiple Representation*, Moscou, août 2000

[Theng, 2006] Theng, J. *Conception d'outils ergonomiques pour la navigation au sein de données géographiques à représentation multiple*, Rapport de stage, Université Marne-La-Vallée, 2006

[Uitermark *et al.*, 1998] Uitermark, H.T., Oosterom, P.J.M., van Mars, N.J.I. et Molenaar, M. *Propagating updates: finding corresponding objects in a multi-source environment*. In : *Proceedings of the 8th International Symposium on Spatial data Handling (SDH'98)*, Vancouver, 11-15 juillet 1998, p. 580-591

[Uitermark, 2001] Uitermark, H. *Ontology-Based Geographic Data Set Integration*, Thèse de Doctorat, Université Twente, Pays-Bas, 2001

[Vannoorenberghe et Denœux, 2002] Vannoorenberghe, P. et Denoeux, T. Handling uncertain labels in multiclass problems using belief decision trees. In : Proceedings of IPMU, Annecy, 2002, p. 1919-1926

[Vannoorenberghe *et al.*, 2000] Vannoorenberghe, P., Lefevre, E., Colot, O. *Application de la théorie des fonctions de croyance à la surveillance de l'environnement*. In : *Actes des Rencontres francophones sur la Logique Floue et ses Applications (LFA)*, 2000, p. 229-236

[Vannoorenberghe *et al.*, 2003] Vannoorenberghe, P., Lefevre, E. et Colot, O. *Traitement d'image et théorie des fonctions de croyance*. In : *Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA)*, 2003, p. 287-294

[Vannoorenberghe et Flouzat, 2006] Vannoorenberghe, P. et Flouzat, G. *Segmentation d'image et extraction d'information : une approche basée sur les fonctions de croyance*. In : *Actes des Rencontres francophones sur la Logique Floue et ses Applications (LFA)*, Toulouse, 19-20 octobre 2006, p. 355-362

[Vasco et Caetano, 2006] Vasco, B. et Caetano, M. *Mapping Uncertainty in land cover characterization by comparison of land cover cartographies*. In : *Proceedings from the 7th International Symposium on Spatial Accuracy Assessement in Natural Resources and Environmental Sciences*, Lisbonne, 7-9 juillet 2006, p. 705-715

[Vauglin, 1997] Vauglin, F. *Modèles statistiques des imprécisions géométriques des objets géographiques linéaires*. Thèse de Doctorat, Université Marne-la-Vallée, 1997

[Virrantaus, 2003] Virrantaus, K. *Analysis of the uncertainty and imprecision of the source data sets for a military terrain analysis application*. In : *Proceedings of ISSDQ*, Hong-Kong, 2003

[Voltz, 2005] Voltz, S. *Data-Driven Matching of Geospatial Schemas*. In : *Proceedings of the International Conference on Spatial Information Theory (COSIT)*, Ellicottville (Etats-Unis), september 2005, p. 115-132

[Voltz, 2006] Voltz, S. *An Iterative Approach for Matching Multiple Representations of Street Data*. In : *Proceedings of ISPRS Workshop, Multiple representation and interoperability of spatial data*, Hanovre (Allemagne), 22-24 février 2006, p. 101-110

[Wallerman *et al.,* 2006] Wallerman, J., Coomaren, P., Vencatasawmy et Bondesson, L. *Spatial simulation of forest using Bayesian state-space models and remotely sensed data.* In : *Proceedings of the 7th International Symposium on Spatial Accuracy Assessement in Natural Resources and Environmental Sciences*, Lisbonne, 7-9 juillet 2006, p. 520-530

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

[Wang *et al.*, 2005] Wang, S., Shi, W., Yuan, H. et Chen G. *Attribute Uncertainty in GIS Data*. *FSKD*, 2005, n°2, p. 614-623

[Wang *et al.*, 2002] Wang, S., Li, D., Shi, W., Wang, X. et Li, D. *Rough spatial description*. In *Proceedings of ISPRS Commission II Symposium on Integrated System for Spatial Data Production, Custodian and Decision Support*, Xi´an, 20-23 août 2002, p. 503-509

[Wang et Bell, 2004] Wang, H. et Bell, D. *Extended K-Nearest Neighbours Based on Evidence Theory*. *The Computer Journal*, 2004, vol. 47, n°6, p. 662-672

[Woodcock et Gopal, 2000] Woodcock, C. et Gopal, S. *Fuzzy set - Theory and thematic map*, 2000

[Worboys, 1998] Worboys, M.F. *Computation with Imprecise Geospatial Data*. *Computers, Environment and Urban Systems*, 1998, vol. 22, n°2, p. 85-106

[Worboys et Duckham, 2004] Worboys, M. et Duckham, M. *GIS – computing Perspective*, Second Edition, CRC Press LLC, 2004

[Wright, 2000] Wright, E.J., *Extending Techniques From 'Standard' Error Theory To Categorical Data And Local GIS Operations*, In : *Proceedings of the 2000 American Society of Photogrammetry and Remote Sensing (ASPRS) Conference.* Washington : Spring 2000, May 2000

[Wright, 1998] Wright, E. *Application of Bayesian Networks for Representing Uncertainty in Geospatial Data. ASPRS convention,* 1998

[Wu et Palmer, 1994] Wu, Z. et Palmer, M. *Verb Semantics and Lexical Selection*, In : *Proceedings of the 32nd Annual Meetings of the Associations for Computational Linguistics*, 1994, p. 133-138

[Yager, 1987] Yager, R. *On the Dempster-Shafer framework and new combination rules*. *Informations Sciences*, 1987, n°41, p. 93-138

[Yuan et Tao, 1999] Yuan, S. et Tao, C., Development of Conflation Components. In : Li, B., et al. (éd.), *Proceedings of Geoinformatics Conference*, 1999, p. 1-13

[Zadeh, 1965] Zadeh, L.A. *Fuzzy sets*. *Information and Control*, 1965, vol. 8, p. 338-353

[Zadeh, 1979] Zadeh, L.A. *On the validity of Dempster's Rule of Combination of Evidence*, University of California, Berkeley, 1979, ERL Memo M79/24

[Zadeh, 1978] Zadeh, L.A. *Fuzzy sets as a basis for a theory of possibility*. *Fuzzy sets and systems*, 1978, vol. 1, p. 3-28

[Zadeh, 1979] Zadeh, L.A. *On the validity of Dempster's rule of combination of evidence*. Technical Report UCB/ERL M79/24, 1979, University of California, Berkeley

[Zhan, 1998] Zhan, F.B. Approximate analysis of binary topological relations between geographic regions with indeterminate boundaries. In : *Soft Computing,* Springer-Verlag, 1998, vol. 2, p. 28-34

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

[Zhang et Couloigner, 2004] Zhang, Q. et Couloigner I. *A framework for road change detection and map updating.* In : *Proceedings of the 20th ISPRS Congress*, Istanbul, 12-23 juillet 2004

[Zhang et Goodchild, 2002] Zhang, J. et Goodchild, M. *Uncertainty in geographical Information*. Taylor & Francis, 2002, p. 266

[Zobel et Dart, 1995] Justin Zobel et Philip Dart. Finding Approximate Matches in Large Lexicons. In : *Software-Practice and Experience*, 1995, vol. 25, n°3, p. 331–345

[Zouhal et Denoeux, 1998] Zouhal L. M. et Denoeux, T. *An evidence theoretic k-NN rule with parameter optimization*. IEEE Transactions on Systems, Man and Cybernetics, 1998, vol. 28, n° 2, p. 263-271

Annexes

Annexe 1. Aperçu des théories mathématiques pour modéliser les imperfections

Dans cette annexe nous présentons un aperçu des théories mathématiques qui permettent de modéliser les imperfections.

La théorie des probabilités

La théorie des probabilités est une théorie mathématique qui étudie des phénomènes (« expériences aléatoires ») pouvant avoir plusieurs issues possibles (« résultats »). Elle associe à chaque résultat un réel compris entre 0 et 1 qui correspond à la probabilité que le résultat se produira lorsque l'expérience aléatoire est réalisée.

Soit (Ω,P(Ω)) un espace probabiliste associé à une expérience aléatoire, où Ω représente l'ensemble de référence contenant tous les résultats possibles et P(Ω) représente l'ensemble de toutes les parties de Ω. On appelle probabilité sur Ω toute fonction P de P(Ω) dans l'intervalle [0, 1], vérifiant les critères suivants :

$$\begin{cases} \Pr(\Omega) = 1 \\ \forall A_1 \in \mathrm{P}(\Omega), A_2 \in \mathrm{P}(\Omega), \ \textit{tels que}\ A_1 \cap A_2 = \phi, \ \Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2) \end{cases} \quad (46)$$

Le deuxième critère est le critère d'additivité, défini pour toute collection de résultats $A_1 \ldots A_n$ disjoints.

La fusion dans le cadre de la théorie des probabilités peut se faire en utilisant les probabilités conditionnelles, et elle est basée sur des règles strictes de combinaison des probabilités afin de calculer ou réviser la probabilité d'une hypothèse.

La décision consiste à retenir l'événement associé à la plus grande probabilité a posteriori.

Dans le cadre du modèle probabiliste, en absence de toute information, on associe aux résultats des probabilités égales. En conséquence, l'ignorance totale est mal gérée par le modèle probabiliste. Ce dernier s'appuie sur des probabilités, c'est-à-dire des nombres fournis par l'expert pour décrire les événements. Ces probabilités doivent être précises et non approximatives. En effet, la théorie des probabilités ne gère pas non plus l'imprécision.

Dans le domaine de l'information géographique, parmi les travaux qui utilisent la théorie des probabilités, nous pouvons citer ceux qui utilisent des modèles tels que le modèle bayésien (à chaque probabilité antérieure on attribue une nouvelle information afin d'avoir la probabilité a posteriori) [Wright, 1998 ; Zhang et Goodchild, 2002] ou les champs de Markov [Mellouli *et al.*, 87 ; Zhang et Goodchild, 2002 ; Wallerman *et al.*, 2006].

En conclusion, la théorie des probabilités est adaptée surtout à la modélisation des connaissances statistiques que nous avons sur un problème donné. Elle est limitée, parce qu'elle ne permet pas une modélisation souple de l'incomplétude et de l'imprécision. De plus elle nécessite des connaissances a priori. Enfin, elle peut être vue comme un cas particulier de la théorie de l'évidence, lorsque l'imprécision est nulle et lorsque les éléments focaux sont des singletons [Bouchon-Meunier, 1995].

La théorie des ensembles approximatifs

La théorie des ensembles approximatifs, définie par [Pawlak, 1982] au début des années 1980, s'appuie sur la notion d'ensemble approximatif.

Un ensemble approximatif f est défini par une fonction d'appartenance qui associe à chaque élément $x \in X$, une valeur dans l'ensemble E de la manière suivante :

$$f : X \rightarrow E \text{, avec } E = \{T, M, F\} \quad (47)$$

où T, M et F signifient respectivement « peut être », « appartient » et « n'appartient pas ».

Les concepts sont représentés par leurs approximations inférieure et supérieure. Ainsi, l'approximation supérieure d'une classe X est l'ensemble des objets n'appartenant probablement pas à X. L'approximation inférieure d'une classe X est l'ensemble des objets appartenant sûrement à X.

La théorie des ensembles approximatifs est un outil pour manipuler l'imprécision et l'incertitude inhérentes aux situations de décision. A la différence de l'approche des probabilités conditionnelles, elle n'impose pas la connaissance a priori des données.

Dans le domaine de l'information géographique, les objets imprécis sont définis en utilisant une région positive, négative et une région de frontière, la région de frontière contenant les objets qui ne peuvent être classés avec certitude ni à l'intérieur de *X*, ni à l'extérieur de *X*. La théorie des ensembles approximatifs est utilisée surtout lorsque nous devons manipuler des phénomènes indiscernables.

La théorie des sous-ensembles flous

La théorie des sous-ensembles flous est une théorie non probabiliste introduite par [Zadeh, 1965]. Ce dernier a introduit la notion de sous-ensemble flou afin de résoudre le problème posé par les connaissances exprimées symboliquement, présentant des imprécisions ou ayant un caractère vague. Cette théorie repose sur l'idée d'appartenance partielle à une classe, de gradualité dans le passage d'une limite à une autre, ainsi que sur le fait qu'une classe donnée n'a pas de bornes strictes. Elle est considérée comme une interface entre les connaissances décrites symboliquement telles que grand, moyen ou petit, et les connaissances décrites numériquement [Bouchon-Meunier, 1995 ; Gascône, 1997].

Définition : un sous-ensemble classique A de Ω est défini par une fonction caractéristique, $\chi_A(x)$, prenant la valeur 0 ou 1, de la façon suivante :

$$\chi_A(x) : \Omega \rightarrow \{0,1\} \quad (48)$$

Définition : un sous-ensemble flou A de Ω est défini par une fonction d'appartenance qui associe à chaque élément $x \in \Omega$, le degré $f_A(x)$ compris entre 0 et 1 avec lequel x appartient à A.

$$f_A(x) : \Omega \rightarrow [0,1] \quad (49)$$

Lorsque $f_A(x)$ ne prend que des valeurs égales à 0 ou 1, le sous-ensemble flou devient un sous-ensemble classique. Un sous-ensemble flou n'impose pas la normalisation par rapport à la théorie des probabilités et la théorie des possibilités.

Le noyau de A, noté Noy(A), est défini comme l'ensemble des éléments $x \in \Omega$, appartenant totalement à A, c'est-à-dire :

$$\text{Noy}(A) = \{x \in \Omega, f_A(x) = 1\} \quad (50)$$

Le support de A, noté supp(A), est l'ensemble ayant un degré d'appartenance non nul, c'est-à-dire :

$$\text{Supp}(A) = \{x \in \Omega, f_A(x) \neq 0\} \quad (51)$$

Cette théorie des sous-ensembles flous présente l'avantage de pouvoir modéliser les opérateurs du type ET, OU et également de les combiner.

La théorie des sous-ensembles flous ne permet pas de modéliser l'incertitude. Or, comme d'ailleurs Bouchon-Meunier l'affirme [Bouchon-Meunier, 1995], l'imprécision et l'incertitude sont fortement liées. Par conséquent, [Zadeh 1978] a introduit la théorie des possibilités.

La modélisation des entités géographiques basée sur la théorie des sous-ensembles flous a été utilisée par de nombreux auteurs [Hansen, 2003 ; Fonte et Lodwick, 2004]. L'étude de l'occupation du sol est l'exemple typique d'implémentation, parce que les objets géographiques sont vagues par définition et qu'ils n'ont pas de contours bien définis [Comber *et al.*, 2005a ; Woodcock et Gopal, 2000 ; Akyürek et Kıvanç, 2006].

La théorie des possibilités

La théorie des possibilités dérive de celle des ensembles flous [Dubois et Prade, 1988 ; Benferhat *et al.*, 2003]. Elle permet de modéliser les incertitudes sur des événements, sans avoir besoin de connaissances a priori. Elle permet de dire dans quelle mesure un événement est possible et dans quelle mesure on en est certain. Ces deux affirmations sont formalisées grâce à deux mesures : la mesure de possibilité et la mesure de nécessité. Ces deux mesures sont des cas particuliers de la plausibilité et de la crédibilité.

La mesure de possibilité

Etant donné un ensemble de référence Ω, contenant un certain nombre d'événements, on associe à chaque événement, c'est-à-dire à tout élément de l'ensemble P(Ω) des sous-ensembles ordinaires de Ω, une mesure évaluant à quel point cet événement est possible au moyen d'un coefficient compris entre 0 et 1.

Définition : une mesure de possibilité [Bouchon-Meunier, 1995], notée Π, est une fonction définie sur l'ensemble P(Ω), prenant des valeurs dans l'intervalle [0,1], telle que :

$$\begin{cases} a)\ \Pi(\phi) = 0,\ \Pi(\Omega) = 1 \\ b)\ \forall A_i \in P(\Omega),\ i = 1,2\ldots\ \Pi(\cup_{i=1,2,\ldots} A_i) = \sup_{i=1,2,\ldots} \Pi(A_i) \end{cases} \quad (52)$$

où sup représente la plus grande valeur dans le cas fini.

Etant donnés seulement deux sous-ensembles de Ω, la propriété b) de l'équation précédente s'écrit de la manière suivante :

$$\forall A_1 \in P(\Omega),\ A_2 \in P(\Omega),\ \Pi(A_1 \cup A_2) = \max(\Pi(A_1),\ \Pi(A_2)) \quad (53)$$

La mesure de nécessité

La mesure de nécessité indique le degré avec lequel la réalisation d'un événement est certaine, attribuant un coefficient compris entre 0 et 1 à toute partie de Ω. Les propriétés de la mesure de nécessité, différentes de celles de la mesure de possibilité, sont les suivantes :

Définition : une mesure de nécessité [Bouchon-Meunier, 1995], notée *N*, est une fonction définie sur l'ensemble P(Ω) des parties de Ω, prenant des valeurs dans l'intervalle [0,1], telle que :

$$\begin{cases} a)\ N(\phi) = 0,\ N(\Omega) = 1 \\ b)\ \forall A_1 \in P(\Omega),\ i = 1,2,...,\ N(\cap_{i=1,2,...} A_i) = \inf_{i=1,2,...} N(A_i) \end{cases} \quad (54)$$

où inf représente la plus petite valeur parmi toutes les valeurs, dans le cas fini.

Etant donnés seulement deux sous-ensembles de Ω, la propriété b) de l'équation (32) s'écrit de la manière suivante :

$$\forall A_1 \in P(\Omega),\ A_2 \in P(\Omega),...N(A_1 \cap A_2) = \min(N(A_1),\ N(A_2)) \quad (55)$$

Ainsi, une mesure de nécessité associe à l'intersection de deux événements, la plus petite valeur des coefficients attribués à chacun des événements.

La mesure de nécessité est liée à la mesure de possibilité par la formule :

$$\Pi(A) = 1 - N(\overline{A}) \quad (56)$$

Etant donné un événement E, dans la théorie des probabilités, on évalue les chances que celui-ci se réalise en calculant sa probabilité P(E), alors que dans la théorie des possibilités, la connaissance sur l'occurrence de l'événement E est évaluée par la mesure de possibilité Π(E). Afin d'indiquer dans quelle mesure l'événement E se réalise, lorsque nous ne disposons pas de sa probabilité de réalisation P(E), l'intervalle [Π(E), N(E)] est utilisé.

La théorie des possibilités propose un grand nombre d'opérateurs de fusion [Dubois et Prade, 1994]. Ainsi, on peut citer l'opérateur minimum, qui effectue une combinaison conjonctive lorsque les sources d'information sont concordantes, l'opérateur maximum, qui réalise une combinaison disjonctive lorsque les sources sont discordantes, et l'opérateur adaptif qui réalise une combinaison adaptive en fonction du degré de compatibilité des deux sources d'information.

Au niveau de la décision, nous pouvons citer le maximum de possibilité et le maximum de nécessité.

On peut remarquer la différence suivante entre la théorie des possibilités et la théorie des probabilités : les produits de la théorie des probabilités deviennent des recherches de minimum dans la théorie des possibilités et les sommes deviennent des recherches de maximum.

Une différence majeure entre la théorie des probabilités et la théorie des possibilités concerne un événement et son complément. La probabilité d'un événement détermine complètement son complémentaire, alors que la possibilité et la nécessité d'un événement et son complémentaire sont faiblement liées.

La théorie des possibilités modélise la préférence que l'on a pour une hypothèse tandis que la théorie des fonctions de croyance formalise plutôt un degré de vérité plus ou moins certain.

Annexe 2. Distances sémantiques entre différents concepts pour les points remarquables du relief

Nous présentons dans cette annexe le test que nous avons réalisé auprès des experts afin de déterminer les distances sémantiques entre les natures des objets représentant les points remarquables du relief définies dans la BDCARTO et dans la BDTOPO de l'IGN, et les résultats obtenus. Les distances sémantiques obtenues en utilisant la taxonomie de domaine réalisée au laboratoire COGIT sont également présentées en fin de cette annexe.

Le test présenté aux experts est le suivant :

Test pour les experts

Etant données deux bases de données géographiques, la BDCARTO et la BDTOPO, nous souhaitons les apparier en nous appuyant sur la nature des objets.

L'objectif du test est de réaliser un appariement interactif basé sur la sémantique des objets. Il s'agit de comparer chaque valeur de l'attribut nature de la BDCARTO avec chaque valeur de l'attribut nature de la BDTOPO et d'attribuer une distance sémantique comprise entre 0 et 1, où 0 signifie une ressemblance sémantique totale et 1 signifie qu'il n'y a aucune ressemblance. Les distances sémantiques sont ensuite reportées dans le Tableau 26.

Les différentes valeurs de l'attribut nature présentes dans la BDCARTO et la BDTOPO sont énumérées ci-dessous :

BDCARTO :

- Cap, pointe : cap, pointe, promontoire,
- Cirque,
- Col, passage : col, passage, cluse,
- Coteau, falaise : coteau, versant, falaise,
- Cuvette, dépression : cuvette, bassin fermé, doline, dépression,
- Dune, plage : dune, isthme, cordon littoral, plage, grève,
- Ile, îlot, presqu'île,
- Pic : pic, aiguille, piton,
- Plaine, plateau,
- Récifs : récifs, brisants,
- Rocher : chaos, rocher, escarpement rocheux,
- Sommet, crête, colline : crête, arrête, ligne de faîte, chaîne de montagne, montagne, massif rocheux, mont, colline, mamelon, sommet,
- Vallée : défilé, gorge, canyon, val, vallée, vallon, ravin, thalweg, combe,
- Volcan, cratère,

BDTOPO :

- Cap : cap, pointe, promontoire,
- Carrière : carrière, sablière,
- Cirque,
- Col : col, passage,
- Crête : crête, arête, ligne de faîte,
- Coteau, falaise : coteau, versant, falaise,
- Dépression : cuvette, bassin fermé, doline, dépression,
- Dune,
- Escarpement : barre rocheuse, escarpement rocheux, face abrupte, falaise,
- Gorge : canyon, cluse, défilé, gorge,
- Grotte : aven, cave, gouffre, grotte, habitation troglodytique,
- Ile : île, îlot, presqu'île,
- Isthme : cordon littoral, isthme,
- Montagne,
- Pic : pic, aiguille, piton,
- Plage : plage, grève,
- Plaine, plateau,
- Récifs : récifs, brisants, rochers marins,
- Rocher : chaos, rocher, pierrier, éboulis,
- Sommet : mont, colline, mamelon, sommet,
- Vallée : val, vallée, vallon, ravin, thalweg, combe,
- Volcan : cratère, volcan.

Résultats du test

Les valeurs moyennes des distances sémantiques attribuées par les douze experts que nous avons interrogés (après la diagonalisation de la matrice précédente) sont présentées dans le Tableau 27.

A titre d'exemple, nous donnons dans le Tableau 28 les valeurs des distances sémantiques calculées automatiquement à partir de la taxonomie de domaine définie au laboratoire COGIT [Abadie et Mustière, 2008].

BDTOPO / BDCARTO	Cap	Carrière	Cirque	Col	Crête	Dépression	Dune + Isthme	Escarpement	Gorges	Grotte	Ile	Montagne	Pic	Plage	Plaine, plateau	Récifs	Rochers	Sommet	Vallée	Versant	Volcan
Cap, pointe																					
Cirque																					
Col, passage																					
Coteau, falaise																					
Cuvette, dépression																					
Dune, plage																					
Ile																					
Pic																					
Plaine, plateau																					
Récifs																					
Rocher																					
Sommet,crête, colline																					
Vallée																					
Volcan, cratère																					

Tableau 26. Tableau à remplir par les experts avec les distances sémantiques entre la nature des objets issus de la BDCARTO et de la BDTOPO

BDTOPO / BDCARTO	Cap	Dune, Isthme	Plage	Ile	Récifs	Cirque	Col	Versant, coteau, falaise	Escarpement	Dépression	Pic	Sommet	Montagne	Crête	Plaine, plateau	Rochers	Vallée	Gorges	Volcan	Grotte	Carrière
Cap, pointe	**0,00**	0,83	0,90	0,98	0,83	1,00	1,00	0,91	0,95	1,00	0,94	0,98	1,00	0,98	1,00	0,93	1,00	1,00	1,00	0,99	1,00
Dune, plage	0,90	**0,10**	**0,11**	0,93	0,83	1,00	1,00	0,99	1,00	1,00	0,99	0,97	0,98	0,99	1,00	0,98	1,00	1,00	1,00	1,00	1,00
Ile	0,91	0,89	0,83	**0,00**	0,79	1,00	1,00	0,98	0,98	1,00	1,00	1,00	1,00	1,00	0,99	0,91	1,00	1,00	0,96	0,98	1,00
Récifs	0,88	0,93	0,88	0,76	**0,03**	1,00	1,00	0,96	0,88	1,00	0,99	0,99	1,00	1,00	1,00	0,83	1,00	1,00	1,00	0,96	1,00
Cirque	1,00	1,00	1,00	1,00	1,00	**0,00**	0,98	0,84	0,85	0,76	0,95	0,94	0,86	0,93	0,97	0,96	0,92	0,96	0,85	0,99	0,91
Col,passage	0,99	0,98	1,00	1,00	1,00	0,95	**0,03**	0,97	0,95	0,95	0,97	0,97	0,93	0,95	0,98	0,98	0,87	0,69	0,99	0,99	0,99
Coteau, falaise	0,98	1,00	1,00	1,00	0,98	0,86	0,98	**0,11**	0,51	0,94	0,94	0,95	0,88	0,91	0,99	0,92	0,98	0,88	0,98	0,99	0,87
Cuvette, dépression	1,00	1,00	1,00	1,00	1,00	0,67	0,98	0,98	0,94	**0,00**	0,99	0,99	0,96	0,99	0,98	0,97	0,88	0,83	0,86	0,97	0,86
Pic	1,00	0,98	1,00	0,99	1,00	0,95	0,94	0,95	0,81	0,99	**0,00**	0,57	0,71	0,77	0,99	0,82	0,99	0,98	0,84	0,99	0,99
Sommet,crête,colline	0,93	0,91	1,00	1,00	1,00	0,93	0,96	0,92	0,84	0,99	0,60	**0,14**	0,48	**0,18**	0,99	0,86	0,99	0,91	0,84	0,99	0,98
Plaine, Plateau	1,00	1,00	1,00	1,00	1,00	0,99	0,98	0,96	0,99	0,97	0,99	0,96	0,99	0,98	**0,00**	0,99	0,83	0,97	0,99	0,99	0,98
Rochers	0,93	0,99	0,99	0,94	0,83	0,97	0,98	0,88	0,67	0,99	0,93	0,94	0,93	0,98	0,99	**0,10**	0,99	0,96	0,94	0,91	0,93
Vallée	1,00	1,00	1,00	1,00	1,00	0,95	0,99	0,98	0,97	0,94	0,99	0,99	0,99	0,99	0,89	0,99	**0,03**	0,53	0,99	0,99	0,98
Volcan	1,00	1,00	1,00	0,98	1,00	0,78	0,99	0,96	0,89	0,77	0,84	0,85	0,83	0,97	0,98	0,98	0,98	0,99	**0,00**	0,99	0,98

Tableau 27. Valeurs moyennes des distances sémantiques fournies par les experts

BDTOPO / BDCARTO	Cap	Dune, Isthme	Plage	Ile	Récifs	Cirque	Col	coteau, falaise	Escarpement	Dépression	Pic	Sommet	Montagne	Crête	Plaine, plareau	Rochers	Vallée	Gorges	Volcan	Grotte	Carrière
Cap, pointe	**0,00**	**0,50**	**0,60**	**0,50**	**0,50**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**
Dune, plage	**0,50**	**0,00**	**0,00**	**0,50**	**0,50**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**
Ile	**0,50**	**0,50**	**0,60**	**0,00**	**0,50**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**
Récifs	**0,50**	**0,50**	**0,60**	**0,50**	**0,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**
Cirque	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**0,00**	**0,72**	**0,66**	**0,72**	**0,72**	**0,43**	**0,33**	**0,15**	**0,43**	**0,71**	**0,66**	**0,66**	**0,72**	**0,67**	**0,66**	**1,00**
Col,passage	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**0,66**	**0,00**	**0,50**	**0,40**	**0,50**	**0,40**	**0,50**	**0,40**	**0,40**	**0,50**	**0,50**	**0,50**	**0,40**	**0,60**	**0,50**	**1,00**
Coteau, falaise	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**0,72**	**0,66**	**0,20**	**0,33**	**0,60**	**0,66**	**0,60**	**0,66**	**0,66**	**0,60**	**0,60**	**0,60**	**0,66**	**0,60**	**0,60**	**1,00**
Cuvette, dépression	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**0,66**	**0,60**	**0,50**	**0,60**	**0,00**	**0,60**	**0,50**	**0,60**	**0,60**	**0,50**	**0,50**	**0,50**	**0,60**	**0,67**	**0,50**	**1,00**
Pic	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**0,43**	**0,66**	**0,60**	**0,66**	**0,60**	**0,00**	**0,20**	**0,33**	**0,33**	**0,60**	**0,60**	**0,60**	**0,66**	**0,60**	**0,60**	**1,00**
Sommet,crête,colline	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**0,33**	**0,60**	**0,50**	**0,60**	**0,50**	**0,20**	**0,00**	**0,20**	**0,20**	**0,50**	**0,50**	**0,50**	**0,60**	**0,50**	**0,50**	**1,00**
Plaine, Plateau	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**0,66**	**0,60**	**0,50**	**0,60**	**0,50**	**0,60**	**0,50**	**0,60**	**0,60**	**0,00**	**0,50**	**0,50**	**0,60**	**0,50**	**0,50**	**1,00**
Rochers	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**0,66**	**0,60**	**0,50**	**0,60**	**0,50**	**0,60**	**0,50**	**0,60**	**0,60**	**0,50**	**0,00**	**0,50**	**0,60**	**0,50**	**0,50**	**1,00**
Vallée	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**0,70**	**0,60**	**0,50**	**0,60**	**0,50**	**0,60**	**0,50**	**0,60**	**0,60**	**0,50**	**0,50**	**0,00**	**0,20**	**0,50**	**0,50**	**1,00**
Volcan	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**0,66**	**0,60**	**0,50**	**0,60**	**0,50**	**0,60**	**0,50**	**0,60**	**0,60**	**0,50**	**0,50**	**0,50**	**0,60**	**0,00**	**0,50**	**1,00**

Tableau 28. Valeurs des distances sémantiques calculées automatiquement à partir de la taxonomie de domaine [Abadie et Mustière, 2008]

Annexe 3. Distances sémantiques entre différents concepts pour les réseaux routiers

Dans cette annexe nous présentons les distances sémantiques déterminées entre les différents types de route définis dans la BDCARTO de l'IGN et dans MultiNet de TeleAtlas.

Le Tableau 29 montre les distances sémantiques entre différents types de route de la BDCARTO et de MultiNet. Comme nous l'avons expliqué dans le chapitre D au paragraphe D.3.2.1, afin de calculer les distances sémantiques, nous avons utilisé les attributs « Type de classement » et « Vocation » de la BDCARTO (lignes du Tableau 29) et l'attribut « FRC » de MultiNet (colonnes du Tableau 29).

BDCARTO \ MultiNet		0	1	2	3	4	5	6	7	8
Autoroute		0	0.25	0.33	0.5	0.5	0.56	0.5	0.77	0.79
Nationale		0.5	0.25	0.33	0.25	0.5	0.56	0.5	0.77	0.79
Départementale		0.5	0.5	0.33	0.25	0	0.39	0.25	0.77	0.79
Autre	Voc 1	0.25	0.37	0.42	0.5	0.5	0.56	0.5	0.77	0.79
	Voc 2	0.43	0.29	0.36	0.325	0.5	0.51	0.5	0.73	0.77
	Voc 6	0.5	0.5	0.42	0.37	0.25	0.47	0.25	0.77	0.79
	Voc 7	0.5	0.5	0.42	0.37	0.25	0.47	0	0.77	0.79
	Voc 8	0.25	0.37	0.42	0.5	0.5	0.56	0.5	0.72	0.79

Tableau 29. Distances sémantiques entre les types de route de la BDCARTO et de MultiNet

L'attribut « Type de classement » possède les valeurs suivantes :

1- autoroute,
2- route nationale,
3- route départementale,
4- autres.

Lorsque la valeur de l'attribut « Type de classement » est égale à « autres », nous analysons l'attribut « Vocation ». Ce dernier renseigne sur la hiérarchisation du réseau routier. Les valeurs possibles de cet attribut sont les suivantes :

1- type autoroutier,
2- liaison principale,
6- liaison régionale,
7- liaison locale,
8- bretelle.

L'attribut « FRC » de MultiNet possède les valeurs suivantes :

0- autoroute, route prioritaire,
1- route prioritaire moins importante qu'une autoroute,
2- autre route prioritaire,
3- route secondaire,
4- liaison locale,
5- route locale d'importance élevée,
6- route locale,
7- route locale d'importance moins élevée,
8- autre route.

Nous remarquons dans le Tableau 29 (cadres verts) que les correspondances suivantes ont une distance sémantique faible, c'est à dire qu'il y a une ressemblance sémantique entre ces types de routes :

- -autoroutes-0,
- -route nationale-1 ; route nationale-2 ; route nationale-3,
- -route départementale-2 ; route départementale-3 ; route départementale-4 ; route départementale-5 ; route départementale-6,

Ces distances sémantiques faibles coïncident avec l'analyse des données réalisée, par exemple les objets de MultiNet de type 1, 2 et 3 sont souvent des routes nationales.

En ce qui concerne l'attribut « Autre », nous remarquons que les distances sémantiques sont de moins bonne qualité, dans le sens où elles ne reflètent pas toujours la réalité. Ainsi, les distances sémantiques dans le cadre rouge sont élevées, alors que de nombreux objets de MultiNet du type 7 ou 8 correspondent dans la réalité à des objets de la BDCARTO du type « Autre » et de vocation « Voc2 », « Voc6 » ou « Voc7 ».

Les distances sémantiques pour les objets du type « Autre » sont parfois justifiées. Par exemple, la distance sémantique entre un objet de la BDCARTO du type de classement « Autre » et de vocation « Voc1 » et un objet de MultiNet de type 0 est faible (0,25), parce que les carrefours dénivelés et les bretelles dans la BDCARTO de vocation « Voc1 », sont au même niveau que les autoroutes dans la taxonomie pour les réseaux routiers.