Alba Fuga
Laboratoire Sisyphe, UMR 7619, Université Pierre et Marie Curie Paris VI,
e-mail : alba.fuga@neuf.fr

ZONES D'INTERFAÇAGE GEOGRAPHIQUE ET METHODE DE COMPARAISON AUTOMATIQUE DE DONNEES

GEOGRAPHIC INTERFACE AREAS AND METHODS OF AUTOMATIC DATA COMPARISON

RÉSUMÉ. Dans le cadre de l'analyse d'un territoire sur le plan géophysique, et dans le but d'en identifier les ressources naturelles, de nombreuses informations sont acquises. Il s'agit de classifier, caractériser, et interpréter des mesures obtenues par campagnes de navigation sismique, par carottage, acquises dans des puits de forage, ou encore par campagnes de prélèvement d'échantillons.

La problématique qui accompagne cette analyse de territoire concerne d'une part la gestion des données complexes et volumineuses dans leurs lieux de stockage. D'autre part la question de l'aide à l'interprétation est posée lorsqu'il s'agit de classifier et comparer de la manière la plus automatique possible ces représentations et caractérisations du territoire. Dans ce contexte ont été développés la méthodologie et les programmes LAC (Logiciel Automatique de Comparaisons).

L'un des mécanismes mis en place dans cette méthodologie concerne l'interaction entre un système de filtrage à tamis de critères de comparaison et un système de seuillage pour définir une résolution de comparaison et de regroupement. Cette résolution représente un élément clé de l'analyse car elle permet de détecter des zones d'interfaçage, de frontière, ou de changement de milieu, tout en qualifiant un caractère plus ou moins progressif de ces frontières. Après une première description de la méthodologie LAC, nous voyons de quelle manière elle s'applique aux données de géosciences et comment on peut la décliner sur le plan géographique.

MOTS CLÉ: Ressemblance, métrique de similarité, groupe de similarité, résolution, seuil de tolérance, zone d'interfaçage

ABSTRACT. Numerous data are registered when a territory has to be analyzed in its geophysical configuration and behavior, in order to identify natural resources. All this information acquired through seismic navigation surveys, drilling, or sample withdraw, needs to be classified, characterized and interpreted.

This territory analysis process goes along with the management of complex and voluminous data. Automatic interpretation assistance is moreover needed to find a methodology that could optimize and automatize big geoscience data comparisons in order to automatically characterize a territory. In this context LAC (Logiciel Automatique de Comparaisons) methodology and programs have been designed and coded.

The interaction between comparison criteria filtering system and a threshold adjustment

system is one of the mechanisms involved in this methodology, in order to define a comparison and clustering resolution. This resolution is a key analysis element which makes geographic interface areas, border areas, or environment changes be detected. After a first description of LAC methodology, the way it is applied to geoscience data, and how it can be developed in the geographic field, will be explained.

KEY WORDS: Similarity, similarity metrics, similarity groups, resolution, tolerance threshold, geographic interface area.

INTRODUCTION : L'APPROCHE LAC DE COMPARAISON AUTOMATIQUE DE DONNÉES

La méthodologie générale est basée sur des algorithmes de classification automatique couplés à une série de mesures de similarité hiérarchisées.

L'objectif de la méthodologie LAC est de fournir un outil d'analyse de l'information territoriale et géophysique par la mesure de ressemblance, et par la réalisation de comparaisons suivant au plus près les raisonnements pouvant être faits par les experts métier.

La ressemblance peut être perçue comme le jugement ou l'évaluation de la proximité conceptuelle entre objets, selon une résolution donnée. On peut prendre l'exemple de la ressemblance en positionnement. Deux objets sont semblables en positionnement s'ils sont "suffisamment" proches, selon la distance euclidienne, orthodromie, ou loxodromie… L'adverbe "suffisamment" est relatif à la résolution à laquelle on juge de la ressemblance, ou encore au facteur de tolérance utilisé. On peut considérer que deux points de l'espace sont semblables si la distance euclidienne qui les sépare est inférieure à 1 unité de mesure. Il s'agit d'une notion extrêmement proche de la notion de limite en analyse mathématique. Deux objets sont proches s'ils tendent l'un vers l'autre. Ici, la différence est que l'on ne se place pas sur l'ellipsoïde, ou dans l'espace euclidien à 3 dimensions, mais on se place dans un "espace de similarité", c'est-à-dire un espace ayant autant de dimensions que le nombre de critères de comparaison, et régi par la métrique induite un l'arbre de décision.

Un arbre de décision est ici défini comme un arbre binaire, ou automate, permettant de savoir quelle métrique de similarité et quels facteurs de tolérance on applique si à l'étape précédente les objets comparés ont été identifiés comme similaires, et quelle métrique et facteurs on applique sinon. Un tel arbre permet la comparaison multicritères, aux facteurs de tolérance près. Il permet également de donner un indicateur signifiant le degré de ressemblance des objets comparés. Cet indicateur est relatif au cheminement de la ressemblance dans l'arbre. Cet arbre ou automate constitue lui-même un algorithme et donc une métrique de similarité évoluée, ou composée.

L'automatisation informatique de ces mesures de ressemblance et de ces comparaisons assure le croisement multicritères d'un grand nombre d'informations en peu de temps. Par exemple, le système croise en 5 minutes de comparaison 4000 lignes de navigation sismique. Une ligne de navigation sismique est un objet géo-scientifique complexe car mettant en jeu plus de 20 critères de comparaison hétérogènes pour la plupart.

LA MODÉLISATION ET LES CRITÈRES DE COMPARAISON

Le phénomène que l'on souhaite analyser, ou le territoire que l'on souhaite explorer est modélisable par un ensemble d'attributs et de caractéristiques. Certaines de ces caractéristiques constituent des critères pertinents de comparaison entre les différentes configurations d'un même phénomène sur ce territoire. Une acquisition est une mesure, un enregistrement, de l'un de ces critères de comparaison. Elle est une réalisation physique du critère. Cependant, comme toute mesure et tout enregistrement, elle a une nature et une unité de mesure. Par "donnée", on entend ici l'ensemble des informations, acquisitions

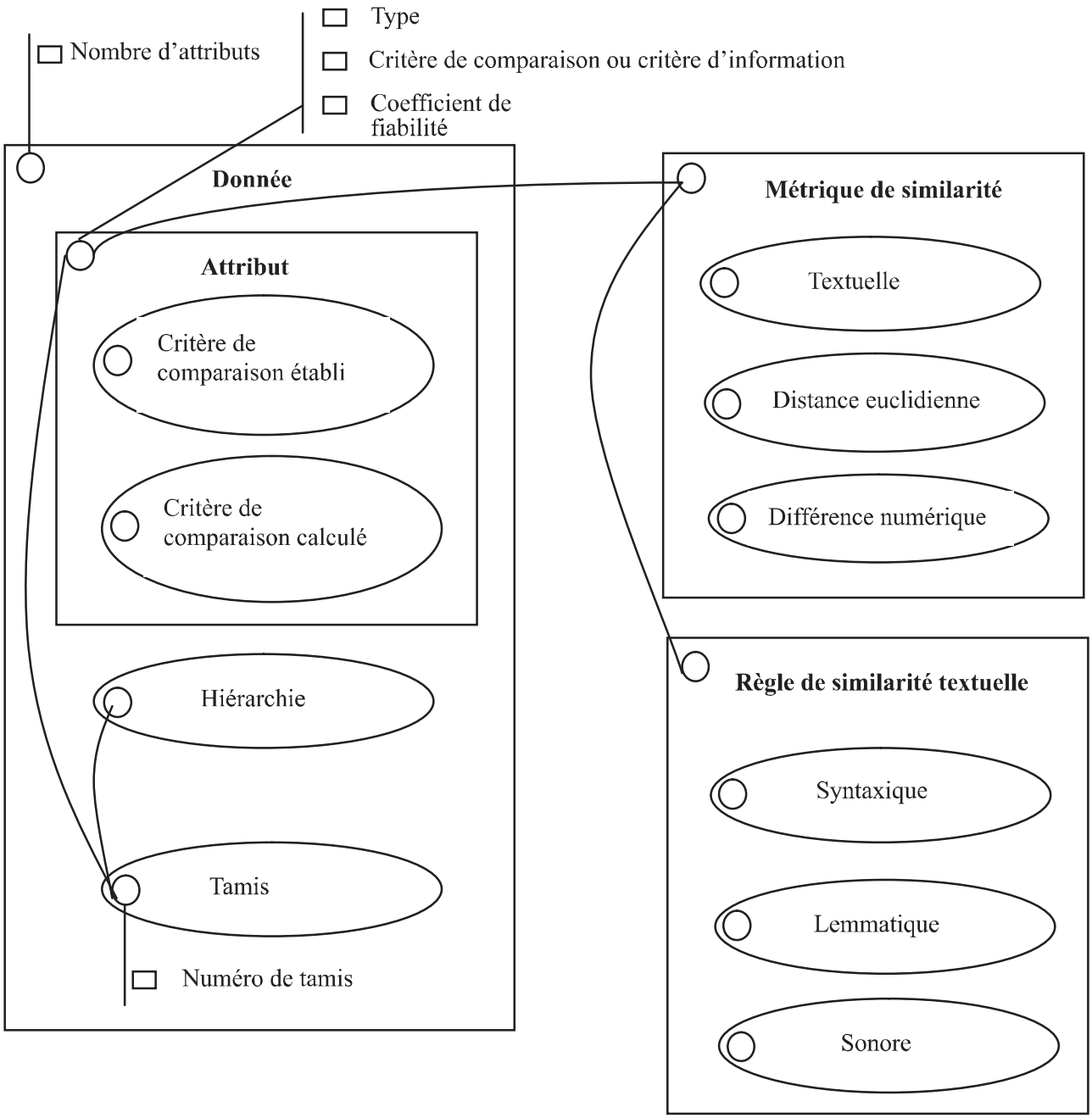

Fig. 1 : Schéma HBDS de la méthodologie LAC

directes ou déduites réalisant un modèle d'un phénomène physique, ou une configuration physique.

Dans l'approche LAC, la première phase de la méthode de comparaison est la modélisation de la donnée ou du phénomène. Ensuite, il est nécessaire d'identifier les différentes catégories d'attributs, en distinguant les attributs critères de comparaison des attributs de renseignement qui ne serviront pas à la comparaison des données.

Dans une deuxième phase méthodologique, les critères de comparaison établis ou calculés sont classifiés eux-mêmes. Il s'agit de les ranger par catégorie, par tamis. Les critères de la classification attributaires sont la fiabilité mathématique, la pertinence de comparaison (ou degré de caractérisation de la donnée), la robustesse du facteur de tolérance.

La fiabilité mathématique concerne la formule mathématique de calcul de l'attribut et son adéquation plus ou moins grande avec le phénomène modélisé. Par exemple, une longueur pour une ligne de navigation sismique peut être calculée soit par interpolation et abscisse curviligne, soit par sommation des segments entre points

de tir. L'abscisse curviligne est le moyen mathématique approchant le plus la réalité de la longueur d'une ligne, mais ce n'est pas forcément l'appareil mathématique utilisé, notamment s'il faut faire face à une problématique de performances temporelles des comparaisons et calculs d'attributs.

Le critère de pertinence de comparaison vise à ordonner les attributs selon leur capacité à caractériser le phénomène ou la donnée modélisée. Par exemple, pour une modélisation de cours d'eau, la couleur des pierres dans l'eau serait un critère moins caractérisant que sa profondeur ou son débit à des points précis, ou en moyenne.

Concernant la robustesse des facteurs de tolérance qui sont ces seuils de résolution nous permettant de savoir à partir de quelle proximité les objets sont suffisamment similaires, leur fiabilité peut dépendre de la zone géographique où est prise la donnée. Par exemple, les noms donnés aux objets géographiques peuvent être plus ou moins éloignés d'un nom standard selon le pays dans lequel ils sont saisis. Alors on devra prendre en compte dans le seuil de tolérance le pays d'acquisition des données. Pour l'exemple de modélisation d'un cours d'eau, le débit change en fonction de la saison, selon le climat de la zone géographique ciblée. Le seuil de tolérance choisi pour comparer des débits de cours d'eau n'est donc pas aussi robuste et générique qu'un attribut que serait le nombre de barrages sur le cours d'eau à une date donnée, par exemple.

Une fois que les attributs sont classés dans des tamis, et que les tamis sont eux-mêmes hiérarchisés, on peut aborder la question des métriques de similarité attributaire et élémentaire.

LES MÉTRIQUES ATTRIBUTAIRES DE SIMILARITÉ – SPÉCIALISATION EN FONCTION DES CRITÈRES DE COMPARAISON

A chaque nature d'acquisition correspondent une unité de mesure et une métrique. On peut considérer qu'une métrique de similarité est une formule mathématique, ou algorithme permettant de comparer deux objets selon un critère unique afin d'évaluer s'ils sont similaires ou non. La comparaison se fait au facteur de tolérance près. Un simple comparateur ">" (supérieur ou égal) peut constituer une métrique de similarité. Par exemple, pour comparer deux profondeurs totales de puits de forage, on utilise une simple soustraction métrique. Pour comparer deux positions géo-référencées, on utilise une distance euclidienne si on a à faire à des coordonnées planes, ou bien une orthodromie ou une loxodromie si on manipule des coordonnées géographiques.

De la même manière, on définit dans la méthodologie LAC un ensemble de comparateurs. Ils sont applicables aux métadonnées que représentent les conditions d'acquisition comme un nom de campagne sismique, un nom de ligne de navigation sismique ou des noms de documents et rapports techniques, d'avis donnés sur les conditions d'acquisition ou sur la fiabilité des mesures. Les comparateurs s'appliquent aussi aux métadonnées déduites, c'est-à-dire calculées à partir d'acquisitions. Il peut s'agir de préfixes, suffixes déduits, de noms d'auteurs extraits, d'enveloppes convexes, centroïdes, azimuts et autres éléments pouvant être déduits et calculés depuis les acquisitions. Ces comparateurs et métriques concernent donc aussi bien des données numériques, géométriques que textuelles.

Un ensemble de comparateurs peut donc être attribué à chaque tamis de critères de comparaison, en fonction de la nature et de la fonction des critères qu'il contient. Par exemple, afin de rechercher certains noms de roches dans des titres de documents divers, il est nécessaire :

- De disposer d'un dictionnaire de synonymes, ou abréviations connues de roches que l'on souhaite chercher
- De prendre en compte le fait que ces noms peuvent être écrits dans les titres avec des insertions de caractères spéciaux

– De prendre en compte qu'il arrive parfois qu'on trouve dans ces titres un système de numérotation avec la présence potentielle de zéros non significatifs

L'analyse de l'information dépend donc d'une part de la modélisation du territoire et du phénomène, d'autre part de métriques de similarité spécifiques aux différents attributs. Elle dépend également de trois autres éléments : du type de classification que l'on effectue, de la hiérarchie des critères de comparaison, et du paramétrage des seuils de tolérance pour les comparateurs. Par la suite, on portera l'attention sur la notion de résolution que contient cette méthodologie, ainsi que sur ce qu'elle implique en termes d'analyse de l'information.

TROIS STRATÉGIES DE CLASSIFICATION – PRINCIPE DE RÉSOLUTION

On peut distinguer dans cette approche trois types de regroupements : les couples, les groupes asymétriques, et les clusters. Chacune de ces méthodes est appropriée à une problématique spécifique. Tout comme l'élaboration de métriques de similarité sur mesure selon la nature et la fonction des critères de comparaison, on attribue une méthode de classification spécifique à une problématique donnée. L'approche LAC est donc une approche adaptative. On répertorie les différentes problématiques, et pour chacune, on préconise une configuration donnée de LAC.

Par exemple, dans certaines bases de données, les informations peuvent être lacunaires, c'est-à-dire que tous les attributs caractérisant une donnée ne sont pas renseignés. Comment traiter alors ces attributs vides ? Dans les métriques de similarité attributaire, on peut considérer que les deux attributs comparés dont l'un vide sont soit exactement similaires, soit exactement différents. Cependant, selon la position hiérarchique de l'attribut dans son tamis, et du tamis parmi les autres tamis, si les attributs lacunaires ne sont pas bloquants, cela peut causer une baisse de précision dans la comparaison. Il faut donc encore choisir le comportement à adopter par rapport aux données lacunaires selon le contexte, la configuration des données, et la problématique visée.

Les différentes problématiques envisagées jusqu'à présent dans le cadre de la gestion du territoire, de l'analyse des risques et d'une meilleure exploitation des ressources naturelles, sont :

– l'harmonisation des bases de données afin d'enlever des doublons et de ne garder que les données les plus précises

– la réconciliation d'informations et de différents supports

– la reconstitution et le rattachement documentaire

– le géo-référencement

– le croisement multicritère pour l'analyse et l'interprétation des phénomènes

Couples et réconciliation de sources

Les couples sont des regroupements deux à deux de données, pouvant suivre des contraintes de regroupement comme la règle "on ne doit jamais retrouver dans un même couple deux données provenant d'une même source". Ce type de procédé est nécessaire lorsqu'on l'on souhaite fusionner ou réconcilier des bases de données. Il peut servir aussi lors du chargement de nouvelles données dans une base de référence, pour savoir si les données à charger ne sont pas déjà contenues en base. Ce procédé est aussi utile pour savoir quelles sont les données qu'on veut comparer une base que l'on possède déjà aux métadonnées d'une base qu'on souhaiterait acheter, afin de voir quelles données il est réellement nécessaire d'acheter.

Fusionner des informations concernant une même zone géographique demande de résoudre les cas de recouvrement

d'informations. Parfois des acquisitions peuvent avoir été faites par différentes technologies, ou méthodologies, différents traitements, notamment d'analyse du signal, peuvent avoir été appliqués sur les données. Doit-on alors fusionner les différentes bases de données en réalisant simplement une union d'ensembles, ou doit-on les fusionner de manière plus sélective afin de ne garder que les informations les plus complètes, de la meilleure qualité ?

La seconde solution permet d'optimiser notre aptitude à lire et analyser, puis prendre des décisions sur ces données. La classification par couplage permet donc la réconciliation de différentes sources de données. Elle permet aussi de vérifier s'il n'y a pas eu perte de données lors de migrations de bases ou de changements de support de stockage de l'information.

Groupes asymétriques et rattachements

Les groupes asymétriques correspondent à une situation où l'on souhaite rattacher des informations par recoupements afin de former un puzzle complet des données dont on dispose. Par exemple, on peut posséder d'un côté des cartes géo-référencées d'une zone, d'un autre côté des identifiants de puits de forage, des noms de puits, des rapports techniques de forage, des rapports et études de zones à risques naturels. Toutes ces données et ces rapports peuvent être nombreux, volumineux, et la liaison des éléments se référant aux mêmes phénomènes ou objets physiques peut être quasi impossible de manière manuelle. Notre approche permet ici de se baser par exemple sur un puits de forage et de lui rattacher l'ensemble des documents techniques dans les titres desquels on retrouve le nom de ce puits approximativement écrit, ou ses coordonnées plus ou moins exactement saisies. Le terme "approximativement" fait référence aux seuils de résolution exposés plus haut. Nous pouvons alors rattacher à une zone connue pour un risque naturel spécifique tous les puits qui ont un positionnement, ou des caractéristiques permettant de considérer qu'ils sont rattachables à cette zone. On construit donc de manière successive des relations 1-N d'appartenance, c'est-à-dire un puits lié à plusieurs documents, une zone liée à plusieurs puits.

Clustering, propagation, harmonisation

Le troisième type de regroupement utilisé est la classification hiérarchique ascendante par densité. La notion de densité utilisée est basée sur la mesure de similarité élémentaire suivant l'approche LAC. Il s'agit d'une mesure de similarité sur les objets, les futurs éléments de groupes, avec tous les attributs.

Dans cet algorithme de classification, on attribue dès le départ à chaque donnée un numéro de cluster. Au début, elles ont toutes le numéro du cluster inexistant. Ensuite, on compare la première donnée de la liste aux autres. Si on trouve des données qui lui sont suffisamment similaires (une donnée suffit), alors on affecte à ces données le même numéro de cluster, différent du numéro du cluster vide. Lorsqu'on en a fini avec la première donnée de la liste, on continue avec la deuxième, "donnée courante", seulement si elle porte toujours le numéro du cluster inexistant, donc si elle n'a pas déjà été affectée à un cluster existant. Si une donnée est suffisamment similaire à une donnée déjà contenue dans un cluster, alors nous avons deux possibilités. Soit la donnée courante porte le numéro du cluster inexistant, et n'est pas dans un cluster avec

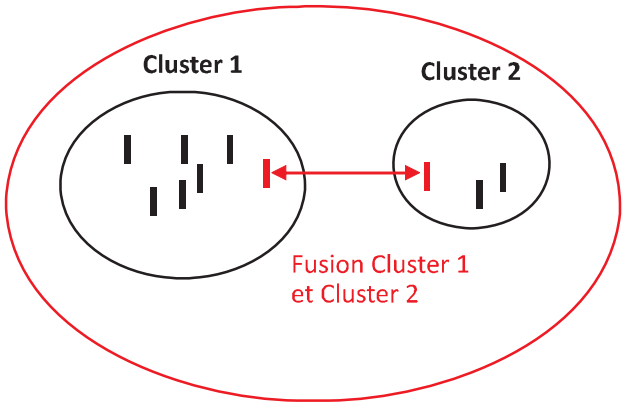

Fig. 2. Exemple de fusion entre deux clusters. Phénomène de "contagion"

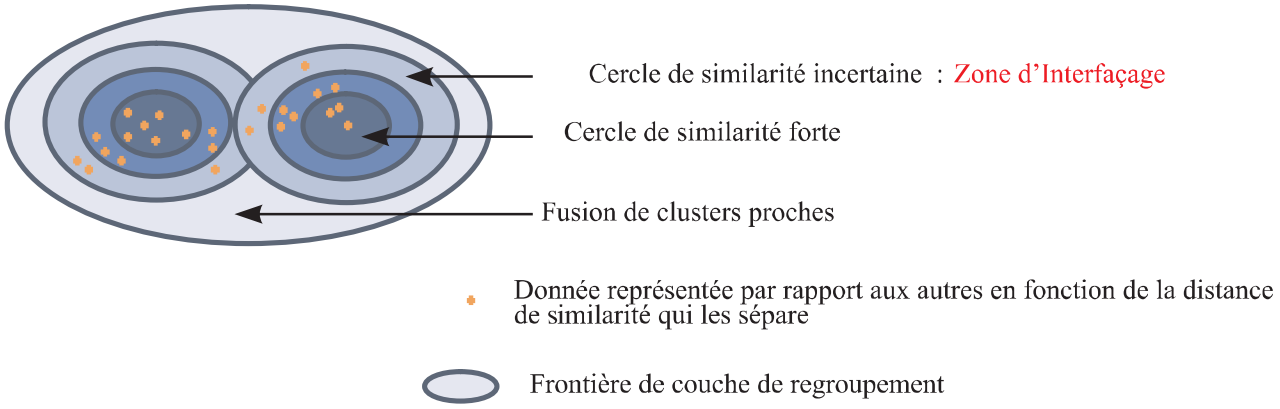

Fig. 3. Représentation de données dans un graphique de similarité, avec les contours de clusters formés selon différents vecteurs de similarité

d'autres données. On peut alors directement l'affecter au cluster de la donnée qui lui ressemble. Soit la donnée courante est déjà dans un cluster différent du cluster de la donnée qui lui est similaire. Il faut alors fusionner les deux clusters.

Quelles que soient les sources des informations et leur nombre, il s'agit de regrouper les objets similaires selon le vecteur de résolution. Il est possible de fusionner des groupes dont les données extrêmes sont au-delà du seuil de résolution, si d'autres données sont suffisamment proches. Ici on aborde une notion de continuité entre objets composites, dans le domaine de la similarité. Ces fusions peuvent se comporter comme des phénomènes de propagation par voisinage. Selon cette cartographie d'entités complexes et composites, il est possible de prendre, ou non, des décisions de fusion, d'intégration ou de séparation. Une autre possibilité est de choisir, ou construire un représentant d'un cluster, comme c'est le cas lorsqu'on raccorde en continuité deux lignes de navigation sismique si l'une des lignes possède des coordonnées de points de tir légèrement translatés.

La particularité de ces algorithmes de classification est leur couplage avec un système de filtrage qui correspond à la mise en tamis hiérarchisés des critères de comparaison dans les premières phases méthodologiques, ainsi qu'à l'affectation de métriques attributaires spécialisées à aux différentes natures de critères. La classification est automatique, ainsi que l'application du système de filtrage et des mesures. Par contre, il faut mettre l'accent sur le fait que la modélisation préalable du phénomène et la mise en tamis des critères dépendent d'un travail d'intelligence humaine.

Résolution et zone d'interfaçage

Le choix entre deux lignes de navigation sismique par exemple dépend de la résolution à laquelle on regarde ces lignes. Il s'agit de la résolution de similarité. Selon la méthodologie LAC, la similarité est mesurée de manière attributaire par les métriques de similarité spécifiques à chaque nature de critère de comparaison, mais également de manière élémentaire par association de ces mesures attributaires. La mesure élémentaire dépend d'une hiérarchie et d'un classement que l'on effectue entre les critères de comparaison, selon leur potentiel discriminatoire, et leur fiabilité.

On considère alors comme similaires deux objets dont la mesure de similarité est supérieure à un seuil de résolution pouvant être défini par l'utilisateur souhaitant analyser les données.

Ce seuil de résolution peut être défini comme un vecteur de seuils de résolution attributaire, chacun définissant une résolution attributaire sur un critère de comparaison du modèle. Par exemple, pour analyser une zone géographique sur laquelle

on a obtenu des traces de signaux à partir de géophones et sismographes, on pourra considérer que pour deux signaux similaires (même positionnement, mêmes longueurs d'ondes reçues), la trace la plus longue sera la plus complète, et la trace ayant le pas d'étalonnage le plus petit sera la plus précise. Si on considère que la précision est plus importante pour une étude de territoire, on considèrera que la trace la plus précise, même si elle est moins complète, prédominera.

Le paramétrage du vecteur de résolution permet de définir non seulement le moment à partir duquel on discrimine différents groupes, mais aussi de définir la limite d'interfaçage entre les différents groupes. Si l'on compare donc des données dans le but de les harmoniser, retirer les redondances, faire varier le vecteur de similarité permet de mettre en évidence des caractéristiques de dispersion de celles-ci, et de distinguer différentes cercles de certitude dans un même cluster.

En outre, il est nécessaire de remarquer que ces traitements appliqués à des données géographiques et géophysiques vont bien au delà du traitement de positionnement des données en deux ou trois dimensions. Dans cette approche, on est capable de simuler des phénomènes de regroupement en prenant en compte des paramètres comme des débits, des descriptions textuelles, des couleurs, des profondeurs, un âge, des types de roches et tout autre élément caractérisant la donnée.

Il s'agit d'un "positionnement" géographique étendu. Ce qui nous permet de reconnaître un objet, de le distinguer des autres est la représentation que l'on s'en fait, et notre manière de le placer, de le positionner par rapport aux autres. Dans cette approche, les coordonnées géographiques, projetées ou non, sont complétées par autant d'autres "coordonnées", critères de comparaison, qui nous permettent de construire une représentation plus proche du territoire ou du phénomène réel.

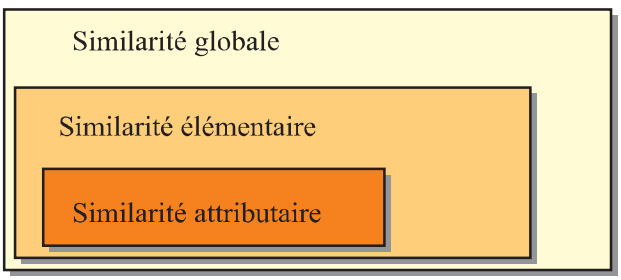

Fig. 4. Les trois échelles de mesure de Ressemblance

Ici, il est intéressant d'aborder la question de la visualisation des ces données complexes car constituées de nombreux attributs servant au traitement. La carte à deux dimensions serait un premier outil de représentation, mais très rapidement limité. En effet, si l'objectif d'un traitement est de retrouver les erreurs de positionnement géographique des données grâce à des comparaisons sur d'autres critères les caractérisant, alors sur une carte à deux dimensions, certains éléments appartenant au même cluster seraient "regroupés" de manière spatialement discontinue.

Ce type de représentation par carte répond à un besoin de placer les groupes les uns par rapport aux autres. Il s'agit d'un besoin d'une vision globale du traitement et des groupes. Dans le cas de la carte géographique classique, une représentation possible serait sorte une anamorphose en fonction des mesures de similarité entre les clusters. Il s'agirait d'adopter en premier le point de vue du groupe, de placer tous les objets du groupe les uns par rapport aux autres selon les mesures de similarité, comme si la similarité élémentaire (par opposition ou attributaire) était une "force d'attraction".

Une fois cette dispersion par cluster effectuée, il s'agirait de placer les clusters les uns par rapport aux autres selon des mesures de similarité entre clusters. Ces distances entre groupes sont également liées aux zones d'interfaçage entre lesdits groupes. Un autre intérêt de ce type de représentation peut être trouvé dans le fait que les zones d'interfaçage, zones frontalières sont alors représentées selon la dispersion des éléments et des groupes dans cet espace de similarité.

La question que l'on peut se poser concerne alors le lien faisable entre une carte

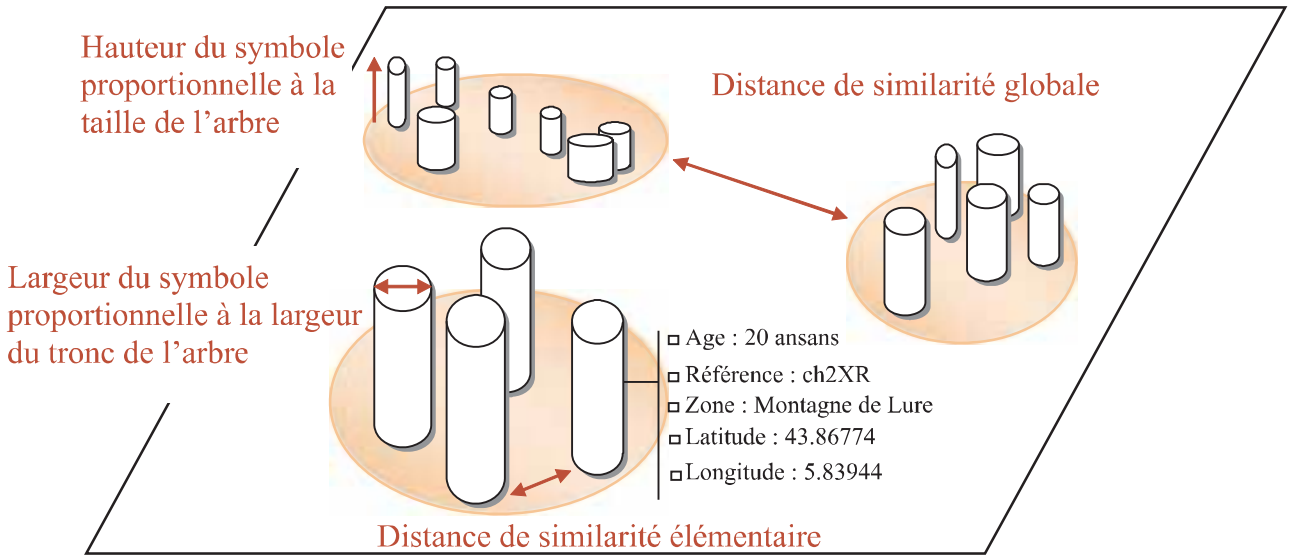

Fig. 5. Une possibilité de visualisation de clusters. Exemple sur des arbres que l'on a classifié selon leur taille, la largeur de leur tronc et leur positionnement géographique

géographique construite par projection de coordonnées, et une telle carte de similarité.

Le premier lien possible consiste en la réalisation d'une carte de similarité numérique en trois dimensions, où la troisième dimension permettrait l'affichage d'attributs nous permettant de faire le lien avec les cartes en projection géographique. Ces attributs seraient des toponymes, des coordonnées géographiques, les langues parlées dans les zones concernées, par exemple, l'idée étant pour le lecteur de pouvoir se repérer entre les différents modes de représentation.

CONCLUSION – L'APPLICATION À DES ZONES D'INTERFAÇAGE DE DONNÉES GÉOPHYSIQUES

L'approche LAC permet aujourd'hui d'harmoniser, réconcilier des bases de données géophysiques, et rattacher des informations les unes aux autres, dans le but d'un meilleur accès à la donnée, et à un gain de place de stockage de l'information. En l'appliquant sur des données géographiques, océanographiques et géophysiques, les croisements et classifications que l'on obtient peuvent permettre d'étudier l'évolution de limites et frontières naturelles entre différents écosystèmes, de repérer des caractéristiques identitaires prédominantes, des zones frontalières d'interfaçage plus ou moins progressives ou des variations brutales du territoire. Des méthodes de visualisation de ces zones et de ces traitements où la mesure de ressemblance à différentes échelles serait la règle de représentation sur ces "cartes de similarité". Ces méthodes de visualisation sont en cours d'étude car elles demandent la définition de stratégies de projection de la similarité en trois dimensions au maximum. ■

Alba Fuga – Ingénieure en Systèmes d'Information Géographiques, elle met actuellement en production le système expert qu'elle a conçu pour le Data Management chez TOTAL, basé sur la méthodologie LAC. Elle poursuit ses recherches en mesure de ressemblance entre données géo-scientifiques dans le cadre d'un Doctorat à l'Université française Pierre et Marie Curie.