

APPRENTISSAGE DE CONCEPTS POUR L'AIDE À L'INTERPRÉTATION DES DIFFÉRENCES DE REPRÉSENTATION D'UN MÊME PHÉNOMÈNE GÉOGRAPHIQUE

Par David SHEEREN, Laboratoire COGIT, Institut Géographique National

1 INTRODUCTION

Il existe généralement plusieurs bases de données géographiques (BDG) relatives à une même portion du territoire, chaque base étant définie pour un domaine d'application spécifique. Le contenu et la représentation des phénomènes géographiques peuvent différer d'une base à l'autre, celles-ci ayant leur propre point de vue sur le monde et étant définies pour une gamme d'échelle d'utilisation particulière. Les représentations des bâtiments, par exemple, ne sont pas les mêmes dans les bases utilisées par les urbanistes et les agronomes. Les routes sont également modélisées différemment selon qu'on les exploite à l'échelle du 1/25 000 ou à l'échelle du 1/100 000 (figure 1). Un même espace géographique peut donc être décrit différemment en fonction des besoins de l'application pour laquelle on doit le représenter et du niveau d'analyse auquel on souhaite l'exploiter.



Fig. 1 Une orthophotographie représentant le monde réel et la saisie du réseau routier selon les spécifications de trois bases de données de résolutions différentes.

L'intégration de ces différentes sources de données est aujourd'hui motivée par plusieurs raisons. D'une part, les utilisateurs cherchent à exploiter au mieux l'information géographique disponible pour leurs applications. Ils souhaitent combiner les différentes bases de données afin de tirer profit de chacune de celles-ci et effectuer des analyses multi-niveaux [1]. D'autre part, les producteurs de données désirent limiter leurs coûts de production et de maintenance. Leur préoccupation est de mettre en correspondance leurs différentes bases indépendantes pour faciliter la propagation des mises à jour et les contrôles qualité [18].

Intégrer des BD géographiques suppose, d'une part, d'étudier les différences au niveau des schémas conceptuels de données et, d'autre part, d'examiner les différences au niveau des données géométriques elles-mêmes. En plus des conflits classiques d'hétérogénéité sémantique, il est nécessaire de comprendre si la différence de représentation est justifiée ou non, pour éviter d'intégrer des erreurs dans le nouveau système et de présenter des

situations incohérentes à l'utilisateur. En générale, les différences de représentation découlent des critères de saisie et de contenu différents des BDG et sont donc parfaitement normales. Cependant, le processus de saisie n'est pas exempt d'erreurs et des différences sont également susceptibles d'apparaître pour cette raison. Par ailleurs, l'intégration peut porter sur des bases qui présentent des actualités différentes, ce qui constitue une autre origine des différences.

Il existe une abondante littérature concernant l'intégration et la fusion de bases de données classiques. Quelques contributions peuvent être trouvées dans [2,17,27]. Dans le contexte des bases de données géographiques, ce domaine de recherche est également très actif. Le processus d'intégration et son adaptation aux bases de données géographiques a été étudié [3,8]. Des algorithmes permettant d'explicitier les liens entre les objets géométriques des différentes bases (appariement) ont été développés [8,18]. Certains langages supportant la représentation multiple ont également été définis [22], de même que des nouvelles structures de données [15].

Si le besoin de justifier les différences de représentation avant d'unifier les données géométriques a également été identifié depuis longtemps [4], il n'existe pas aujourd'hui de solutions opérationnelles capables d'y répondre. Les travaux sur ce sujet ne traitent généralement que des relations spatiales et présupposent l'existence d'un ordre entre les représentations, ce qui n'est pas adapté pour des BDG de résolutions équivalentes, modélisées selon des points de vue différents [11,12,21]. C'est dans ce contexte que s'inscrit notre travail de recherche.

Notre objectif est de concevoir un système informatique capable de détecter et d'interpréter de manière automatique chaque différence de représentation d'un même phénomène géographique et ce, dans un contexte d'intégration de bases de données spatiales.

Une première version du processus que nous avons défini peut être trouvée dans [26]. Il ne sera pas détaillé dans cet article. L'approche que nous proposons est fondée sur l'utilisation des spécifications de chaque BDG pour identifier l'origine des différences. L'automatisation de l'interprétation est rendue possible en explicitant ces connaissances dans un système à base de règles [7]. La mise en correspondance des données est, quant à elle, réalisée à l'aide d'outils d'appariement, principalement géométrique. La comparaison des représentations, leur enrichissement et leur rapprochement fait appel à des outils d'analyse spatiale.

L'explicitation des connaissances provenant des spécifications et leur introduction dans le système à base de règles n'est pas toujours immédiat. Les critères de saisie et de contenu sont décrits de manière relativement formelle, dans des documents volumineux, présentant parfois des ambiguïtés ou un manque d'exhaustivité. Une analyse des spécifications ne permet donc pas toujours d'acquérir l'ensemble des connaissances nécessaires à l'interprétation. D'autres informations doivent être obtenues auprès d'experts du domaine.

Comment peut-on obtenir ces informations ? Il est bien souvent difficile pour les experts de formuler explicitement leur raisonnement et les connaissances qu'ils utilisent. C'est le problème bien connu du « goulot d'étranglement de l'acquisition des connaissances ». Dans le domaine de l'intelligence artificielle, une technique a été proposée pour y faire face : l'apprentissage supervisé [13,19,20,25]. L'utilisation de cette technique pour la question qui nous préoccupe fait l'objet de cet article.

Nous présentons dans la section 2 le problème d'apprentissage que nous avons choisi de traiter et les caractéristiques des données étudiées. Nous détaillons ensuite les différentes étapes réalisées pour détecter les différences de représentation et recueillir les exemples d'apprentissage (section 3). Le processus d'induction est exposé dans la section suivante (section 4). Il aboutit à la production de règles d'interprétation. Nous discutons des résultats et concluons ensuite l'article sur les perspectives de recherche (section 5).

2 LE PROBLÈME D'APPRENTISSAGE

Les expérimentations qui ont été menées pour cette étude ont porté sur les bâtiments de deux BDG de l'IGN : la BDTPO et la BDCARTO. La zone d'étude se situe aux environs d'Orléans (figure 2).

La BDTPO est une base de résolution métrique. Elle provient de la restitution de photographies aériennes et a été définie pour produire les cartes topographiques à l'échelle du 1/25 000. Le thème que nous avons utilisé pour effectuer nos tests date de 1998. Les spécifications de saisie sont décrites de la manière suivante (extrait) [28] : la classe « bâtiment quelconque » contient les « bâtiments en « dur » dont l'architecture ou l'aspect n'est pas industriel, agricole ou commercial. La modélisation est de type surfacique. En règle générale, les bâtiments ne sont pas généralisés, leur individualité est conservée jusqu'aux limites de la précision planimétrique (1m).[...] ».

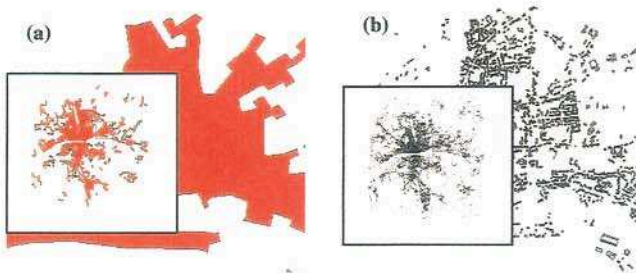


Fig. 2 Illustration des zones d'habitat de la BDCARTO (a) et des bâtiments de la BDTPO (b).

La BDCARTO est une base de résolution décimétrique. Elle est surtout utilisée pour effectuer des analyses au niveau régional et départemental. On l'utilise dans une gamme d'échelles allant du 1/100 000 au 1/250 000. Les éléments de cette base proviennent de deux sources différentes : des images SPOT pour l'occupation du sol et des cartes au 1/50 000 pour le reste des objets. Les spécifications relatives au thème du bâti sont définies de la manière suivante (extrait) [29] : « *Bâti : surface à prédominance d'habitat. Tissu urbain dense, noyaux urbains et faubourgs anciens, bâtiments formant un tissu homogène et continu [...]. Villages et hameaux importants en milieu agricole y compris les aménagements associés. La superficie minimale est fixée à 8ha. Pour les petites parcelles de bâti (surfaces inférieures à 8 hectares) : elles sont regroupées si elles sont distantes les unes des autres de moins de 100 mètres, de manière à atteindre les 8 hectares. Les parcs, bois, et forêts inférieurs à 8 hectares et associés ou inclus à une zone de bâti de plus de 8 hectares sont classés dans le poste 11 (bâti)* ». Les zones d'habitat constituent donc un poste du thème relatif à l'occupation du sol. Elles proviennent d'une interprétation des images SPOT et datent de 1993.

Vu les différences de spécifications et d'actualités entre les données, des différences de représentation entre les jeux sont susceptibles d'apparaître. Notre objectif est de pouvoir identifier l'origine de ces différences de manière automatique. L'explicitation des connaissances, en terme de règles de classification, permettant d'identifier cette origine, semble néanmoins difficile à réaliser, du moins en se limitant à l'utilisation des spécifications. D'autres connaissances sont également nécessaires et nous allons utiliser l'apprentissage supervisé pour les acquérir. Les règles qui seront obtenues automatiquement seront ensuite introduites dans un système-expert.

De manière formelle, le problème d'apprentissage supervisé peut être décrit de la manière suivante. A partir d'un ensemble d'exemples ou couples $(x_i, y_i) = (x_i, f(x_i))$, on cherche à deviner la fonction de classification $y = f(x)$. Les x_i correspondent à des observables ou dans notre contexte, des différences de représentations décrites par un ensemble d'attributs. Les y_i correspondent aux étiquettes ou classes des observables, c'est-à-dire à l'origine de chaque différence de représentation (les spécifications, les erreurs, l'actualité). A partir de ces exemples, on cherche donc à apprendre un classifieur, une fonction $f(\cdot)$, qui permettra de classer ensuite toute nouvelle différence de représentation non encore étiquetée.

Nous utilisons trois termes pour spécifier l'origine des différences de représentation. On parle d'équivalence lorsque les différences de représentation s'expliquent par des différences de spécifications. Le terme d'incohérence est réservé aux différences découlant d'erreurs de saisie ou d'appariement. Les mises à jour concernent les représentations qui ne présentent pas la même actualité.

Les différentes étapes réalisées pour détecter les différences entre les jeux de données et construire les exemples d'apprentissage sont décrites dans la partie suivante.

3 RECUEIL DES EXEMPLES D'APPRENTISSAGE

Avant de construire les exemples, il a d'abord été nécessaire de détecter les différences entre les données. Plusieurs étapes ont été réalisées pour y parvenir.

3.1 Contrôle Intra-base des spécifications

Nous avons, dans un premier temps, vérifié les spécifications pour chaque jeu de données. Nous avons ainsi contrôlé que la superficie des bâtiments saisis dans la BDTOPO était systématiquement supérieure à 1m² et que celle des zones d'habitat saisis dans la BDCARTO était supérieure à 8 hectares. Ce contrôle avait pour objectif d'identifier d'éventuelles erreurs avant la mise en correspondance des données. Aucune erreur à ce stade n'a été détectée.

3.2 Appariement des données

Les données ont ensuite été appariées. Nous avons considéré simplement que chaque bâtiment individualisé de la BDTOPO était associé à une zone de la BDCARTO s'il intersectait celle-ci. Les bâtiments appariés ont été directement considérés comme des représentations *équivalentes* (représentations dont les différences se justifient par les spécifications). Les objets non appariés constituaient les différences à interpréter.

Parmi ces différences, il était encore possible d'en interpréter directement, à l'aide des spécifications des BD. En effet, plusieurs groupes de bâtiments dans la BDTOPO n'apparaissent pas dans la BDCARTO en raison des critères d'existence différents (parcelles de bâti < 8ha). Pour vérifier ce critère il fut cependant nécessaire de transformer la représentation des objets de la base la plus détaillée. Plus précisément, nous avons simulé la BDCARTO à partir des bâtiments individualisés de la BDTOPO pour vérifier les spécifications de la première BD.

3.3 Simulation de la BDCARTO

Cette transformation a été réalisée de la manière suivante. Tous les bâtiments individualisés de la BDTOPO distants de moins de 100m ont été regroupés, de manière à former une zone d'habitat de superficie au moins égale à 8ha (critère de saisie de la BDCARTO). Ce regroupement a été réalisé à l'aide de *buffers* correspondant à une expansion des bâtiments avec un rayon de 50m (figure 3).

A l'issue de ce regroupement, nous avons éliminé les zones inférieures à 8ha afin de respecter les spécifications de la BDCARTO et terminer la simulation. Nous avons donc à ce niveau deux ensembles d'équivalences : les objets appariés, et les zones simulées inférieures à 8 ha.

Il restait néanmoins des différences de représentation entre les données qui ne pouvaient pas être justifiées directement (figure 3). Il s'agissait d'incohérences, de mises à jour (les données ayant des actualités différentes) et d'équivalences. C'est avec ces différences que nous avons poursuivi le processus pour obtenir des exemples d'apprentissage.

3.4 Agrégation de bâtiments de la BDTOPO

Les différences qu'il restait à interpréter étaient essentiellement des différences d'existence. Il s'agissait de bâtiments individualisés de la BDTOPO qui n'existaient pas dans la BDCARTO. Pour identifier leur origine, il a fallu procéder à nouveau à leur agrégation (en repartant de la représentation la plus détaillée de la BDTOPO). En effet, notre objectif était de construire des exemples pour apprendre des règles. Or, pris un à un, les bâtiments ne sont pas des exemples pertinents. C'est seulement un groupe entier de bâtiments que l'on peut interpréter. Il fut donc nécessaire de procéder à nouveau au regroupement des bâtiments de la BDTOPO, de changer de niveau de détail pour construire des représentations adéquates pour le processus d'apprentissage supervisé.

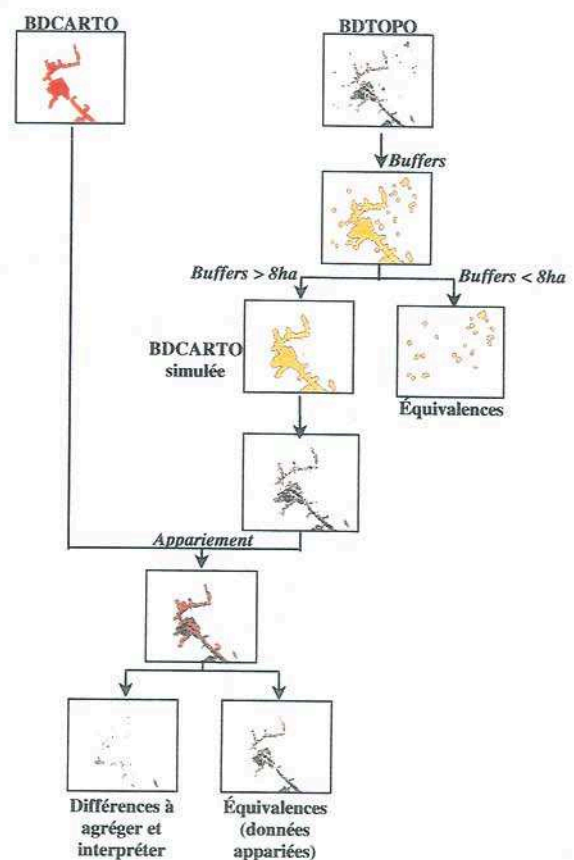


Fig. 3 Transformation de la BDTOPO en respectant les spécifications de la BDCARTO afin de mettre en évidence les différences à interpréter.

Deux méthodes différentes ont été mises en œuvre pour agréger les bâtiments. La première méthode d'agrégation est similaire à celle utilisée lors de l'étape de simulation. Nous avons regroupé les bâtiments de la BDTOPO en utilisant des *buffers* de 50 m de rayon et fusionné les éléments ayant des frontières connectées. Nous avons ensuite procédé à leur érosion en utilisant également un *buffer*

fixé cette fois à 35m. Cette érosion a été faite dans le souci d'élaborer le plus grand nombre de groupes de classes homogènes en termes de différences. Nous avons ainsi obtenu un ensemble d'agrégats qu'il suffisait ensuite de caractériser pour construire les exemples.

Après analyse des différents groupes, il nous a semblé nécessaire d'avoir recours à une autre méthode d'agrégation. Malgré la phase d'érosion, les objets créés présentaient une trop forte hétérogénéité (figure 4). Plusieurs bâtiments avaient été agrégés alors que certains auraient dû rester séparés. Ceci était particulièrement gênant pour l'interprétation car certains groupes mélangeaient à la fois des équivalences et des incohérences. Nous avons donc testé une autre méthode de regroupement.

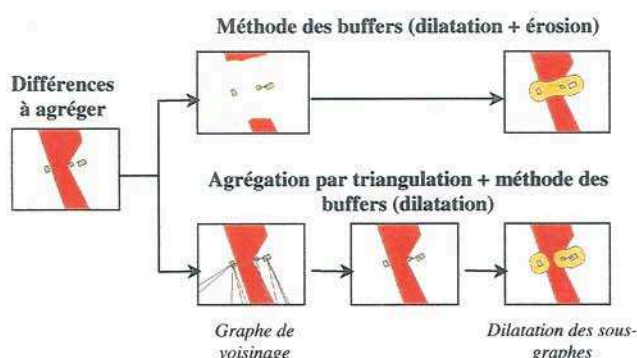


Fig. 4. Les deux méthodes testées pour agréger les bâtiments de la BDTPO.

La seconde méthode est fondée sur l'utilisation d'une triangulation de Delaunay. Elle peut être comparée aux approches de *clustering* basées sur la création de graphes [1,14]. Après avoir calculé le centre de gravité de chaque bâtiment (uniquement ceux relatifs aux différences à interpréter), nous avons réalisé la triangulation sur l'ensemble de ces nœuds. Ensuite, nous l'avons filtré en utilisant deux critères : un critère de longueur sur les arêtes, et un critère d'intersection avec la BDCARTO (figure 4 et 5). Ceci nous a permis d'obtenir des groupes plus homogènes sur lesquels un *buffer* de 15 m de rayon fut calculé. Il restait alors à caractériser ces groupes pour former les exemples d'apprentissage.

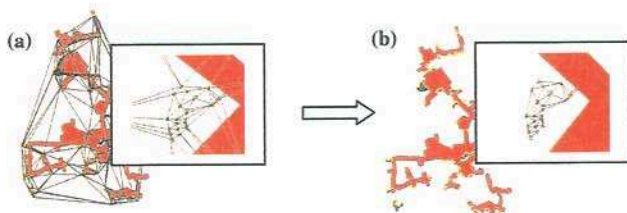


Fig. 5. La méthode de triangulation utilisée et son filtrage.

3.5 Caractérisation des agrégats

Chaque agrégat ainsi créé a été caractérisé à l'aide de huit descripteurs :

- le nombre de maisons individualisées dans le groupe ;
- la superficie du groupe ;
- le périmètre du groupe ;
- la densité des bâtiments individualisés dans le groupe ;
- la distance entre le centre de gravité du groupe et la zone d'habitat la plus proche (BDCARTO) ;
- la distance entre le bâtiment du groupe le plus proche de la zone d'habitat (BDCARTO) et cette zone d'habitat ;
- la distance entre le bâtiment du groupe le plus éloigné de la zone d'habitat (BDCARTO) et cette zone d'habitat ;
- la compacité du groupe.

Nous avons donc au terme de cette caractérisation, un ensemble d'exemples (les groupes) décrits par différents attributs.

3.6 Classification des agrégats

Avant d'utiliser les algorithmes d'apprentissage, il restait à préciser l'origine de chaque différence pour les groupes créés. Tous les observables devaient être classés pour que la production de règles (ou de l'arbre de décision) soit possible.

Pour interpréter nos exemples, nous avons utilisé les données suivantes :

- Une carte topographique à l'échelle du 1/25 000 révisée en 1991.
- Une carte à l'échelle du 1/50 000 datant de 2000.
- Une carte à l'échelle du 1/100 000 datant de 2001.

La première carte est donc plus ancienne que les deux jeux de données utilisés. Les deux autres cartes sont plus récentes. En comparant ces différentes informations, nous avons pu identifier chaque catégorie de différences pour chaque groupe créé, et compléter ainsi les exemples pour apprendre. Les différences sont soit des équivalences (déterminées ici grâce aux connaissances de l'expert, les spécifications n'étant pas assez précises à ce stade de l'analyse), des mises à jour, ou des incohérences (essentiellement des erreurs d'appariement dans ce cas). On peut trouver un extrait de quelques exemples utilisés dans le tableau 1.

Tableau 1. Extrait de quelques exemples d'apprentissage

Attributs	Valeurs			
	2	1	8	1
<i>Nb</i>	2	1	8	1
<i>Surface</i>	4501.166	1937.692	11303.30	1654.495
<i>Périmètre</i>	333.149	160.202	608.507	146.191
<i>Densité</i>	0.035	0.114	0.119	0.099
<i>Distance c gravité</i>	52.400	38.191	5.332	8.271
<i>Distance +proche</i>	8.585	40.221	6.447	8.039
<i>Distance +loin</i>	78.819	40.221	55.992	8.039
<i>Compacité</i>	0.648	1.207	0.488	1.238
<i>Classe</i>	Équivalence	Équivalence	Incohérence	Mise à jour

4 INDUCTION

4.1 Les résultats de l'apprentissage

Les hypothèses permettant de relier les classes de différences et les mesures caractérisant les groupes ont été apprises à l'aide de deux algorithmes : C4.5. [23] et Ripper [6]. Nous donnons un exemple de règle produite avec C4.5. ci-dessous :

*Si le nombre de bâtiments individualisés > 4
Et la densité des bâtiments dans le groupe ≤ 0.117
Et la compacité du groupe > 0.488
Et la distance entre le centre de gravité du groupe et la zone BDCARTO > 53.28
Et la distance entre le centre de gravité du groupe et la zone BDCARTO < 93.88
Alors, le groupe est classé en mise à jour.*

Plusieurs expérimentations ont été faites avec les deux algorithmes. Nous avons d'abord entrepris un apprentissage direct sur l'ensemble des exemples (183 dont 67 équivalences, 21 mises à jour et 95 incohérences). Nous avons également réalisé un apprentissage en deux étapes en distinguant d'abord les incohérences des autres différences, et ensuite en apprenant des règles différenciant les équivalences et les mises à jour. Nous avons enfin testé le *boosting* en fixant le nombre de classificateurs à 10. Les résultats sont présentés dans le tableau 2.

Tableau 2. Résultats des tests d'apprentissage

Algorithmes	Apprentissage direct	Apprentissage en deux étapes	Apprentissage direct avec <i>boosting</i> (10)
C4.5.	Taux d'erreurs (LVO) : 41,4%	Incohérence et autres. Taux d'erreurs (LVO) : 29,8%	Taux d'erreurs (LVO) : 34,8%
		Équivalences et mises à jour. Taux d'erreurs (LVO) : 28,4%	
Ripper	Taux d'erreurs (LVO) : 26,5%	Incohérence et autres. Taux d'erreurs (LVO) : 28,4%	Taux d'erreurs (LVO) : 26,14%
		Équivalences et mises à jour. Taux d'erreurs (LVO) : 23,3%	

Le taux d'erreur donné a été calculé par la méthode *Leave One Out*. Le principe est le même que celui de la validation croisée mais ici le nombre de passes est égal au nombre des exemples (on retire donc le premier exemple et on apprend avec les autres ; on retire ensuite le

second exemple en remettant le premier et on apprend à nouveau ; on fait ceci pour chaque exemple).

On constate que les résultats d'apprentissage sont assez différents suivant l'algorithme qu'on utilise. Ainsi, C4.5. fournit un taux d'erreurs d'environ 40% pour un apprentissage direct contre 26,5% avec Ripper. Un gain d'environ 10% est obtenu lorsqu'on réalise un apprentissage en deux étapes avec C4.5. Par contre, le résultat reste assez stable avec Ripper. La performance du classifieur obtenu avec C4.5. n'est pas réellement concluante. Ripper semble donner de meilleurs résultats mais le taux d'erreur reste quand même assez élevé (environ 25%).

4.2 Discussion

Que peut-on conclure au vu des résultats obtenus ? Plusieurs hypothèses peuvent être émises pour expliquer la performance relativement faible des classificateurs. D'abord, il est possible que les exemples soient trop peu nombreux en comparaison du nombre de descripteurs utilisés ou ces derniers sont en trop grand nombre. Il est également possible que la distribution des exemples ne soit pas suffisamment homogène et qu'il eût été préférable d'utiliser un même nombre d'exemples dans les différentes classes (équivalence, mise à jour, erreur). Les exemples peuvent aussi être trop bruités. Ce bruit peut provenir d'erreurs durant la phase de recueil des exemples et se traduire par une classification inexacte. L'expert aurait mal interprété les exemples ou ceux-ci pourraient être classés dans plusieurs catégories. Il est également possible que les exemples ne soient pas suffisamment bien caractérisés. Ils auraient une description trop pauvre et les mesures ne seraient pas assez pertinentes.

Nous pensons que la qualité des résultats est liée ici à deux raisons principales : l'existence de groupes mixtes, c'est-à-dire des exemples qui auraient pu être classés dans différentes catégories ; le manque de mesures pertinentes, qui auraient permis de mieux discriminer les exemples.

L'existence de groupes mixtes s'explique par le fait que, dans certains cas, il est difficile de distinguer clairement les équivalences des erreurs d'appariement. Il y a bien souvent une incertitude sur la classe du groupe, et la classification globale est susceptible de varier légèrement d'un expert à l'autre. On pourrait envisager de régler ce problème en créant explicitement les classes mixtes. Ceci suppose néanmoins un nombre suffisant d'exemples dans chaque classe ce qui n'est pas le cas ici. C'est pour cette raison que nous avons testé l'apprentissage en deux étapes [9,24]. Il a permis d'améliorer les résultats obtenus avec C4.5.

L'utilisation d'autres descripteurs plus pertinents semble néanmoins nécessaire car les performances des classificateurs restent faibles. Nous pensons qu'il serait intéressant par exemple d'intégrer une information relative à l'orientation des groupes. En effet, les groupes correspondant aux erreurs d'appariement sont généralement parallèles aux zones d'habitat de la BDCARTO contrairement aux équivalences (figure 6). On pourrait également introduire des informations relatives au contexte dans lequel se situe le groupe. L'apparition de nouveaux bâtiments peut s'accompagner de l'apparition de nouvelles routes et ce type d'information pourrait être utilisé pour classer les exemples avec plus de certitude.

On peut aussi envisager de faire de la stratification en associant différents coûts aux classes [10]. Ceci permettrait de tenir compte de la distribution hétérogène des exemples.

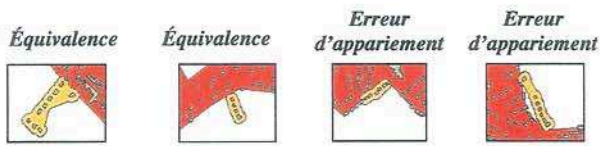


Fig. 6. L'orientation du groupe : un descripteur susceptible d'améliorer la performance des résultats d'apprentissage.

Nous pensons que les techniques d'apprentissage supervisé sont adaptées à notre contexte mais qu'il est nécessaire de poursuivre les tests et d'identifier les descripteurs pertinents pour améliorer les résultats.

5 CONCLUSION

Après avoir exposé la problématique du maintien de la cohérence entre les données lors d'un processus d'intégration de BD géographiques, nous avons mis en évidence

le manque de solutions opérationnelles existant aujourd'hui. Ayant ainsi justifié l'intérêt de cette recherche, nous nous sommes fixés comme objectif de définir un système automatique capable de détecter et d'interpréter les différences de représentation d'un même phénomène géographique. L'approche que nous proposons est fondée sur l'utilisation d'un système à base de connaissances qui s'appuie sur les spécifications de chacune des BDG. Ces spécifications peuvent cependant être imprécises et incomplètes. Dans certains cas, elles ne suffisent pas pour justifier toutes les différences entre les bases. L'explicitation d'autres connaissances du domaine est également nécessaire. Pour les acquérir, nous avons étudié les possibilités d'utiliser l'apprentissage automatique. Des expérimentations ont été mises en œuvre sur deux jeux de données différents de l'IGN. Ces techniques d'apprentissage semblent bien adaptées à notre contexte mais des investigations supplémentaires doivent être entreprises pour mieux caractériser les exemples et améliorer les résultats. Une meilleure analyse du raisonnement suivi par l'expert pourrait être envisagée pour décomposer le processus d'interprétation en plusieurs étapes et apprendre plus facilement.

A terme, l'étude des différences de représentation devrait permettre, en plus de l'intégration cohérente des données, d'améliorer la qualité de chaque BD, et d'enrichir la description des spécifications.

Remerciements

Nous tenons à remercier Sébastien Mustière du laboratoire COGIT (IGN) et Jean-Daniel Zucker du LIM&BIO (Université Paris Nord) pour leurs commentaires et suggestions relatifs aux expérimentations menées.

Bibliographie

- [1] Anders K-H., Sester M. and Fritsch D. 1999. Analysis of settlement structures by graph-based clustering, In *Proceedings of the Semantic Modelling for the Acquisition of Topographic Information from Images and Maps Conference (SMATI'99)*.
- [2] Batini C., Lenzerini M and Navathe S.B. 1986. A comparative analysis of methodologies for database schema integration, *ACM Computing Surveys*, 18(4), pp. 323-364.
- [3] Branki T. and Defude B. 1998. Data and Metadata: two-dimensional integration of heterogeneous spatial databases, In *Proceedings of the 8th International Symposium on Spatial Data Handling (SDH'98)*, Vancouver, Canada, pp. 172-179.
- [4] Buttenfield B.P. and Delotto J.S. 1989. *Multiple representations*, Report for the specialists meeting, National Center for Geographic Information and Analysis (NCGIA), Technical paper 89-3, 1989.
- [5] Car A. and Frank A. 1994. Modelling a Hierarchy of Space Applied to Large Road Networks. In *Proc. of Int. Workshop on Advanced Research in Geographic Information Systems*, Springer-Verlag, pp.15-24.
- [6] Cohen W. 1995. Fast Effective Rule Induction, In *Proceedings of the 12th International Conference on Machine Learning (ICML'95)*, Lake Tahoe, USA, Morgan Kaufmann, pp. 115-123.
- [7] David J.-M., Krivine J.-P. et Simmons R. 1993. *Second generation Expert Systems*. Berlin : Springer-Verlag.
- [8] Devogele T. 1997. *Processus d'intégration et d'appariement de bases de données Géographiques. Application à une base de données routières multi-échelles*, Thèse de doctorat en Informatique, Université de Versailles, 205 p.

- [9] Dietterich T.G. and Bakiri G. 1995. Solving multiclass learning problems via error-correcting output codes, *Journal of Artificial Intelligence Research*, 2, pp. 263-286.
- [10] Domingos P. 1999. Metacost : a general method for making classifiers cost-sensitive, In *Proceedings of Knowledge Discovery and Data Mining Conference (KDD'99)*, New York.
- [11] Egenhofer M.J., Clementini E. & Di Felice P. 1994. Evaluating inconsistencies among multiple representations, In *Proceedings of the Sixth International Symposium on Spatial Data Handling (SDH'94)*, Edinburgh, Scotland, pp. 901-920.
- [12] El-Geresy B.A. and Abdelmoty A.I. 1998. A Qualitative Approach to Integration in Spatial Databases. In *Proceedings of the 9th International Conference on Database and Expert Systems Applications (DEXA'98)*, Springer-Verlag, Lecture Notes in Computer Science 1460, pp. 280-289.
- [13] Esposito F., Lanza A., Malerba D. and Semeraro G. 1997. Machine learning for map interpretation : an intelligent tool for environmental planning, *Applied Artificial Intelligence*, 11(7-8), pp. 673-696.
- [14] Estivill-Castro V. and Lee I. 2002. Multi-level clustering and its visualization for exploratory spatial analysis, *Geoinformatica*, 6(2), pp. 123-152.
- [15] Kidner D.B. and Jones C. B. 1994. A Deductive Object-Oriented GIS for Handling Multiple Representations, In *Proc. of the 6th International Symposium on Spatial Data Handling (SDH'94)*, Edinburgh, Scotland, pp. 882-900.
- [16] Kim W. and Seo J. 1991. Classifying schematic and data heterogeneity in multidatabase system, *IEEE Computer*, 24(12), pp. 12-18.
- [17] Larson J., Navathe S.B. and R. Elmasri 1989. A theory of attribute equivalence in databases with application to schema integration, *IEEE Transaction on Software Engineering*, 15(4), pp. 449-463.
- [18] Lemarié C. and Badard T. 2001. Cartographic database updating. In *Proc of Int. Cartographic Conference (ICC'2001)*, vol. 2, pp.1376-1385.
- [19] Mitchell T.M. 1997. *Machine Learning*, McGraw-Hill International Editions.
- [20] Mustière S. 2001. *Apprentissage supervisé pour la généralisation cartographique*, Thèse de doctorat en Informatique, Université Pierre et Marie Curie, Paris VI, 241 p.
- [21] Paiva J.A. 1998. *Topological equivalence and similarity in multi-representation geographic databases*, PhD Thesis in Spatial Information Science and Engineering, University of Maine, 188 p.
- [22] Parent C., Spaccapietra S., Zimanyi E., Donini P. and Plazanet C. 1998. Modeling spatial data in the MADS conceptual model, In *Proceedings of the 8th International Symposium on Spatial Data Handling (SDH'98)*, Vancouver, Canada, pp. 138-150.
- [23] Quinlan J.R. 1993. *C4.5 : Programs for machine learning*. Morgan Kaufmann.
- [24] Ricci F. and Aha D.W. 1997. *Extending local learners with error-correcting output codes*, Technical Report, Naval Center for Applied Research in Artificial Intelligence.
- [25] Sester M. 2000. Knowledge Acquisition for the Automatic Interpretation of Spatial Data, *International Journal of Geographical Information Science*, 14(1), pp. 1-24.
- [26] Sheeren D. 2002. L'appariement pour la constitution de bases de données géographiques multi-résolutions. Vers une interprétation des différences de représentations, *Revue Internationale de Géomatique*, 12(2), pp. 151-168.
- [27] Sheth A. and Larson J. 1990. Federated database systems for managing distributed, heterogeneous and autonomous databases, *ACM Computing Surveys*, 22(3), pp. 183-236.
- [28] Spécifications détaillées de la BDTOPO", version 3.1., IGN, Saint- Mandé.
- [29] Spécifications de contenu de la BDCARTO", version 2.0., IGN, Saint-Mandé.
- [30] Walter V. and Fritsch D. 1999. Matching Spatial Data Sets: a Statistical Approach, *International Journal of Geographical Information Science*, 13(5), pp. 445-473.
- [31] Zucker J.-D. (to appear). *A Grounded Theory of abstraction in AI*, *Special Issue on Abstraction*, Zeki and Saitta (Eds), Philosophical Transactions of the Royal Society of London, Series B.