Sommaire

Remerciements 2
Sommaire 3
Introduction 5
1) Un aperçu de l'environnement scientifique 6
1.1) Les avancées technologiques 6
1.2) L'accès et la gestion de l'information 8
1.3) Les différentes bases de données chez TOTAL et les problématiques que leur diversité engendre 10
2) Méthodologie applicative 12
2.1) Le données : des profils de navigation sismique 13
2.2) Modélisation, typologie d'attributs – Phase 1 de l'AMR 16
2.3) Classification d'attributs – Phase 2 de l'AMR 20
2.4) Calcul d'attributs de synthèse – Phase 3 de l'AMR 24
2.5) Traitement segmenté de fichiers 25
3) Mesures de Ressemblance 33
3.1) Le workflow existant d'harmonisation des données 33
3.1.1) Le workflow d'harmonisation existant avant la mise en application de l'AMR 33
3.1.2) Les attributs calculés et les méthodes de calcul utilisées avant l'AMR 34
3.1.3) Les méthodes de croisement via InnerLogix (ILX : Logiciel Schlumberger de contrôle qualité des bases de données) 36
3.1.4) Optimisation du workflow d'harmonisation des bases de données chez TOTAL par l'AMR 38
3.2) Les métriques attributaires de similarité – spécialisation en fonction des critères de comparaison 40
3.2.1) Différentes approches pour mesurer la similarité 44
3.2.2) Précision et exhaustivité 49
3.2.3) Les $(Mesuresk)k\epsilon 17$ *50*
3.3) Etalonnage pour les mesures de similarité textuelle dans l'implémentation de l'AMR pour l'harmonisation des bases de données de TOTAL 58
3.4) Similarité contextuelle 62
3.5) Arbre de filtrage à tamis 64
3.6) Résolution et zone d'interfaçage 68
4) Regroupements 71
4.1) Différents types de classification automatique de l'AMR 74
4.1.1) Trois stratégies de classification – principe de résolution 74
4.1.2) Couples et réconciliation de sources 81
4.1.3) Groupes asymétriques et rattachements 82
4.1.4) Clustering, propagation, harmonisation 83
4.1.5) Les différences entre les trois algorithmes de classification et leur combinaison 86
4.2) Tests : Exemple de résultats obtenus sur le Brésil 90
4.2.1) Etape 1 : Egalité exacte entre les noms de lignes 91
4.2.2) Etape 2 : Egalité exacte entre centroides 91
4.2.3) Etape 3 : Egalité exacte entre les longueurs linéaires 92

4.2.4) Etape 4 : Egalité exacte entre les longueurs linéaires et les centroïdes tolérance 100 ..... *93*
4.2.5) Performances temporelles pour les données Brésil (4411 lignes, et 1381137 SP) 94
5) Dimension système expert – automatisation 95
5.1) Graphe d'appel de LAC et enchaînement des mécanismes de l'AMR 97
5.2) L'apport des mécanismes d'intelligence artificielle au fonctionnement de LAC 101
6) Eléments de visualisations 107
6.1) Visualisation géographique 108
6.2) Visualisation par analyse en composantes principales (ACP) 109
6.2.1) Analyse pour l'ensemble des attributs numériques *110*
6.2.2) Analyse pour l'ensemble des attributs numériques, en retirant les coordonnées cartésiennes *117*
6.3) Visualisation « gravitationnelle » par mesures de similarité 121
6.3.1) Algorithme de visualisation en graphe éclaté – Etape 1 ..... *121*
6.3.2) Algorithme de visualisation en image colorée – Etape 2 ..... *127*
Conclusion 136
Bibliographie 139
Glossaire 143
Article paru dans GES Journal en janvier 2014 (Geography, Environment, Sustainability) 146
Algorithmes 159
Réorganisation 159
Classification - Couples 176
Classification – Groupes Assymétriques 178
Clustering 180
Mode d'emploi LAC_DM (pour le Data Management) 184
Généralités 184
LAC_DM pour les lignes de navigation 2D 189
LAC_DM pour les lignes de navigation 3D 194
LAC_DM pour les puits 195
Mode d'emploi de LAC_DT (pour la Documentation Technique) 196
LAC_DT pour l'harmonisation : comparaison de fonds Siège-Filiale, harmonisation de fonds 196
LAC_DT pour le géo-référencement : comparaison d'une liste de mots IHS avec une liste de mots e-Search 198
LAC_DT pour le rattachement : rattachement de noms de puits à des documents techniques 200
Mode d'emploi de LAC_DP (pour la Données Puits) 202
LAC_DT pour la comparaison entre la base LogDB et un disque de stockage 202
LAC_DP pour la recherche de doublons dans LogDB 205
Table des illustrations 208
Table des tableaux 210

Introduction

Les besoins d'harmonisation et de comparaison des données ont trois sources majeures. La première est la nécessité de disposer d'une base de données de référence dans laquelle on stocke l'ensemble de l'information que l'on possède sous sa meilleure version, la plus précise et la plus complète. A chaque fois que de nouvelles données sont reçues, il faut les comparer aux anciennes pour savoir si on les enregistre dans la base de référence ou bien, en cas de conflits, quelles données on garde entre celles déjà en base de référence et celles entrantes.

La deuxième source du besoin d'harmonisation concerne les bases sous forme de projets-études. Les demandes d'harmonisation provenant notamment de filiales sont souvent urgentes. On reçoit des projets où l'information a été enregistrée sans tri. Il faut analyser chaque projet pour retirer les doublons ou les données trop biaisées, pour ensuite fusionner différents projets.

Enfin, l'achat de données est une source indirecte alimentant les besoins en harmonisation. Avoir des données triées et harmonisées permet de ne pas les acheter deux fois, ou dans une version moins précise ou moins exhaustive.

Le gros volume de données à harmoniser par rapport à la demande, ainsi que l'urgence de certaines requêtes provenant de filiales créent le besoin d'automatiser, d'optimiser et d'accélérer le processus d'harmonisation dans le Data Management.

Ce mémoire a pour but de définir et expliquer quelle est la démarche proposée pour mettre en place une solution au problème d'harmonisation des données industrielles.

1) Un aperçu de l'environnement scientifique

Dans ce premier chapitre, on souhaite donner un cadre scientifique permettant de comprendre ce qui a fait émerger les besoins en harmonisation des données.

1.1) Les avancées technologiques

Pour apporter une vision épistémologique de la problématique d'harmonisation des données, on peut rappeler que parfois les besoins théoriques sont amorcés par ces expériences, des choses « qui marchent », qui produisent de la valeur sans avoir été réfléchies ou structurées au départ. D'autres fois, la théorie mathématique, économique, géologique, sociologique etc., sera formulée avant les applications pratiques.

L'amélioration des performances : hardware et high performance computing (HPC)

La gestion de l'information est un exemple parlant de cette course fluctuante entre théorie et pratique car, comme le disent les ingénieurs en HPC, lorsque les limites physiques des ordinateurs sont atteintes, c'est l'algorithmique qui progresse pour résoudre les blocages. Et lorsque l'algorithmique stagne, on cherche des solutions architecturales voire physiques pour agglomérer des machines, augmenter les performances des processeurs, ou inventer des réseaux de stockage de l'information. C'est ce qui est arrivé depuis les années soixante. En particulier depuis les années 2000, des clusters de machines ont commencé à être utilisés, au début il s'agissait de clusters IBM à 8 nœuds, donc quatre de 2Go, et deux de 1Go de mémoire, et 36Go sur disques. Deux nœuds seulement étaient multi-cœurs : l'un de 8 cœurs, et l'autre de 4 cœurs.

En 2012, on a fini par grandement multiplier les capacités de calcul en passant à 594 nœuds de calcul chez Airain Bull, avec 64Go par nœud, et 1204 processeurs, et un stockage sur disque par fichiers global de 2Po. Cela illustre l'immense avancée dans les technologies de stockage et de traitement de l'information.

L'utilisation de processeurs multi-cœurs, ainsi que l'utilisation des clusters a également permis le développement des techniques de calcul vectoriel et de parallélisation des algorithmes, permettant de faire plus de calculs, plus vite, mais aussi d'aboutir à des résultats plus précis.

C'est dans cette atmosphère technologique que les groupes comme Google ont décidé de changer le modèle architectural logiciel en distribuant les programmes aux données au lieu d'envoyer les données aux programmes. Pour cela, ces derniers doivent être adaptés à une démarche en deux temps itérable : une phase de classification/mapping des informations à traiter puis une phase de factorisation/reducing de celles-ci. Notons que l'on retrouve des principes de segmentation des programmes et des données dans ces techniques HPC.

Des modifications dans les techniques d'acquisition des données sismiques

L'informatique n'est pas la seule à avoir évolué vers de plus gros volumes de données : les équipements d'acquisition sismique comme les géophones ainsi que les différentes manières de les déployer sur le terrain évoluent également, et permettent aux différentes compagnies de passer des commandes pour des campagnes de prospection gigantesques mettant au défi aussi bien les modules de traitement que la logistique. On peut passer aujourd'hui d'une acquisition de 0.4 M traces /km^2 à 18 M traces /km^2, soit obtenir une densité 45 fois plus grande en traces sismiques.

1.2) L’accès et la gestion de l’information

Les travaux sur l’accès à l’information viennent en réponse à des besoins provenant de réglementations, de la gestion des territoires et du développement des systèmes d’information géographiques (SIG).

L’accès à la donnée

En effet, en mettant à dispositions des outils comme le GeoPortail ou Google Maps, la cartographie numérique est devenue un moyen de diffuser l’information, en plus de la stocker et d’y faire des traitements. Dans le monde industriel, les SIG se sont aussi multipliés, alliant bases de données propriétaires et visualisation sur mesure pour des problématiques spécifiques au métier, comme pour l’acquisition de données sismiques. On verra par la suite qu’une partie de l’harmonisation des lignes de navigation sismiques était réalisée sous le SIG ArcGIS Cependant l’intérêt du SIG a évolué au-delà du caractère géographique de la donnée. On cherche aujourd’hui à en faire un point d’entrée sécurisé de bases très volumineuses et dont les coordonnées géographiques ne sont pas l’attribut prépondérant, mais plutôt une sorte d’index de stockage. On envisage donc une organisation des bases de données métier fondées sur une segmentation géographique. Or, en les rangeant selon des catégories spatiales, d’une part, on peut avoir des difficultés à comparer ou associer des données de zones géographiques différentes. D’autre part, comme nous le verrons dans le dernier chapitre, ce filtre géographique, tout en étant aujourd’hui indispensable, ne permet pas un stockage harmonisé de l’information d’une donnée complète d’un point de vue métier.

La structuration et les systèmes experts

La question d’accès à la donnée est l’un des sujets moteurs du Data Management : elle évoque aussi bien la rapidité d’accès que les droits de la lire ou de la modifier, relatifs à des notions de degrés de confidentialité. Il s’agit aussi de gérer au mieux le rapport entre la liberté d’accéder ou traiter les données et la garantie de la qualité de celles-ci. Pour ces raisons, des architectures du type Master Data Management ont été conçues. Dans ces architectures, il s’agit de centraliser les données les moins variables du système d’information dans une base qui sert

de noyau, ensuite, il est nécessaire d'identifier les meilleures versions de chaque donnée pour garantir la qualité de celles-ci. Généralement on appelle ces éléments des Golden Records, générés par des règles assez simples de sélection de l'un ou de l'autre des champs attributaires candidats à être le meilleur représentant. Cependant la définition de ces représentants n'est pas encore bien éprouvée et reste relativement empirique, même si on commence à utiliser d'efficaces métriques de similarité comme celles présentées dans le chapitre 3. Dans les chapitres 4 et 5 de ce mémoire, on proposera une nouvelle manière d'obtenir des représentants abstraits ou réels du groupe. Notons que la qualité des données est d'autant plus importante que des normes réglementées contrôlent certains organismes où la donnée a de fortes implications légales ou économiques par exemple.

Dans les architectures de Master Data Management, différentes stratégies sont proposées en fonction des modes d'utilisation des données. Pour choisir la bonne approche, il est indispensable de réaliser un modèle complet des données et des flux d'utilisation, d'où l'importance de modéliser les données, notamment avec des structures d'hyper-classes comme en HBDS, reflétant la factorisation des attributs prépondérants sur un ensemble d'autres classes et pouvant être centralisés.

La structuration des données pour des Master Data Management, outre l'aspect d'accès et de qualité de la donnée, a pour objectif d'améliorer et accélérer la prise de décision. C'est en cela que l'on s'approche de la notion de système expert. En effet, Matthieu Beard et d'autres experts des systèmes d'information définissent un système expert comme un logiciel utilisant des connaissances, des faits et des techniques d'inférence pour résoudre des problèmes ou prendre des décisions. « An expert system is a computer program that uses knowledge, facts, and reasoning techniques to solve problems or aid in making decisions. » M. Beard, *Experts Systems, An Introduction*, 2016.

Entre la structuration des bases de données et le besoin d'optimiser les prises de décisions, il semble bien qu'il soit aujourd'hui envisageable de passer de la notion de base de données à la notion de base de connaissances dans un cadre applicatif.

1.3) Les différentes bases de données chez TOTAL et les problématiques que leur diversité engendre

Les données sont stockées sous différents formats et sous différents systèmes de gestion de bases de données, souvent mis en concurrence, mais intégrés dans une organisation s'orientant tout de même vers un Master Data Management, notamment avec la maintenance d'une base centrale de données de référence. Toutefois, l'intégration des flux entrants à cette base de référence Master DB n'est pas évidente.

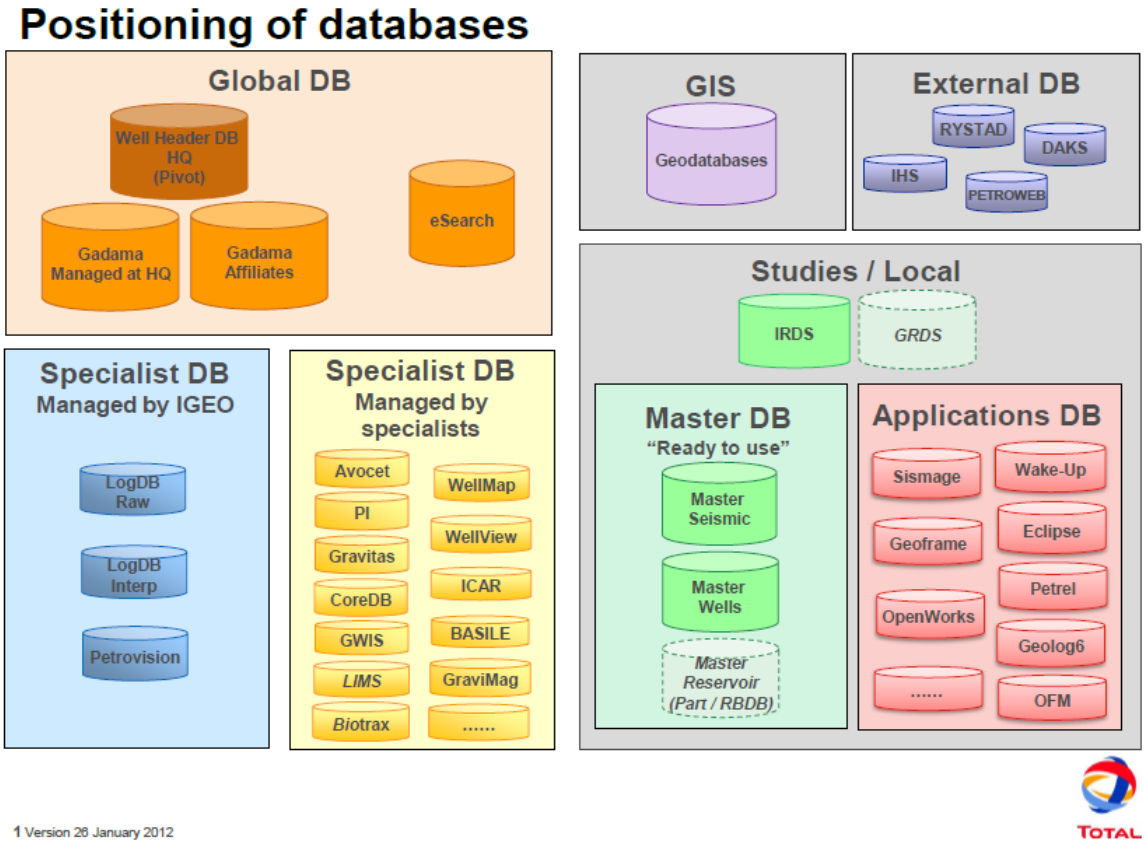


Figure 1 : Illustration des bases de données chez TOTAL. Source : Intranet TOTAL, en 2013

Une partie des données est stockée sous Sismage, le logiciel d'interprétation géologie/géophysique interne développé par TOTAL, et n'acceptant pas de connecteurs à sa

bases de données pour des raisons de sécurité. On y trouve la majorité des projets de navigation sismique. Cependant, d'autres nombreux projets sont stockés sous d'autre bases comme celles de GeoFrame_et d'OpenWorks.

Les données de la base Patrimoniale du Groupe TOTAL, ou données de référence sont stockées sous une base Oracle nommée FINDER pour ce qui est de la partie de navigation sismique (partie « en surface » de la donnée, contenant l'information de positionnement en surface), et les données qui leurs sont liées du point de vue du signal sismique et des traces sismiques (partie « en profondeur » de la donnée) sont stockées dans une base PétroVision.
La migration imminente de Finder dans PetroVision soulève le besoin de réconciliation entre ces deux bases. Il s'agit d'une problématique très proche de celle d'harmonisation. En somme, le but est de trouver quelle ligne de navigation sismique de Finder correspond à quelles traces de PetroVision.

Le fait de stocker la même donnée dans différentes bases (par exemple projets d'origine et base de référence), ou bien deux parties de la même information dans deux bases distinctes pose de nombreuses problématiques. Cette diversité des bases de données est historique, mais aussi reflète l'étendue du Groupe TOTAL et de ses filiales.

Les mécanismes que nécessite l'harmonisation peuvent donc aussi faciliter le suivi et la reconnaissance d'une même donnée dans les différentes bases, sachant que jusqu'à aujourd'hui il n'existe pas de système d'identifiant unique pour une donnée sismique. Quant aux identifiants des données puits, ils ne sont pas encore mis en place pour l'ensemble des données. Les techniques de comparaison et de mesure de similarité conçues dans la méthodologie, d'Automatisation de la Mesure de Ressemblance développée dans le prochain chapitre, appliquée aux bases de TOTAL pourront également être utilisées dans un but d'identification de données complexes.

On entend par donnée complexe une donnée difficile à identifier parce qu'elle porte un nombre de caractéristiques élevé, mais avec incertitudes sur leur renseignement.

2) Méthodologie applicative

Ce chapitre introduit une méthodologie d'automatisation des mesures de ressemblance et ses premières phases.

La méthodologie de l'AMR, Automatisation des Mesures de Ressemblance, présente l'intérêt de combiner plusieurs procédés de mesures de similarité et plusieurs techniques de comparaison afin qu'ils se complètent les uns les autres de manière structurée.

Notre approche d'automatisation des mesures de ressemblance est fondée sur l'articulation entre un système de filtrage à tamis, et des mécanismes de classification automatique.
Pour mettre en place le traitement automatique, il est nécessaire de suivre un protocole préalable de modélisation et de hiérarchisation.

Présentons en premier lieu le type de données à comparer ainsi que le workflow de traitement en place avant l'application de l'AMR.

2.1) Le données : des profils de navigation sismique

Un profil sismique est le résultat d'une campagne d'acquisition sismique soit terrestre, soit marine.

Dans les deux cas, le principe est de générer des microséismes par explosifs ou bien lâché de masses dans l'environnement étudié, et d'acquérir les réponses sismiques de cet environnement grâce à la disposition de géophones autour des sources de microséismes.
On appelle point de tir, ou shot point (SP), les points dans l'espace, de coordonnées mesurées par GPS, où se situe l'émetteur du microséisme.

On appelle point de profondeur commune, ou common depth point (CDP) le point de réflexion de l'onde sismique au niveau d'une couche géologique en correspondance à un SP donné. Un SP quant à lui peut correspondre à plusieurs CDP. Le CDP est donc un attribut caractérisant à la fois les conditions d'acquisition sismique et la composition de l'environnement physique de la campagne d'acquisition. Normalement, pour une campagne donnée, le rapport SP/CDP est constant.

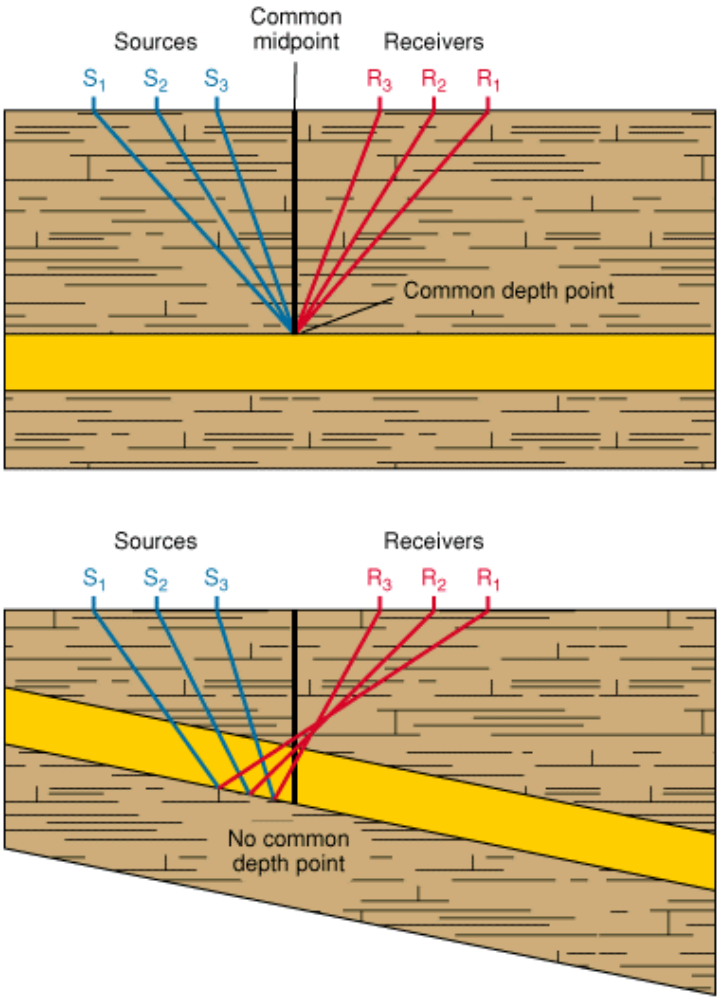


Figure 2 : Schéma de rebondissement de l'onde sismique selon la géométrie du terrain. On y représente les shot point SP, les common depth point CDP, géophones et masses émettrices.
Source : Schéma Schlumberger

Par ailleurs, à chaque couple (SP, CDP) correspond une trace sismique, c'est-à-dire le signal enregistré par le géophone en surface.

Ces traces permettent, après traitement et analyse, de déterminer un modèle du profil sismique 2D, c'est-à-dire la vision d'une coupe géologique du sous-sol de la zone géographique sujette à la campagne.

Dans les données Patrimoniales, on distingue donc les topographies sismiques (stockées dans FINDER pour la base de données de référence) que sont les coordonnées, les noms de lignes et les SP constituant une ligne sismique 2D, ainsi que leurs CDP correspondants aux éléments de sismique que sont les traces, et profils sismiques (stockés dans PetroVision

pour la base de données de référence). Il faut prêter fortement attention aux systèmes de référence géodésiques, ou aux systèmes de projection cartographiques des données, que ce soit pour les exports, les comparaisons ou la visualisation. En effet, des données ne peuvent être analysées l'une par rapport à l'autre que si elles sont exprimées dans le même système de coordonnées.

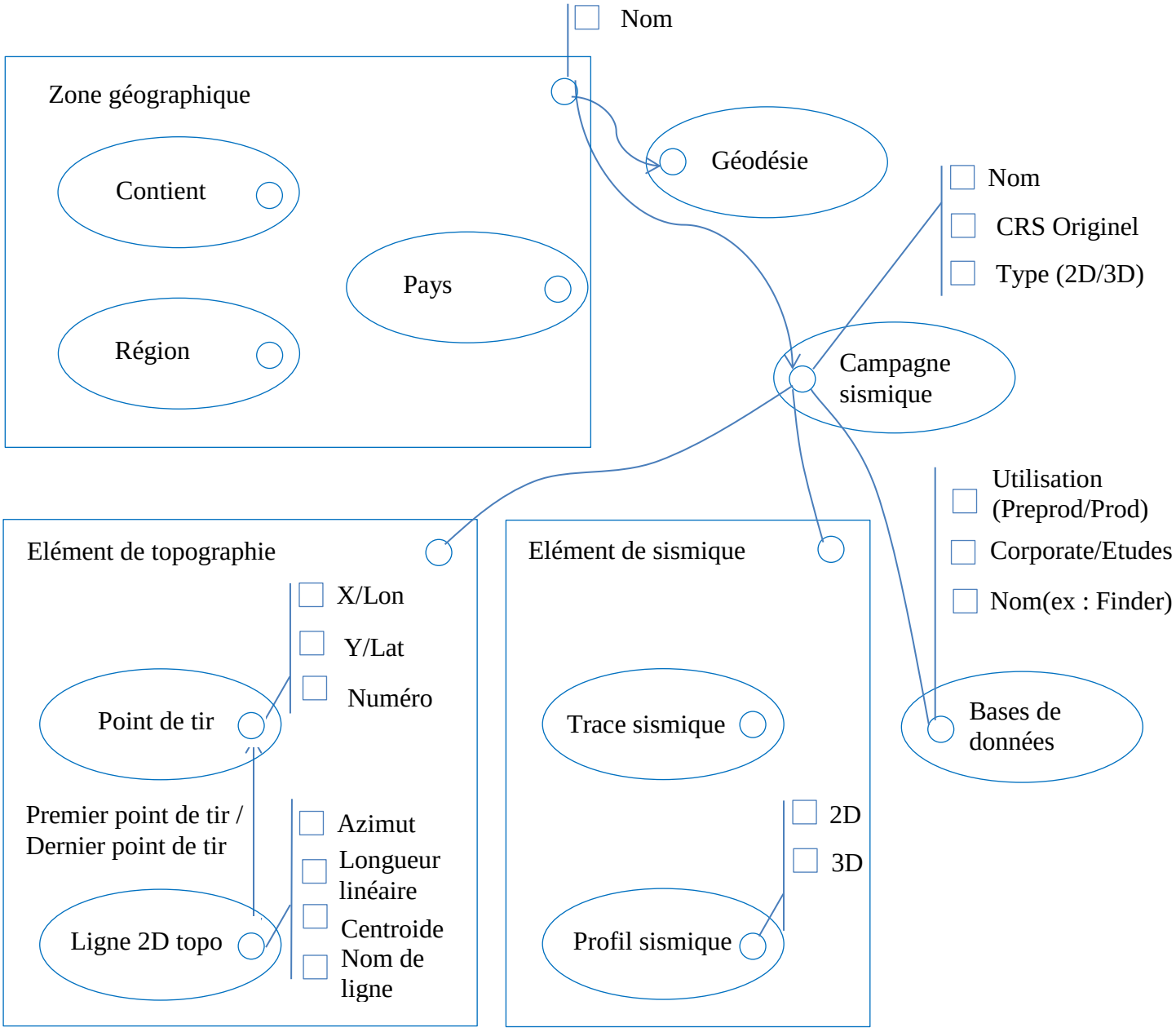


Figure 3 : HBDS des composantes géographiques et géophysiques d'une acquisition sismique et de son stockage

2.2) Modélisation, typologie d'attributs – Phase 1 de l'AMR

La modélisation est possible grâce à l'enregistrement des différentes manifestations d'un phénomène que l'on souhaite analyser, par exemple afin d'explorer le sous-sol d'un territoire. Certaines des caractéristiques de modélisation constituent des critères pertinents de comparaison entre les différentes manifestations. Elles serviraient par exemple à reconnaitre un phénomène à partir de différents enregistrements.

En prenant pour base la méthodologie de structuration HBDS, et en la déclinant dans le thème de la mesure de ressemblance, on considère qu'un phénomène est modélisable comme un ensemble d'attributs relevant de quatre types distincts :

- Les attributs élémentaires bruts qui sont des enregistrements directs d'informations caractéristiques du phénomène. Exemple : enregistrement du signal sismique.

- Les attributs élémentaires traités issus de traitements géophysiques appliqués aux attributs élémentaires bruts. Exemple : l'amplitude du signal sismique calculée à partir du signal brut. Ou un échantillonnage des traces sismiques.

- Les attributs-métadonnées. Exemple : pays d'acquisition de la donnée, date, compagnie d'acquisition.

- Les attributs combinés qui ont été formés à partir de combinaisons d'attributs élémentaires bruts et/ou traités. Un attribut combiné est donc un ensemble d'attributs élémentaires muni d'une règle de combinaison. Exemple : le barycentre d'une ligne de navigation calculé à partir des coordonnées des points constituant cette ligne.

Dans le domaine de l'exploration pétrolière, le caractère brut ou traité des attributs dépend souvent des différentes sources de provenance de la donnée. Par exemple, cela dépendra du fournisseur de la donnée, ou du fait de travailler avec des données nouvelles à charger en base ou des données déjà traitées et retraitées au gré des besoins du binôme interprétation-décision. Plus concrètement, une trace sismique sera considérée comme attribut brut si aucun algorithme de traitement du signal ne lui a été appliqué après enregistrement. Elle sera considérée comme attribut traité dans le cas contraire.
Cette distinction peut avoir une importance non négligeable pour introduire un degré de fiabilité de l'attribut lors de l'analyse des données.

Nous venons de décrire une première distinction typologique des attributs de modélisation d'un phénomène à des fins comparatives.
D'un point de vue structurel, la modélisation d'un phénomène peut être complétée par des liens rattachant ce phénomène à d'autres phénomènes. Par besoin de comparer des manifestations de phénomènes et non des interactions entre phénomènes géophysiques, on considérera que toute relation d'un phénomène à d'autres sera représentée sous forme d'attribut et non de lien. Par exemple, pour un lien de filiation entre classes, on considérera que la classe mère portera un attribut donnant la liste de ses classes filles. On pourra donc tenir compte d'attributs à caractère relationnel et d'attributs à caractère non relationnel. Exemple : si on modélise un puits de forage, la profondeur d'un puits est non relationnelle. La distance moyenne d'un puits aux puits voisins est un attribut relationnel.

Une seconde typologie d'attributs consiste à distinguer les attributs qui seront considérés comme critères de comparaison, donc comparables, des autres attributs donnant des informations sur le phénomène mais non utilisables dans les mécanismes de comparaison, donc non comparables.

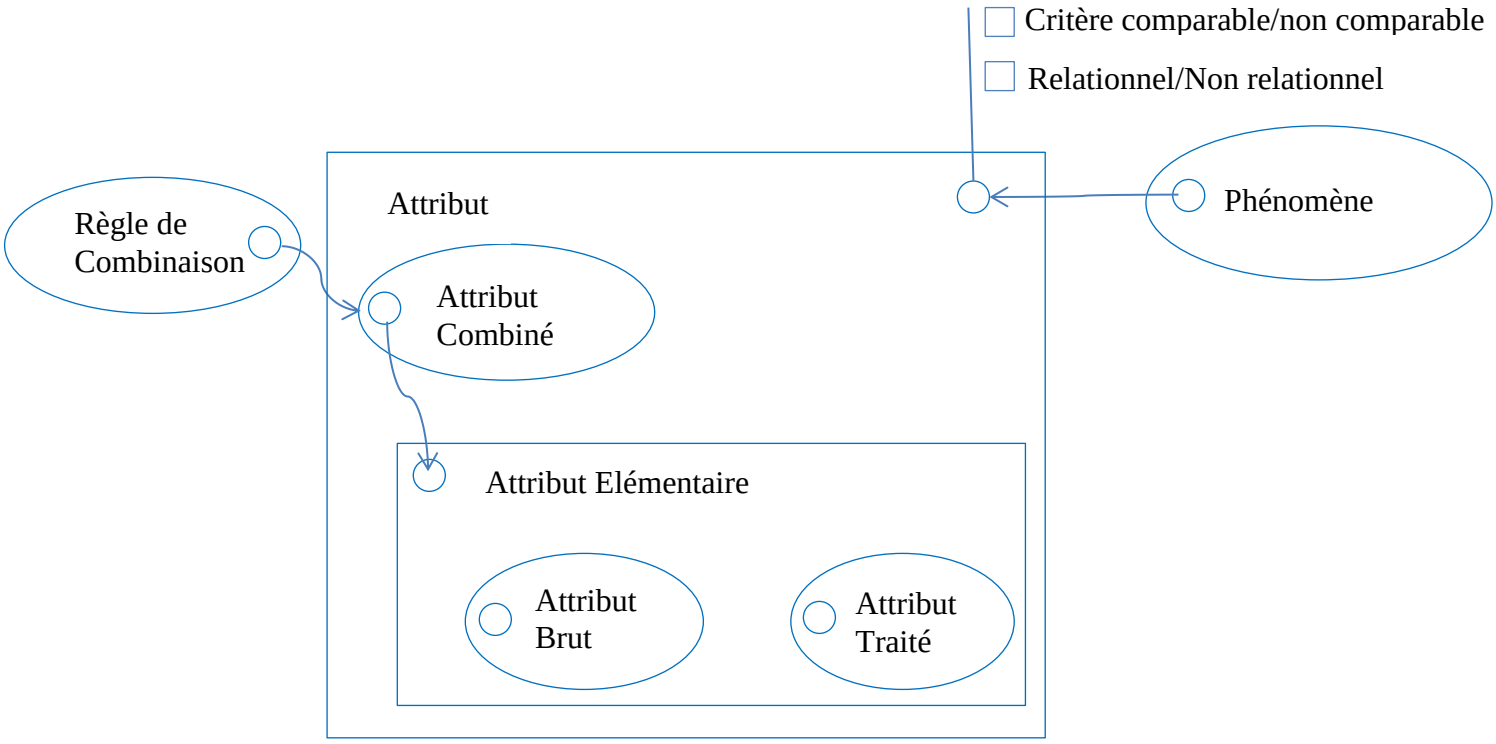


Figure 4 : Modèle HBDS de la typologie attributaire de l'AMR

Ainsi, par « donnée », entend-on l'ensemble des réalisations d'attributs composant le modèle d'un phénomène physique, ou une configuration physique.

La première phase de l'AMR consiste à constituer la liste des attributs qui décrivent le phénomène à modéliser, en identifiant à quel type appartient chaque attribut

L’application de la phase 1 de l’AMR aux lignes de navigation sismiques serait alors la suivante :

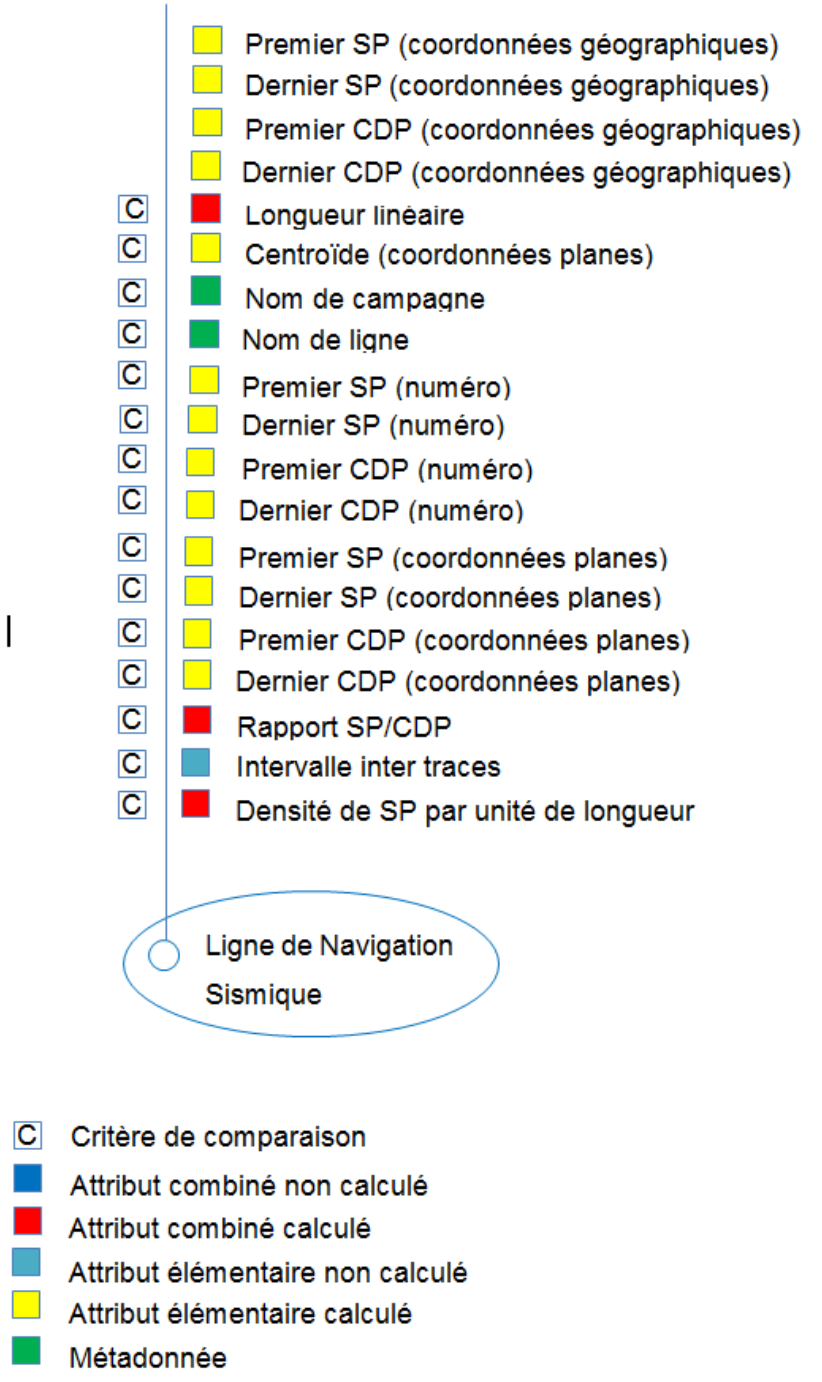


Figure 5 : Composition attributaire d’une ligne de navigation sismique

2.3) Classification d'attributs – Phase 2 de l'AMR

Dans la deuxième phase de la méthodologie, les critères de comparaison de la modélisation de la première phase de l'AMR sont eux-mêmes classés par tamis, c'est-à-dire par groupe d'attributs. Les critères de cette classification sont la fiabilité mathématique, la pertinence de comparaison (ou degré de caractérisation de la donnée), et la robustesse du facteur de tolérance.

La fiabilité mathématique concerne la formule mathématique de calcul de l'attribut, s'il est calculé, 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 somme 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 performance de calcul informatique lors des comparaisons et calculs d'attributs.

Le critère de pertinence de comparaison vise à ordonner les attributs selon leur capacité à caractériser le phénomène modélisé. 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.

Les facteurs de tolérance sont ces seuils de résolution nous permettant de savoir à partir de quelle proximité les objets sont suffisamment similaires. La robustesse des facteurs de tolérance peut dépendre de la zone géographique où est enregistrée 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 le pays dans le paramétrage du seuil de tolérance pour la comparaison des données.

Pour revenir à l'exemple de modélisation d'un cours d'eau, le débit change en fonction de la saison, et du 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. En effet, sur deux fleuves, le nombre de barrages est un critère de comparaison plus stable que le débit, donc le seuil de tolérance pour la comparaison peut être défini plus facilement et avec plus de robustesse : si un seuil est robuste, ce même seuil appliqué à des données différentes correspondra alors à une même exigence sur les comparaisons.

On peut ajouter qu'une bonne harmonisation des données nécessite de s'adapter au contexte englobant ces données, notamment en adaptant les seuils de tolérance s'ils n'ont pas ce caractère de robustesse.

On nomme tamis un groupe d'attributs présentant les mêmes degrés de fiabilité, pertinence et robustesse, ou bien un groupe d'attributs décrivant un aspect spécifique de la donnée. Une fois les attributs classés dans des tamis, et les tamis eux-mêmes hiérarchisés, on peut aborder la question des métriques de similarité élémentaires et globale, thème traité au prochain chapitre.

Pour la modélisation des lignes de navigation sismique, la hiérarchie des attributs de comparaison a été réalisée après analyse des données, et après consultation d'Eric FAGOT enseignant-chercheur à l'IFP. Ci-dessous se trouve le tableau de synthèse qui illustre l'ordre choisi.

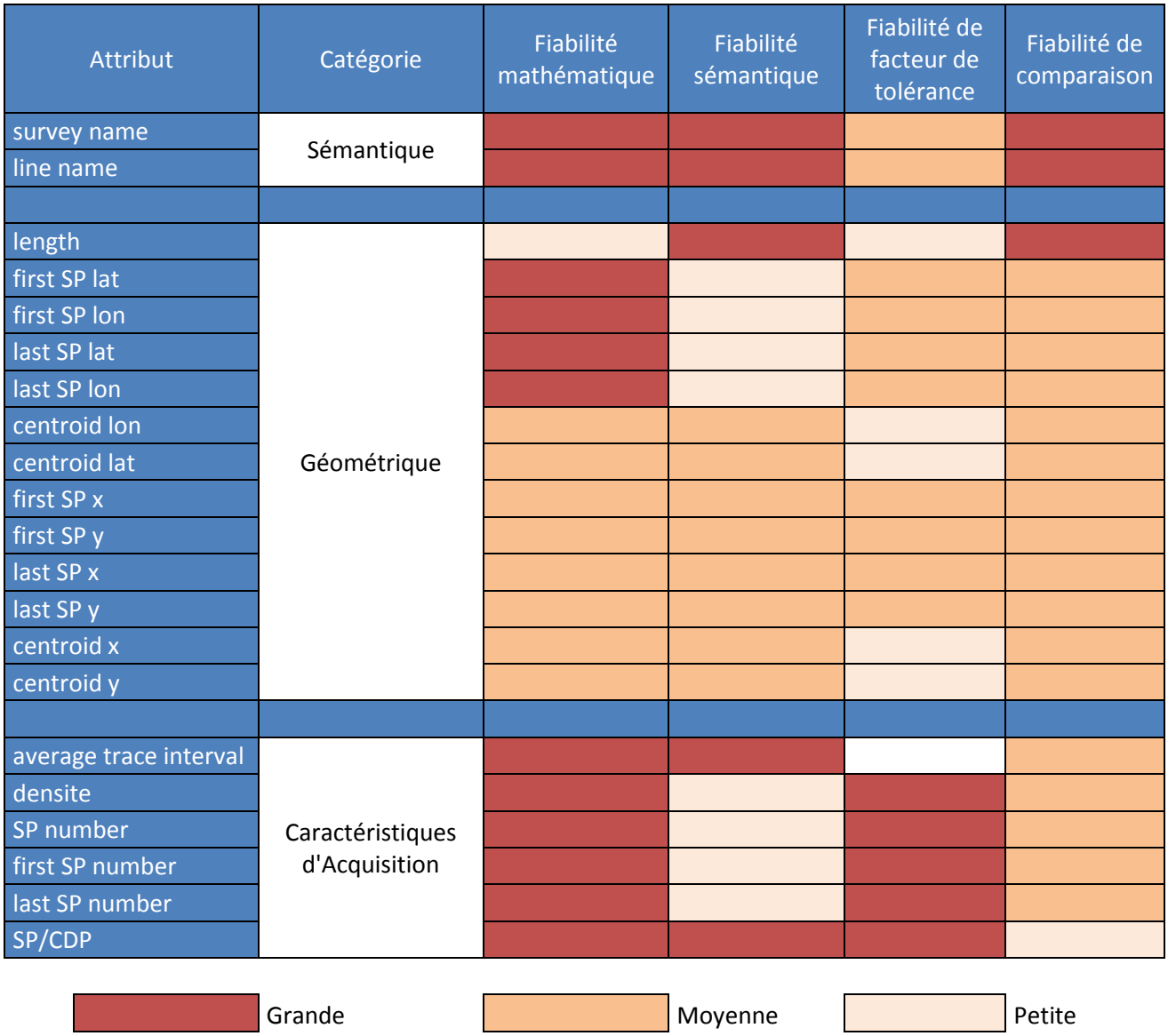


Figure 6 : Classification des critères de comparaison des lignes de navigation sismique, les catégories sont rangées par fiabilité décroissante pour la comparaison

Cette classification des attributs ne suit pas uniquement les critères de fiabilité, pertinence, robustesse, mais aussi le thème descriptif que l'attribut renseigne sur la donnée, le point de vue qu'il reflète. Par exemple pour la ligne de navigation sismique, nous distinguons le tamis sémantique du tamis géométrique et du tamis d'acquisition. Le tamis sémantique concerne des métadonnées de la ligne comme le nom de campagne ou le nom de ligne. Le tamis géométrique contient des informations relatives au positionnement ou à la géométrie de la ligne, et le tamis d'acquisition comporte des informations portant sur les choix d'acquisition comme la distance inter-traces, ou comme le type de prétraitement géophysique des données.

Chaque attribut peut être perçu comme une dimension de modélisation de la donnée ou du phénomène. Mais chaque tamis peut lui aussi être considéré comme une dimension, à une autre échelle.

2.4) Calcul d'attributs de synthèse – Phase 3 de l'AMR

La troisième phase de l'AMR consiste à définir les règles de calcul des attributs combinés.

Dans les bases de données industrielles de géosciences, on peut accéder à des données comme les numéros et coordonnées des SP, et CDP associés, et parfois à l'intervalle inter-traces qui est la distance entre deux sources émettrices du signal sismique.
Ces données peuvent être utilisées comme critères de comparaison, à condition de ne pas produire ou utiliser une information trop dense, trop nombreuse et chaotique car on risquerait d'augmenter la confusion au lieu d'améliorer l'efficacité et la qualité des comparaisons.

Plus il y a d'information et plus on dispose d'éléments d'analyse, mais plus le système est complexe aussi. Pour cette raison, il faut d'une part structurer l'information, d'autre part il faut la synthétiser pour la comprendre, l'interpréter, et l'utiliser de la manière la plus efficace.

C'est pourquoi on choisit de calculer les centroïdes et les longueurs linéaires comme attributs de synthèse représentant l'ensemble des SP dans notre cas d'application de l'AMR à la navigation sismique. Ces attributs représentent de manière schématique la forme de la ligne de navigation sismique. Il est aussi nécessaire de trouver un équilibre entre la complexité des formules de calcul et le temps de traitement lorsqu'on se trouve confronté à des données très volumineuses. Dans l'AMR on tient compte d'un grand nombre de dimensions hiérarchisées menant à des capacités méthodologiques de discernement, permettant d'utiliser des formules plus simples que le kriegeage par exemple, qui demanderait un long temps de calcul. Le caractère multidimensionnel va compenser les méthodes de calcul simples.

Le centroïde, la longueur linéaire, la densité par unité de longueur, premier SP, dernier SP, premier CDP, dernier CDP, rapport SP/CDP sont donc des attributs calculés de la ligne de navigation sismique. Ils sont calculés à partir des coordonnées de SP des bases de données. Les autres attributs sont directement extraits des bases. Le calcul des premiers et derniers SP et CDP se fait simplement par recherche (segmentée) des maximum et minimum des numéros de SP dans une ligne de navigation. Le premier CDP correspond au premier SP, idem pour les derniers.

Par exemple, la densité de SP par unité de longueur consiste à diviser la longueur linéaire calculée de la ligne de navigation sismique par le nombre de SP qui la composent.

Cette phase de spécification des formules de calcul doit être associée au choix des métriques de similarité attributaires que l'on exposera au chapitre suivant.

2.5) Traitement segmenté de fichiers

Nous donnerons deux sens au terme « réorganisation ». Au niveau de la forme, il s'agit du reformatage du fichier en entrée, et d'une restructuration de l'information. En effet, les formats en entrée peuvent provenir de différents systèmes de gestion de bases de données, donc peuvent ne pas porter la même information, et même s'ils portent la même information, elle n'est pas rangée à la même place. Il faut donc déplacer, échanger des colonnes, par exemple, afin d'obtenir en sortie toujours le même format. Il s'agit d'un traitement essentiel d'optimisation, en amont, pour les traitements de classification réalisés en aval.
De plus, c'est à ce niveau que sont gérées les informations non renseignées, ou partiellement renseignées. Leur gestion nous permettra de réaliser des traitements statistiques et de gestion des données lacunaires. Si l'on considère les attributs disponibles dans chaque base de données comme un ensemble d'éléments, alors le format final du fichier retraité est l'union de tous ces ensembles. Tout attribut possède sa place, qu'il soit renseigné ou non. S'il n'est pas renseigné, on trouve dans la case qui lui correspond le mot-clé « NON_RENSEIGNE ».

En Data Management, la question de la standardisation des formats de fichiers d'échange et de stockage de données est délicate car même pour des fichiers standards comme le UKOOA, selon l'évolution des données et des technologies, il y a un besoin de révision, donc de modification du fichier standard. Or, ces mises à jour génèrent elles-mêmes une hétérogénéité de formats dans les domaines industriels où l'historique est exploité.

Au niveau du contenu, la réorganisation est un processus de traitement de l'information car on y calcule les attributs non présents initialement dans la base de données (attributs élémentaires combinés).

De plus le fichier en entrée est un fichier de points de tirs, c'est-à-dire que l'on y trouve, pour chaque ligne de navigation sismique, tous ses SP et CDP associés, avec leurs coordonnées géographiques, cartographiques, et numéros. Une ligne de navigation sismique est donc étendue sur de nombreuses lignes du fichier d'entrée.

Or, nous souhaitons comparer des lignes sismiques et non des points, il est donc nécessaire de factoriser ces points en lignes pour générer un fichier de sortie ne contenant que les lignes de navigation sismique décrites de la manière la plus complète possible pour les mesures de ressemblance qui suivront. Chaque ligne de navigation n'occupera donc qu'une seule ligne du fichier de sortie. Ci-dessous sont présentés deux exemples de procédés de réorganisation d'attributs, l'un en prenant comme fichier de départ un export de la base de données de SISMAGE chez TOTAL et l'autre un fichier standard UKOOA.

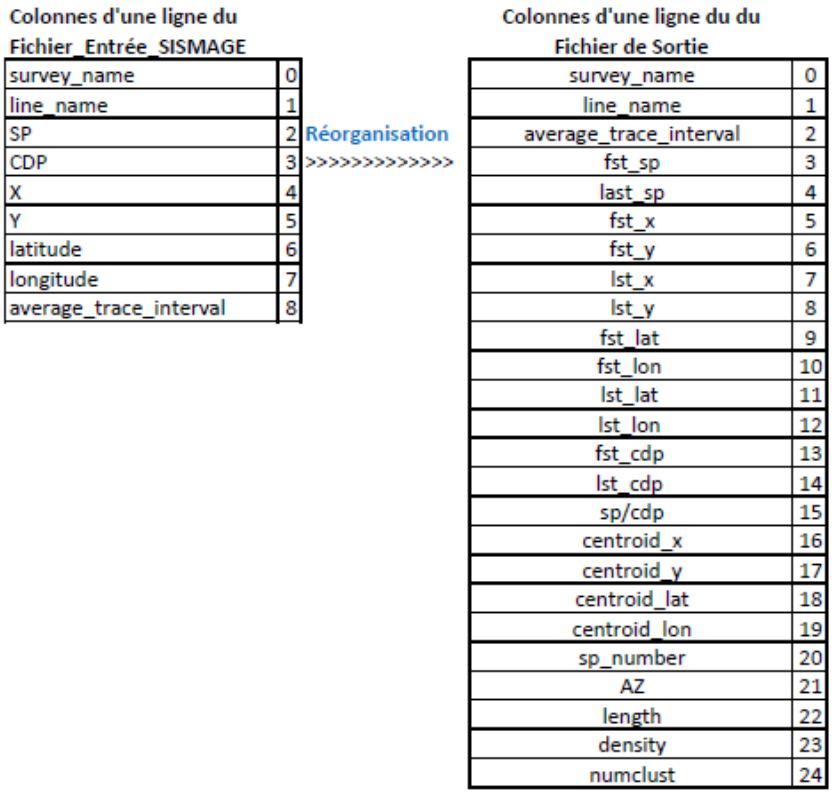


Figure 7 : Schéma représentant le travail de réorganisation des données pour passer de l'ensemble des points sismiques d'un fichier issu de Sismage à la représentation de la ligne de navigation sismique

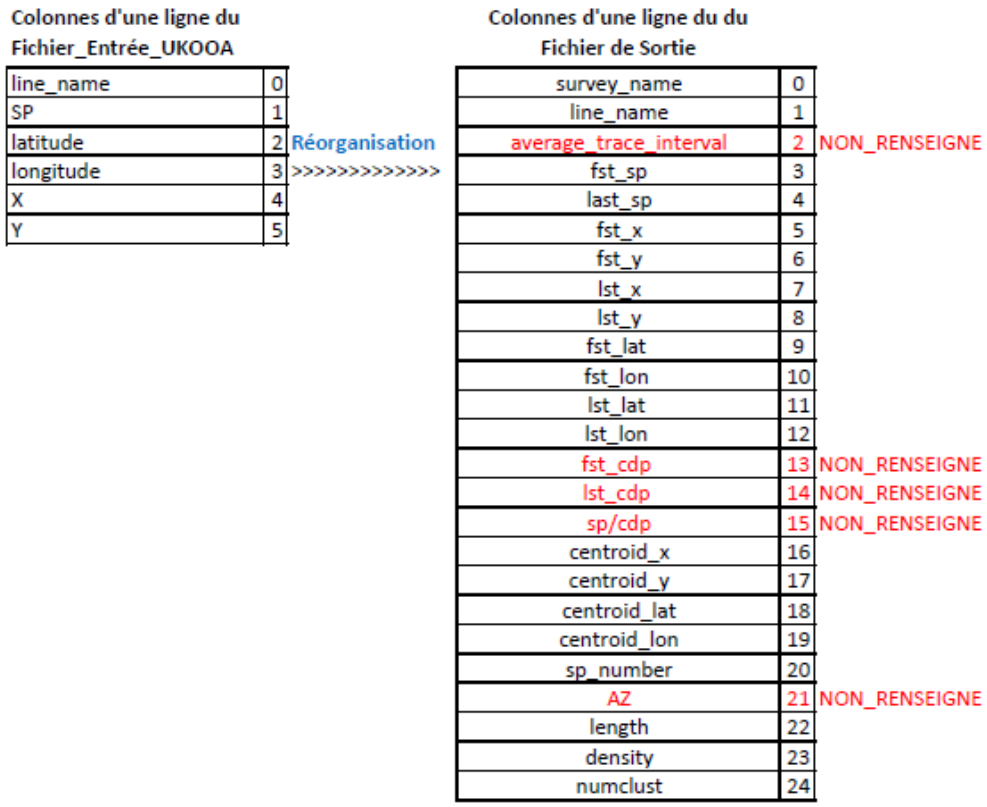


Figure 8 : Schéma représentant le travail de réorganisation des données pour passer de l'ensemble des points sismiques d'un fichier UKOOA à la représentation de la ligne de navigation sismique

Les fichiers en entrée étant souvent de très grande taille, il est impossible de les réorganiser d'une traite. Par conséquent, il a fallu concevoir un algorithme de segmentation permettant de traiter et réorganiser le fichier « par morceaux ». Le facteur limitant est la taille maximale des vecteurs et variables en programmation (quel que soit le langage).

Cet algorithme est basé sur les notions expliquées ci-après.

La segmentation est possible car en entrée aussi bien qu'en sortie de notre programme, se trouvent des fichiers, donc des éléments persistants, contrairement aux variables de programmation, non persistantes. Ces variables nous serviront de buffers.

Dans notre contexte, un buffer se définit comme une variable de programmation dans laquelle on stocke une partie du fichier d'entrée. Cette partie est ce qu'on appelle un segment, c'est-à-dire un nombre de lignes du fichier d'entrée à taille fixe. La taille du segment est déterminée en fonction des capacités d'allocation de mémoire du compilateur utilisé.

Les buffers utilisés pour la réorganisation sont continus, c'est-à-dire que les lignes qu'ils contiennent suivent exactement l'ordre des lignes du fichier, sans exclure de ligne.

Un buffer peut avoir différents types informatiques. On peut utiliser une matrice pour stocker le segment de fichier, avec une ligne du fichier par ligne de matrice et les attributs de la ligne de fichier dans les colonnes de la ligne de la matrice. On peut aussi avoir des buffers scalaires. En effet, on utilisera deux grandes catégories de buffers :

- Les buffers d'entrée pour le fichier d'entrée, dans lesquels on introduit deux sous-catégories :
    - Les buffers de segmentation, classiques sous forme de matrice contenant les segments continus du fichier d'entrée. Ces buffers ont une portée de mémoire d'un segment. A chaque fois que l'on passe au segment suivant, le buffer est effacé puis rempli par le nouveau segment.

    - Des buffers de cumulation dont la mémoire persiste à l'échelle d'une ligne de navigation sismique. Ces buffers ne posent pas de problème de taille mémoire car ce sont des buffers « de calcul », i.e. utilisés pour le calcul des attributs calculés où l'on cumule l'information. Pour prendre un exemple concret, une ligne de navigation sismique peut occuper 1000 lignes de fichier. Or, si la taille du segment est de 500 lignes de fichier, alors le calcul de la longueur ne peut pas se réaliser sur un buffer de segmentation. Il doit se réaliser sur toute la ligne, donc sur les deux buffers qu'elle occupe. Il faut ajouter que dans une campagne sismique, le nombre de SP par lignes de navigation n'est pas constant. Ainsi un buffer de segmentation peut-il contenir plusieurs lignes de navigation, et une ligne de navigation peut occuper

plusieurs buffers de segmentation. Les buffers de cumulation ont donc une persistance plus grande que les buffers de segmentation.

- Les buffers de sortie, où l'on stocke l'information à écrire dans le fichier de sortie, qui est donc écrit de manière segmentée, complété à chaque fois qu'un segment est traité. Ce sont des buffers de segmentation.

De plus, pour assurer une bonne reprise, un bon raccordement entre les segments de fichier, des mécanismes de raccordement sont mis en place. Ils permettent de réaliser un léger recouvrement entre la version ancienne d'un buffer (pour le segment N-1), et sa version actuelle (pour le segment N). Ces clés de segmentation sont des variables dont la persistance est limitée à celle d'un segment, mais elles contiennent une information relative non pas au bloc de fichier traité au moment présent, mais une information relative au bloc passé.

Les cas de figure que doivent gérer les mécanismes de segmentation sont les suivants :

- Un segment peut contenir plusieurs lignes
- Une ligne peut occuper plusieurs segments
- Une ligne peut occuper la fin d'un segment et le début du segment successeur
- Le dernier segment du fichier peut avoir une taille inférieure à la taille fixe conventionnelle d'un segment.

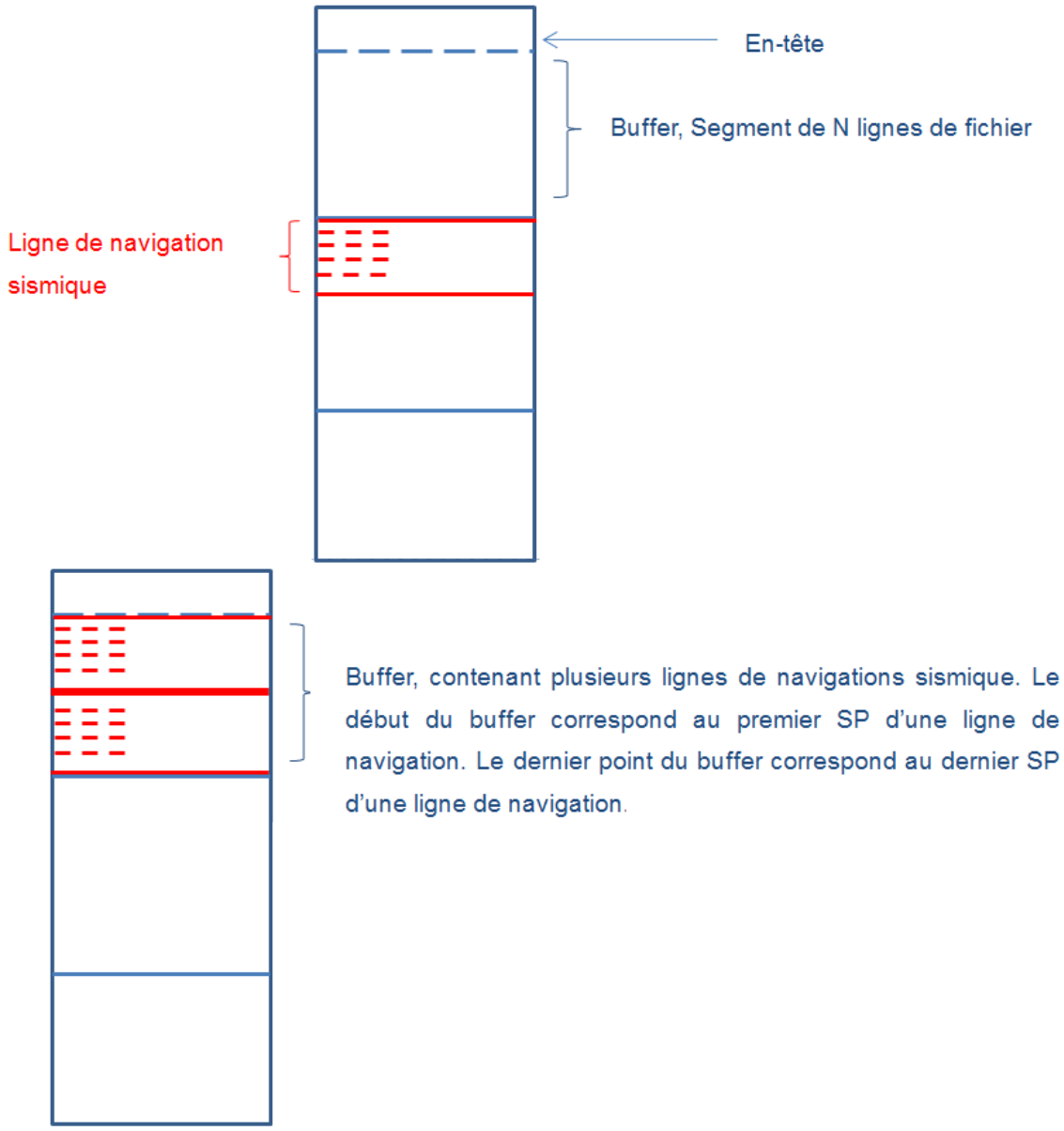


Figure 9 : Exemple de buffer avec lignes de navigation entières

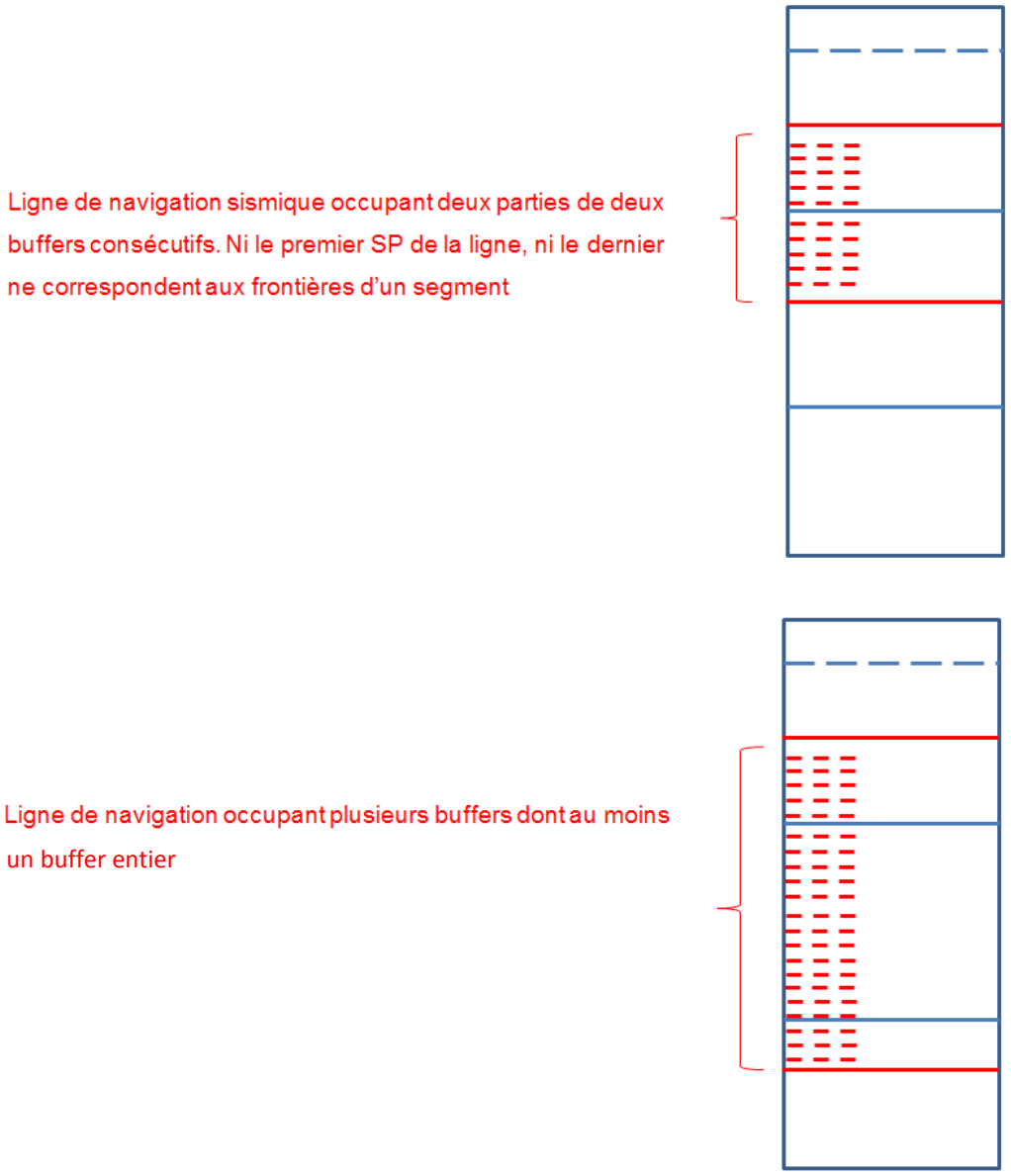


Figure 10 : Exemple de buffer avec lignes de navigation à raccorder entre différents segments de fichier

L'algorithme de factorisation des lignes de navigation sismique, de calcul d'attributs supplémentaires, et de formatage se trouve en annexe.

La méthode mise en place permet de réaliser un traitement segmenté d'un fichier ASCII. Le modèle de fichier utilisé ici est celui d'un export de SISMAGE.

Pour SISMAGE, le fichier d'export doit avoir le format ci-dessous.

```
------Begin Header----------
Export of geometry info for 2D line= LAP-12
Datum: WGS 84
Reference Meridian: Greenwich
Projection System: UTM zone 20N   63W
Unit: m
Average trace interval: 51.04752947062986
------End Header----------
survey_name        line_name        SP      CDP        X        Y        latitude  longitude        average_trace_interval
```

Figure 11 : Exemple de l'en-tête d'un fichier issu de SISMAGE

Ci-dessous se trouve l'interface homme-machine implémentée pour réaliser la réorganisation d'un fichier d'export de données issu d'une base de données industrielle. L'algorithme a été adapté à deux formats de fichier : le standard UKOOA et un format conventionnel de la base de données SISMAGE.

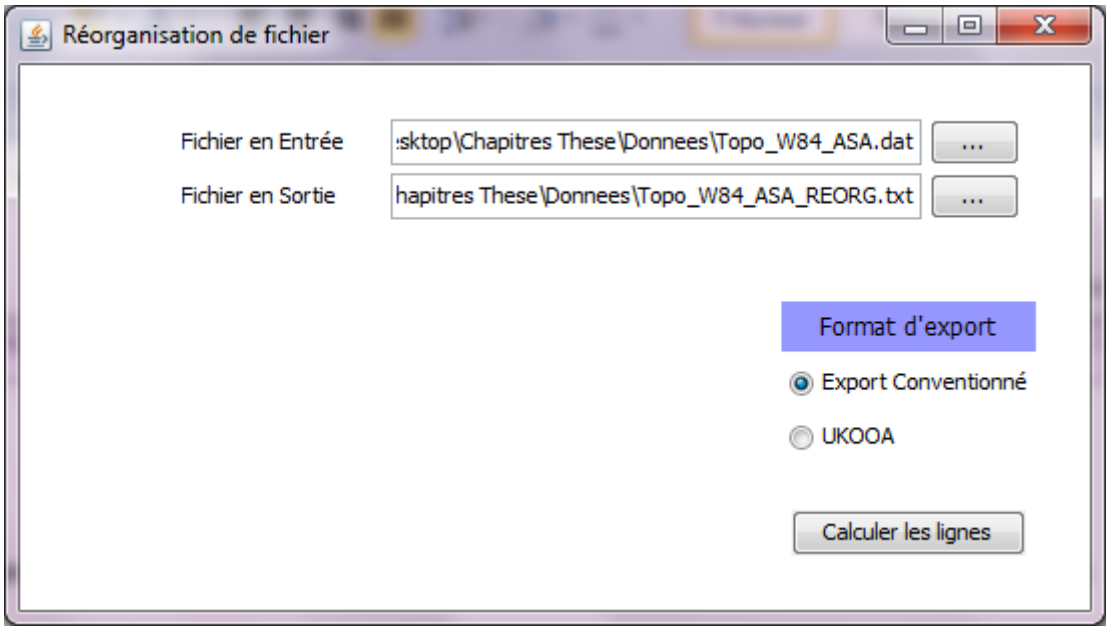


Figure 12 : Interface de réorganisation de fichier

3) Mesures de Ressemblance

Les mesures de similarité ont d'abord été développées pour l'appariement des mots, l'alignement des séquences ADN. Plus tard, elles ont évolué pour la reconnaissance et gestion des doublons dans les bases de données pour la gestion desquelles ont été créées les distances d'édition. Ensuite, viennent les approches plus complexes, mises en place pour comparer et apparier des objets autres que les mots. Des classifications à bases de statistiques apparaissent, puis des considérations déterministes telle la distance de Jaro-Winkler. Les travaux de recherche les plus récents portent sur les appariements d'objets à l'aide d'apprentissage supervisé, à base de statistiques bayésiennes et proposent enfin, comme Kessler, des modélisations pour exploiter le contexte des objets et les relations entre objets afin d'en mesurer la similarité.

3.1) Le workflow existant d'harmonisation des données

3.1.1) Le workflow d'harmonisation existant avant la mise en application de l'AMR

Le processus d'harmonisation déjà en place permettait de réaliser une première phase d'automatisation du travail d'harmonisation pour la fusion de projets pour les demandes d'harmonisation des filiales. L'harmonisation pour la base de référence s'est toujours faite manuellement.

Le processus automatisé s'effectuait via ArcGIS et ProSource. ArcGIS permettait de charger des ShapeFiles de données de navigation sismique exportés depuis les différents systèmes de gestion de bases de données que l'on a cités plus haut. Le processus permettait aussi de calculer les attributs complémentaires pour caractériser de manière synthétique, à travers des attributs de géométrie, les lignes de navigation sismique.

Le croisement des lignes de navigation sismique se faisait à partir de ProSource, logiciel Schlumberger, permettant de se connecter aux bases de données FINDER, GeoFrame et OpenWorks et d'afficher le résultat des requêtes sous forme de tables.

3.1.2) Les attributs calculés et les méthodes de calcul utilisées avant l'AMR

Dans le processus existant, la comparaison des lignes sismiques suivait des attributs très synthétiques pour représenter une ligne. Ces attributs sont :

- Le nom de la campagne sismique
- Le nom de la ligne sismique
- Les coordonnées du centroïde de la ligne calculé par ArcMap
- La longueur linéaire de la ligne calculée par ArcMap
- L'azimut de la ligne calculé par ArcMap

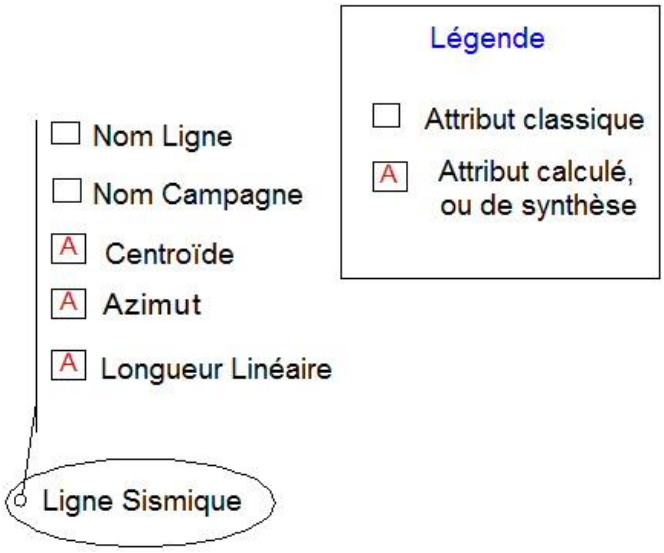


Figure 13 : Modélisation d'une ligne de navigation sismique

Les noms de campagne sismique et de ligne sont des attributs présents dans les bases de données. Centroïde, azimut, longueur linéaire sont des attributs calculés à partir des coordonnées de SP présents en base de données.

On ne peut pas avoir accès aux méthodes de calcul de ces attributs qu'utilisait ArcMap ni à l'algorithme de croisement utilisé par ProSource car les codes sont privés.
Le croisement se faisait toutefois en prenant en compte des facteurs de tolérance définis par l'utilisateur.

A l'issue de ce workflow, on obtenait une succession de groupes de lignes considérées comme étant des doublons.

Ce workflow ne permettait pas un croisement multicritère, c'est-à-dire un croisement réalisé en une passe sur l'ensemble des critères de comparaison. Les utilisateurs devaient donc procéder à plusieurs phases de croisement, et généraient plusieurs fichiers avec les doublons relatifs à chaque phase de croisement. Ils devaient alors faire la synthèse des résultats trouvés pour les différents attributs et pour les différents niveaux de tolérance utilisés à la fin des croisements.

Les différentes phases de croisement étaient par exemple :

- Croisement selon l'égalité des noms de ligne
- Croisement selon l'égalité exacte des longueurs linéaires
- Croisement selon l'égalité exacte des longueurs linéaires et (XOR logique) centroïdes
- Croisement selon l'égalité des longueurs linéaires à 100 m près
- Croisement selon l'égalité des longueurs linéaires et (XOR logique) centroïdes à 100m près
- Croisement selon l'égalité des longueurs linéaires et (XOR logique) centroïdes à 500m près
- Etc.

3.1.3) Les méthodes de croisement via InnerLogix (ILX : Logiciel Schlumberger de contrôle qualité des bases de données)

Il a été demandé à l'équipe Engineering Schlumberger chargée de ILX d'y programmer des méthodes de calcul des attributs de comparaison cités ci-dessus. Les formules que Schlumberger a utilisées pour les calculs sont les suivantes :

Équation 1 : Formules de calcul des coordonnées du centroïde d'une ligne de navigation sismique

$$X\_Centroïde = \frac{1}{nb\ SP\ ligne} \sum_{i=1}^{i=nb\ SP\ ligne} x\_sp(i)$$

$$Y\_Centroïde = \frac{1}{nb\ SP\ ligne} \sum_{i=1}^{i=nb\ SP\ ligne} y\_sp(i)$$

Équation 2 : Formule de calcul de la longueur linéaire d'une ligne de navigation sismique

$$Longueur\_Linéaire = \sum_{i=2}^{i=nb\ SP\ ligne} distance\_euclidienne(sp(i), sp(i-1))$$

Pour la longueur linéaire, cette formule semble très simpliste, car deux lignes exactement similaires, mais dont l'une aurait subi un ré-échantillonnage des SP différent de la première n'auraient pas la même longueur.

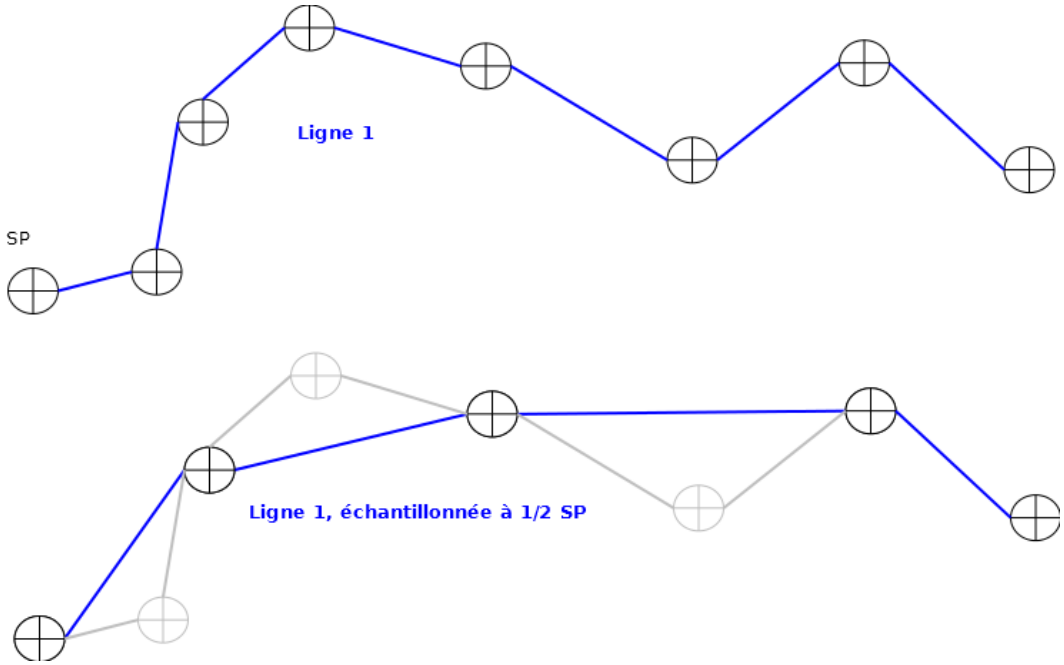


Figure 14 : Exemple des lignes de navigation sismiques ré-échantillonnées

ILX est capable de calculer les attributs et de croiser les données automatiquement. Les méthodes de croisement des données de ILX restent privées.

Les résultats de ILX sont des rapports d'analyse contenant les doublons trouvés, comme pour ProSource. L'avantage de ILX est sa capacité à effectuer une correction directe sur la base de données. Cet avantage paraît dangereux dans la mesure où si l'on ne possède pas de copie de la base, alors il n'y a plus de retour possible en arrière.

ILX permet aussi un contrôle graphique des doublons par visualisation des lignes. On peut noter que ILX fonctionne en se connectant directement aux bases de données via des connecteurs. Par conséquent, il y a un nombre restreint de bases de données auxquelles il se connecte, et TOTAL a demandé le développement de nouveaux connecteurs pour Access et ShapeFile d'ESRI, par exemple. Cela dit, aucune connexion à Sismage n'est permise par TOTAL. Il faut donc trouver un autre moyen pour réussir à traiter les données SISMAGE (contentant la majorité des données de navigation sismique) pour les harmoniser sous ILX qui ne traite pas les fichiers d'exports.

Cette problématique de connexion à Sismage est aussi valable pour l'harmonisation des puits et leur attribution d'identifiants uniques, s'ils sont stockés sous Sismage.

L'utilisateur final d'un outil de croisement multicritères dans un but d'harmonisation, de réconciliation ou d'aide à l'indentification aurait un profil de type géologue pour les aspects puits et géophysicien pour les aspects sismiques. Le fonctionnement d'outils de croisement multicritères doit donc être simple à paramétrer au niveau informatique, mais comporter une composante méthodologique et scientifique poussée.

3.1.4) Optimisation du workflow d'harmonisation des bases de données chez TOTAL par l'AMR

Les métriques de similarité mises en place permettent de mesurer la ressemblance entre deux objets géo-scientifiques. Elles peuvent être basées par exemple sur des méthodes de comparaison de noms, de calculs d'intersections et superpositions, de comparaisons de géométries ou de topologies. Elles peuvent aussi résulter de l'analyse des métadonnées disponibles selon les caractéristiques d'acquisition de la donnée. Le but est de quantifier et, si possible, visualiser la similarité entre deux objets comparés.

Pour optimiser les mesures de similarité pour les lignes de navigation sismiques, la première étape a été d'enrichir le modèle structurel d'une ligne par l'ajout d'attributs qui étaient stockés en bases mais pas utilisés pour les comparaisons. En deuxième étape, on a suivi le protocole de l'AMR en classant les attributs en critères de comparaisons ou simples attributs informatifs, en attributs calculés ou d'accès direct, puis selon les catégories de filtrage décrites dans le chapitre 2. Ainsi, on donne aux données un caractère pluridimensionnel plus étoffé. Par la suite, l'automatisation est réalisée grâces aux algorithmes et programmes de classification automatique décrits dans le chapitre 4.

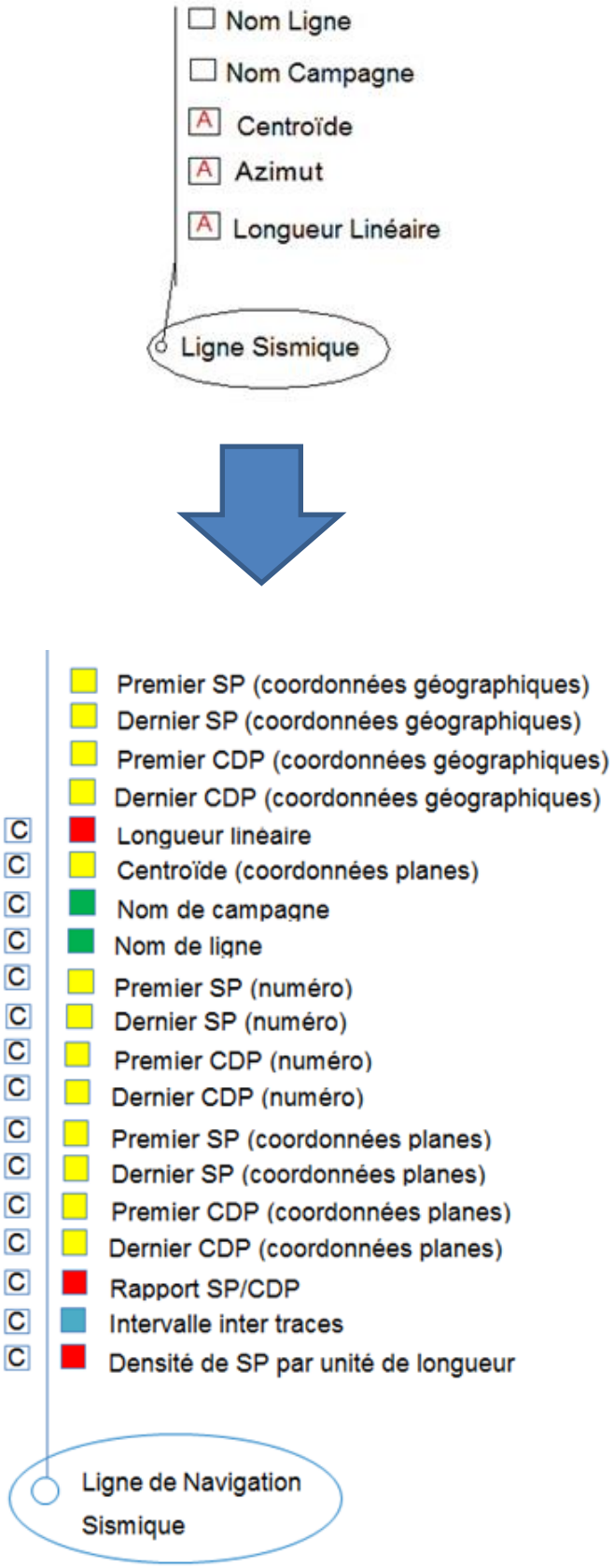


Figure 15 : Passage du modèle ancien de la ligne sismique au modèle AMR

3.2) Les métriques attributaires de similarité – spécialisation en fonction des critères de comparaison

A chaque nature d'acquisition correspond 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.

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 AMR 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.

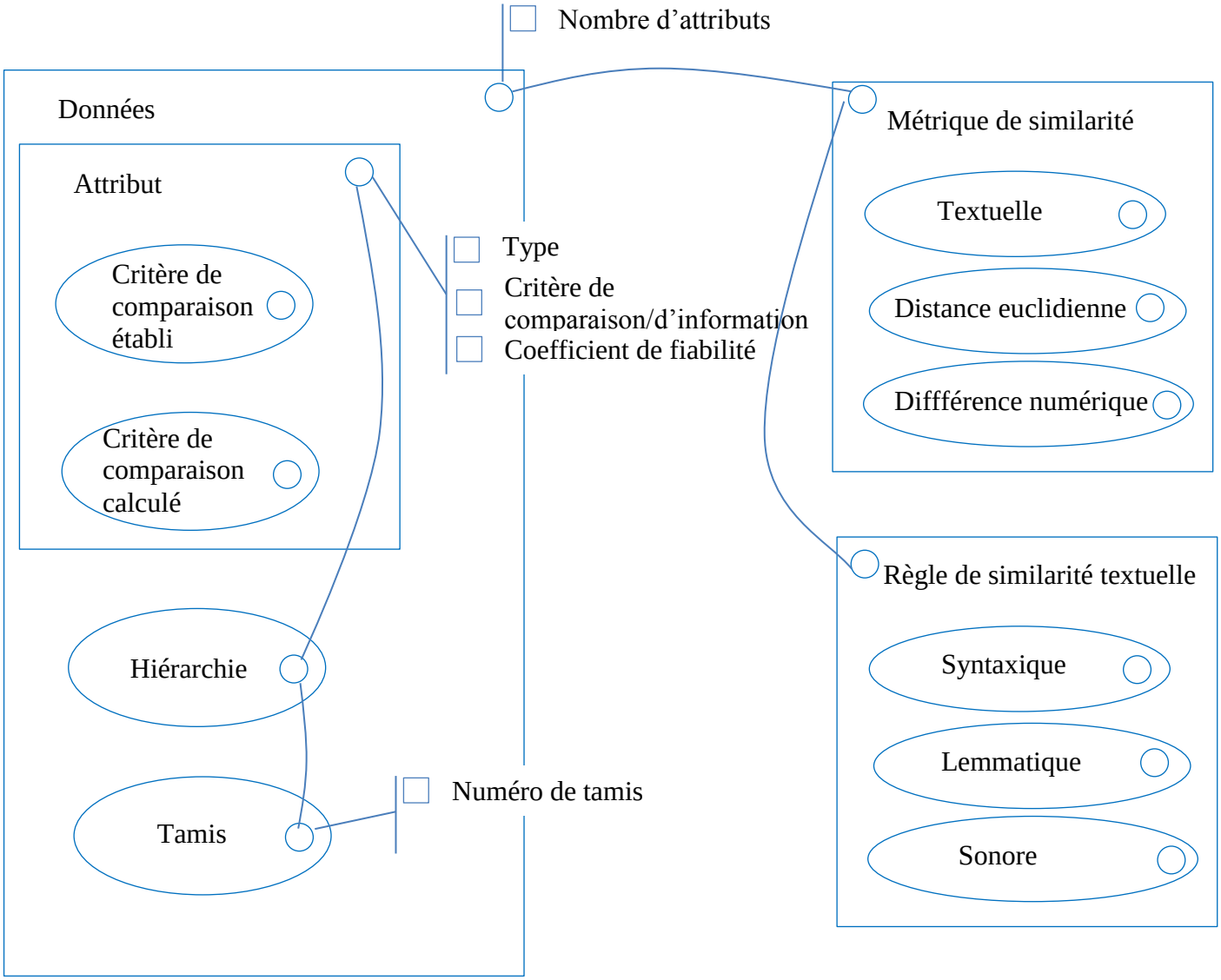


Figure 16 : Synthèse HBDS de l'AMR

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, par exemple :

- 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 le fait 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.

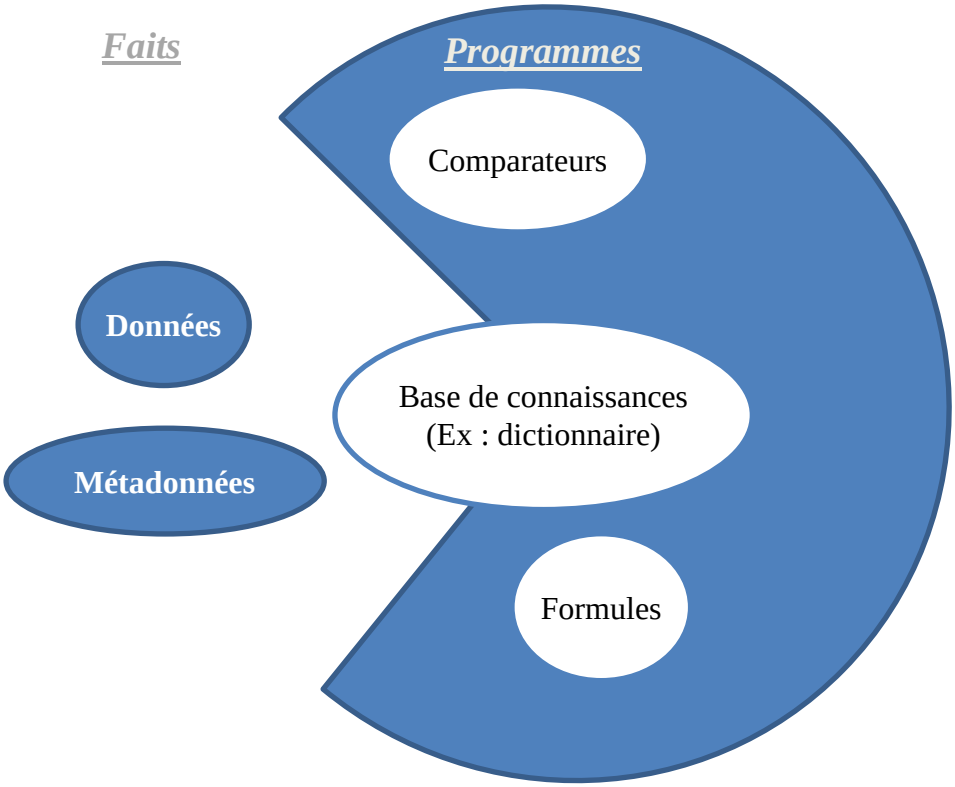


Figure 17 : Diagramme des Faits et des Programmes pour la mesure de ressemblance.

Dans la figure 17, on montre qu'une base de connaissance sera considérée à la fois comme un Fait et un Programme. Ces notions comme le passage d'une base de données à une base de connaissances sera développée dans le chapitre 5 sur les systèmes experts. Il est important toutefois de lier l'infrastructure de stockage de données à leur qualité et aux travaux d'harmonisation que l'on souhaite leur appliquer. En effet, dans les bases de données structurées, il est souhaitable que le modèle conceptuel des données facilite à la fois l'adaptation aux formats standards et le fonctionnement de processus d'harmonisation relatifs à la gestion des flux entrants et sortants de données.

3.2.1) Différentes approches pour mesurer la similarité

Le domaine de la similarité textuelle a été fortement étudié pendant les dernières années dans le domaine des systèmes d'information géographiques, notamment dans les laboratoires étudiant le rapport entre les SIG et Internet. Ainsi, dans des laboratoires comme Cardiff ont été créées des Ontologies de Toponymes (TO) à partir de sources d'information sur Internet. J'ai participé à l'un de ces projets en 2011 au laboratoire de recherches en Informatique de l'Université de Cardiff afin de réaliser une étude comparative des différentes mesures de similarité sur des échantillons de toponymes.

Le parallèle avec nos bases de données industrielles est légitime car les sources de provenance des toponymes étaient diverses elles aussi, sans compter le fait qu'il peut exister pour un même lieu différentes manières de l'écrire, donc différentes versions. L'utilisation d'ontologies de toponymes afin d'étudier les mesures de similarité à utiliser pour les bases de données industrielles est aussi justifiée par le fait que les puits ou lignes sismiques ont des noms qui contiennent souvent des toponymes. De plus, les ontologies de toponymes permettent de prendre en compte des phénomènes de synonymie, de ressemblances phonétiques ou de variations d'écriture que l'on retrouve dans les bases industrielles géophysiques. Un point commun supplémentaire est qu'un toponyme, par définition, est lié à des coordonnées géographiques ou à une zone géographique, tout comme les objets géophysiques.

Pour la suite, on va étudier une famille de mesures de similarité $(Mesures_k)k\epsilon 17$ où l'on étudie 17 mesures différentes.

Une mesure de similarité attributaire, par convention, doit être normalisée et positive (valeurs entre 0 et 1 qui pourraient ensuite être exprimées en pourcentages). Il est d'usage que la mesure augmente lorsque la ressemblance entre les objets mesurés augmente.
Il existe différents types de mesures de similarité. Parmi celles-ci, il s'agit de trouver et améliorer, voire combiner celles les plus adaptées au type d'attributs ciblés. Ces formules ont été développées dans des travaux de recherche en géomatique aussi bien qu'en bio-informatique, par exemple pour l'alignement de séquences d'ADN afin de trouver les gènes équivalents entre espèces.

On peut distinguer trois approches différentes pour réaliser des mesures de similarité.

L'approche textuelle consiste en la comparaison de chaînes de caractères pour savoir combien elles sont similaires. Deux types de mesures de similarité textuelle ont été répertoriés : les distances d'édition et les distances composites. On considère qu'une mesure textuelle est une distance élémentaire, « distance d'édition », si son calcul ne nécessite pas de modifier les chaînes de caractères comparées, et ne nécessite que la connaissance des caractères en commun entre les deux chaînes de caractères. Les caractères communs sont représentés comme l'intersection des alphabets que représentent les deux chaînes. Cependant, selon les versions de ces distances, l'intersection peut être l'ensemble des caractères communs quelles que soient leurs positions dans les deux chaînes, avec ou sans doublons, ou encore, de manière plus stricte, en tenant compte de la position des caractères. Ces variations jouent un rôle considérable dans la performance de la mesure de similarité, en fonction du but de celle-ci, et de la nature des données. Quant aux distances composites, elles peuvent appliquer des traitements spécifiques plus ou moins complexes aux chaînes de caractères, et utiliser les distances d'édition en tant que briques algorithmiques.

L'approche contextuelle des mesures de similarité prend en compte des paramètres autres que les toponymes pour évaluer la similarité entre les objets géographiques. L'une de ses particularités est qu'elle peut aussi utiliser les méthodes textuelles comme étapes de détermination de la similarité.

L'approche sémantique, enfin, peut utiliser des composantes textuelles ou bien contextuelles, mais elle se base surtout sur le sens des attributs/mots/toponymes. Dans cette approche, il s'agit de chercher les synonymes dans les attributs textuels. Dans l'approche textuelle, on considère que deux synonymes représentent deux objets différents. On ne reconnaît alors pas l'équivalence relative entre les mots « parc » et « jardin ». Le choix de l'approche dépend en grande partie des informations disponibles, comme des bases de données de référence, ou des dictionnaires, et de leur fiabilité de ces informations.

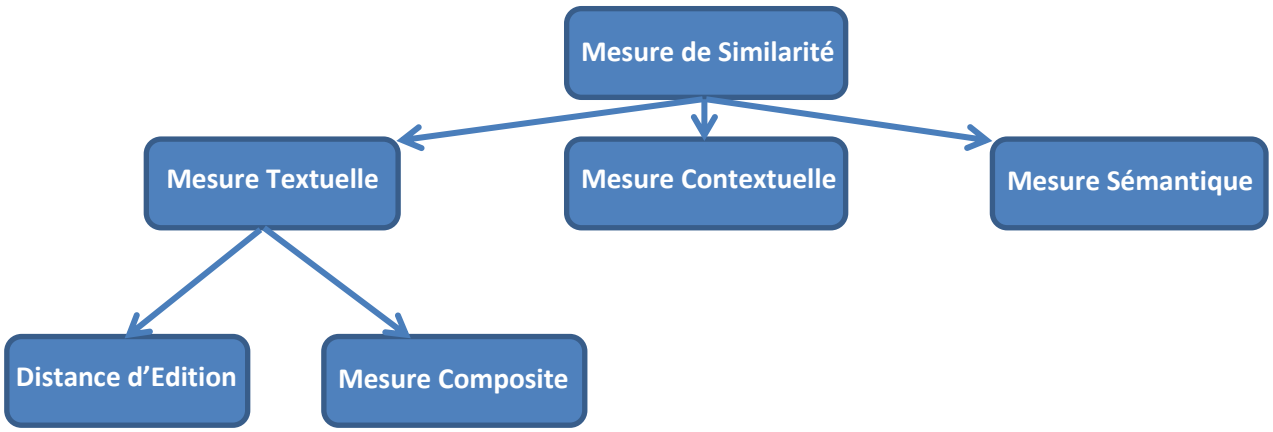


Figure 18 : Classification des mesures de similarité textuelle

On pourrait également mentionner les méthodes de mesure de similarité par corrélation d'images. Mais pour appliquer cela à de vastes bases de données, il faut être capable de d'abord produire ces images.

Pour étudier différentes métriques de similarité et les comparer entre elles, quatre échantillons de référence ont été formés sur Cardiff, Edinburgh, Paris, et en milieu rural en périphérie de Cardiff. En moyenne, chaque échantillon de référence contient 220 entités géographiques.

Nom de l'échantillon	Coordonnées de la requête	Rayon de la requête (m)	Nombre d'entités géographiques	Sources
Cardiff	-3.1804979 51.5828	700	264	Open Street Maps Geonames Wikimapia Foursquare Mastermap
Edinburgh	-3.213689 55.95287833	400	220	Open Street Maps Geonames Wikimapia Foursquare Google Places Point X (Ordinance Survey)
Paris	2.351074 48.857487	700	227	Open Street Maps Geonames Wikimapia Foursquare Google Places
Milieu rural (périphérie de Cardiff)	-3.481979 51.409165	6000	159	Open Street Maps Geonames Wikimapia Foursquare Google Places

Figure 19 : Extrait d'un échantillon de toponymes

Source 1	Name 1	Type 1	Source 2	Name 2	Type 2	Geographic distance	Decision
Open Street Maps	Natwest	Night Life and Business	Foursquare	Zizzi	Italian Restaurant	491.5	/false
Open Street Maps	Natwest	Night Life and Business	MasterMap	Windsor house	10026	152.6	/false
Foursquare	Cardiff Castle	Historic Site	Geonames	Cardiff Castle	S.CSTL	225.0	/true
Foursquare	Cardiff Castle	Historic Site	Wiki Mapia	Cardiff Castle	landmark	69.9	/true
Foursquare	Cardiff Castle	Historic Site	Foursquare	Cardiff University Trevithick Library	College Library	76.0	/false
Foursquare	Cardiff Castle	Historic Site	Foursquare	Cardiff Business School		76.0	/false
Foursquare	Cardiff Castle	Historic Site	Geonames	Cardiff Arms Park	L.PRK	225.0	/false
Foursquare	Cardiff Castle	Historic Site	Geonames	Cardiff Millenium Stadium	S.STDM	331.0	/false
Foursquare	Cardiff Castle	Historic Site	Wiki Mapia	Cardiff Arms Park	landmark	276.8	/false
Foursquare	Cardiff Castle	Historic Site	Wiki Mapia	Cardiff	city	513.2	/false
Foursquare	Cardiff Castle	Historic Site	MasterMap	Cardiff Arms Park	10026	329.3	/false
Foursquare	Cardiff Castle	Historic Site	MasterMap	Caradog House	10026	652.7	/false
Foursquare	Cardiff Castle	Historic Site	Foursquare	The Yard	Pub	76.0	/false
Foursquare	Cardiff Castle	Historic Site	Foursquare	Castle Welsh Crafts	Gift Shop	57.5	/false
Foursquare	Cardiff Castle	Historic Site	Geonames	Cathays Park	P.PPLX	313.0	/false
Foursquare	Cardiff Castle	Historic Site	MasterMap	Capital Tower	10026	347.1	/false
Foursquare	Cardiff Castle	Historic Site	MasterMap	Car Pk	10026	630.0	/false
Foursquare	Cardiff Castle	Historic Site	MasterMap	Castle Court	10026	264.9	/undefined
Foursquare	Cardiff Castle	Historic Site	MasterMap	Castle Mews	10026	351.7	/false

Figure 20 : Extrait d'un résultat de mesure de similarité sur l'échantillon de toponymes de Cardiff, avec prise de décision manuelle sur l'identification d'un doublon.

3.2.2) Précision et exhaustivité

Afin d'évaluer chacune des algorithmes de similarité étudiés, on introduit les deux indicateurs que sont la précision et l'exhaustivité. La précision permet de savoir si l'algorithme donne des valeurs correspondant à ce pourquoi il a été élaboré, donc s'il remplit bien son rôle pour la mesure de similarité. L'exhaustivité de l'algorithme ou de la métrique qu'il représente permet de savoir, on trouve statistiquement l'ensemble des objets censés être trouvés, pas moins, ni plus. On calcule ces indicateurs en pourcentages, 100% valant pour précision maximale et pour exhaustivité maximale. Un algorithme ayant 100% de précision et d'exhaustivité donnera des résultats « exacts ». Notons que précision et exhaustivité sont mesurées en référence à des résultats validés par des personnes expertes du sujet sur le même jeu de données.

Soient : NEM le nombre d'équivalences mesurées, $NEFM$ le nombre d'équivalences que la mesure n'aurait pas dû détecter, NER, le nombre d'équivalences de référence, c'est-à-dire validées par un expert.

La précision d'une mesure de similarité est alors définie par :

$$P(\%) = \frac{NEM - NEFM}{NEM} * 100$$

Et l'exhaustivité est définie par :

$$E(\%) = \frac{NEM - NEFM}{NER} * 100$$

3.2.3) Les $(Mesures_k)_{k\epsilon 17}$

Dix-sept métriques de similarité ont été étudiées afin de sélectionner les plus précises et les plus exhaustives d'entre elles pour des jeux de données à attributs hétérogènes. Parmi ces métriques figurent des méthodes textuelles élémentaires et composites, ainsi que des méthodes contextuelles. Par contre on n'a pas étudié de méthodes sémantiques par manque d'accès à des bases de référence pouvant servir de dictionnaires. Dans la suite de ce chapitre, on présente les différentes métriques et ce qui les caractérise.

Pour commencer, on introduit la distance d'édition de Levenshtein comme étant une métrique élémentaire qui suit un algorithme dans lequel sont comptées les opérations permettant de passer d'un mot à l'autre. Elle donne une précision de 100% sur des échantillons de référence constitués de toponymes mais d'exhaustivité insuffisante (45 %)

Ensuite, on décrit une deuxième métrique nommée Soundex, basé sur les comparaisons phonétiques des mots, et fonctionnant très différemment. Elle est d'une importante exhaustivité (74 %), au détriment de la précision, valant en moyenne 34% sur les mêmes échantillons de référence.

On constate qu'une combinaison linéaire pondérée des deux métriques mentionnées ci-dessus donne des résultats appréciables : 99% de précision et 50% d'exhaustivité en moyenne sur les quatre échantillons de référence.

$$Combinée(mot1, mot2) = \frac{1 * Levenshtein(mot1, mot2) + 4 * Soundex(mot1, mot2)}{5}$$

Les distances de similarité textuelles rendent compte de la proportion de caractères différents entre les deux chaînes de caractères. Ces différences sont dues à quatre types d'opérations : la transposition, la suppression d'un caractère, l'ajout d'un caractère et la substitution d'un caractère par un autre.

A chacune de ces opérations, excepté la transposition, l'algorithme de Levenshtein associe un coût. Lorsque les deux chaînes de caractères sont parcourues, la somme des coûts constitue la mesure de la dissemblance entre les deux mots. La mesure de similarité serait donc le complémentaire de la mesure de dissemblance par rapport à 1. Rappelons qu'une mesure de similarité doit être normée, la somme des coûts doit donc être elle aussi inférieure à 1, ou bien elle doit être normalisée par rapport à la longueur de la plus grande des deux chaînes de caractères comparées. Parmi les distances d'édition comme la mesure de Levenshtein, on trouve également la distance de Jaccard et la distance de Jaro-Winkler, plus appropriée aux langages à déclinaison car elle privilégie l'équivalence des mots ayant un préfixe commun.

$$Jaccard(mot1, mot2) = 1 - \frac{|mot1| \cap |mot2|}{|mot1| \cup |mot2|}$$

Deux caractères (i dans mot1 et j dans mot2) sont dits en correspondance si

$$|indice(i) - indice(j)| < Partie\ Entière\left(\frac{\max(|mot1|, |mot2|)}{2}\right) - 1$$

où $indice(i)$= position du caractère i dans le mot1
et $indice(j)$= position du caractère j dans le mot2

Soient :
l, le nombre de caractères du préfixe commun, avec | l |<4
p, le coefficient servant à privilégier les mots à préfixe commun. D'usage p = 0,1

La distance de Jaro est alors définie ainsi :

$$J(mot1, mot2) = \frac{1}{3} * (\frac{m}{|mot1|} + \frac{m}{|mot2|} + \frac{m - t}{m}$$

avec m, le nombre de caractères en correspondance
et t, le nombre de transpositions. Il s'agit du demi-nombre de fois où deux caractères en correspondance n'ont pas les mêmes indices.

Alors la distance de Jaro-Winkler est la suivante :

$$JaroWinkler(mot1, mot2) = J(mot1, mot2) + l * p(1 - J(mot1, mot2))$$

Quant à la méthode Soundex qui peut être composée avec chacune des différentes distances d'édition, on décrit ci-dessous une version de l'algorithme permettant de générer cette mesure composite :

- certains groupes de lettres sont modifiés selon des conventions phonétiques, ce qui peut dépendre de la langue des données. La première lettre du mot n'est jamais modifiée.

- on ne garde qu'une des lettres répétées, et on retire les voyelles, ainsi que le H, ou W, selon les versions du Soundex.

- un numéro est associé à chaque lettre, et différentes lettres peuvent avoir le même numéro, ce qui correspond à un groupe de sons similaires. Par exemple, b et v peuvent avoir le même numéro

- un code est alors généré pour chaque mot. Il correspond à la première lettre du mot suivie des numéros des lettres restantes après modifications. Ce code est ensuite tronqué au quatrième caractère compris.

- les codes des deux mots sont comparés alors grâce à une distance d'édition

Tableau 1 : Tableau des algorithmes en fonction des objectifs de mesure textuelle

Critère	Elément algorithmique impliqué
Longueur des chaînes de caractères (des toponymes)	Tronquer ou non le code
Langage	Combinaisons de lettres à changer selon les sons formés
But de la mesure : Exhaustivité élevée ? L'utilisateur souhaite extraire toutes les entités équivalentes à tout prix Précision élevée ? L'utilisateur veut être sûr que si une donnée est extraite, alors elle est fiable	 Association de numéros aux lettres Retrait ou non des lettres doublées Choix de la distance d'édition

Les mesures de similarité utilisées proviennent de deux bibliothèques, celle d'Apache, et la bibliothèque Simmetrics. Cependant le code source de la librairie d'Apache n'était pas disponible. On ne pouvait donc pas savoir quelle version de Soundex était utilisée. J'ai donc implémenté une nouvelle version de Soundex, intégrant des améliorations issues de mes expériences sur les données et de suggestions trouvées dans les travaux de recherche de David HOLMES et Catherine MCCABE, ainsi que dans les travaux de William COHEN**.**

Dans cette nouvelle version, la première lettre de chaque mot est maintenue, les autres lettres sont modifiées selon les équivalences phonétiques (c.f. Table 2). Les lettres doublées sont retirées, ainsi que les voyelles et W. H est retiré sauf s'il est précédé d'un A.

Les espaces entre les mots du toponyme et les signes de ponctuation sont aussi retirés. Aucun nombre n'est associé aux lettres afin d'augmenter la précision de la mesure. On considère que les modifications dans les équivalences phonétiques suffisent à obtenir une exhaustivité suffisante.

Afin d'être mieux adaptée aux toponymes, qui sont des chaînes de caractères composées d'un ou plusieurs mots, généralement de longueur supérieure à quatre, le code de chaque mot n'est pas tronqué. On appelle ici « code » la chaine de caractères composée de la première lettre du toponyme et des autres lettres après modifications phonétiques et filtrage de la ponctuation et des espaces. La similarité entre deux toponymes est enfin mesurée par comparaison de leurs codes grâce à la mesure d'édition de Dice.

$$Dice(mot1, mot2) = \frac{2 * (|mot1 \cap mot2|)}{|mot1| + |mot2|}$$

Toutes ces mesures ne sont pas des distances mathématiquement parlant, la mesure de Dice par exemple ne satisfait pas l'inégalité triangulaire. Cependant l'expérience montre que cette mesure est plus sensible que la distance de Jaccard (qui est une distance euclidienne) lorsque les données sont très hétérogènes. On applique Dice sur des mots qui ont déjà été traités via Soundex pour la mesure combinée Dice-Soundex.

L'intersection entre les deux toponymes représente les caractères communs entre les deux mots, quels que soient leurs indices. Les doublons ne sont pas exclus. Une intersection stricte considérerait que deux caractères sont communs seulement s'ils sont égaux et ont la même position dans le toponyme. Des traitements supplémentaires comme le retrait du nom de la ville dans le toponyme, dans le cas où le toponyme contient plus d'un mot ainsi qu'un traitement permettant d'appliquer Soundex séparément sur les différents mots des toponymes ont été appliqués.

Finalement, une précision de 95% et une exhaustivité de 62% en moyenne sur les quatre échantillons de référence sont obtenues. Le traitement des abréviations, par contre, a été testé et exclu car il augmente en effet l'exhaustivité de quelques pourcents, mais tend à diminuer la précision.

Tableau 2 : Les combinaisons de lettres modifiées par Soundex

Combinaison de lettres	Position : Préfixe	Position : Suffixe	Toute position
CA			KA
CC, CK			KK
CE			SE
CH		KK	
CHL, CL			KL
CHR, CR			KR
CI			SI
CO			KO
CS, CZ, TS, TZ	SS		
CU			KU
CY			SY
DG			GG
GH			HH
GN	NN		
HR, WR	RR		
HW	WW		
KN, NG	NN		
MAC, MC			MK
NST			NSS
NT		TT	
PF, PH			FF
RT, RDT		RR	
SCH			SSS
TIO, TIA			SIO
TCH			CHH

Cela dit, Soundex peut encore être amélioré en intégrant le traitement des caractères spéciaux comme pour transformer ElysÃ©e en Elysée, par exemple.

De plus, les tests sur l'effet de la distance d'édition sur cette version de Soundex montrent que Levenshtein, Jaro-Winkler, Dice donnent les mêmes résultats, par contre, les algorithmes d'alignement de séquences comme Needleman et Gotoh ne sont pas adaptés et donnent une très faible précision dans Soundex.

Remarquons enfin que la difficulté pour trouver des métriques de similarité donnant des résultats significatifs sur les données hétérogènes se situe dans un équilibre entre exhaustivité et précision. Dans la plupart des cas, lorsqu'on modifie la métrique pour augmenter l'exhaustivité, on perd en précision et inversement. Cependant, on remarque aussi qu'en introduisant des contraintes contextuelles sur ces mesures textuelles, on compense une partie de l'antagonisme entre précision et exhaustivité.

De plus, on a testé une mesure de similarité uniquement basée sur un algorithme de classification. Celui-ci donne les meilleurs résultats en termes d'exhaustivité, avec une baisse de la précision demeurant cependant bien meilleure que celle des distances d'édition pour une exhaustivité constante.

On peut déduire de ces études des métriques de similarité que la combinaison entre une méthode contextuelle, encapsulant des métriques textuelles, utilisée dans un algorithme de classification est la combinaison la plus adaptée pour détecter les doublons dans les bases de données industrielles pour obtenir une précision et une exhaustivité maximales à antagonisme réduit.

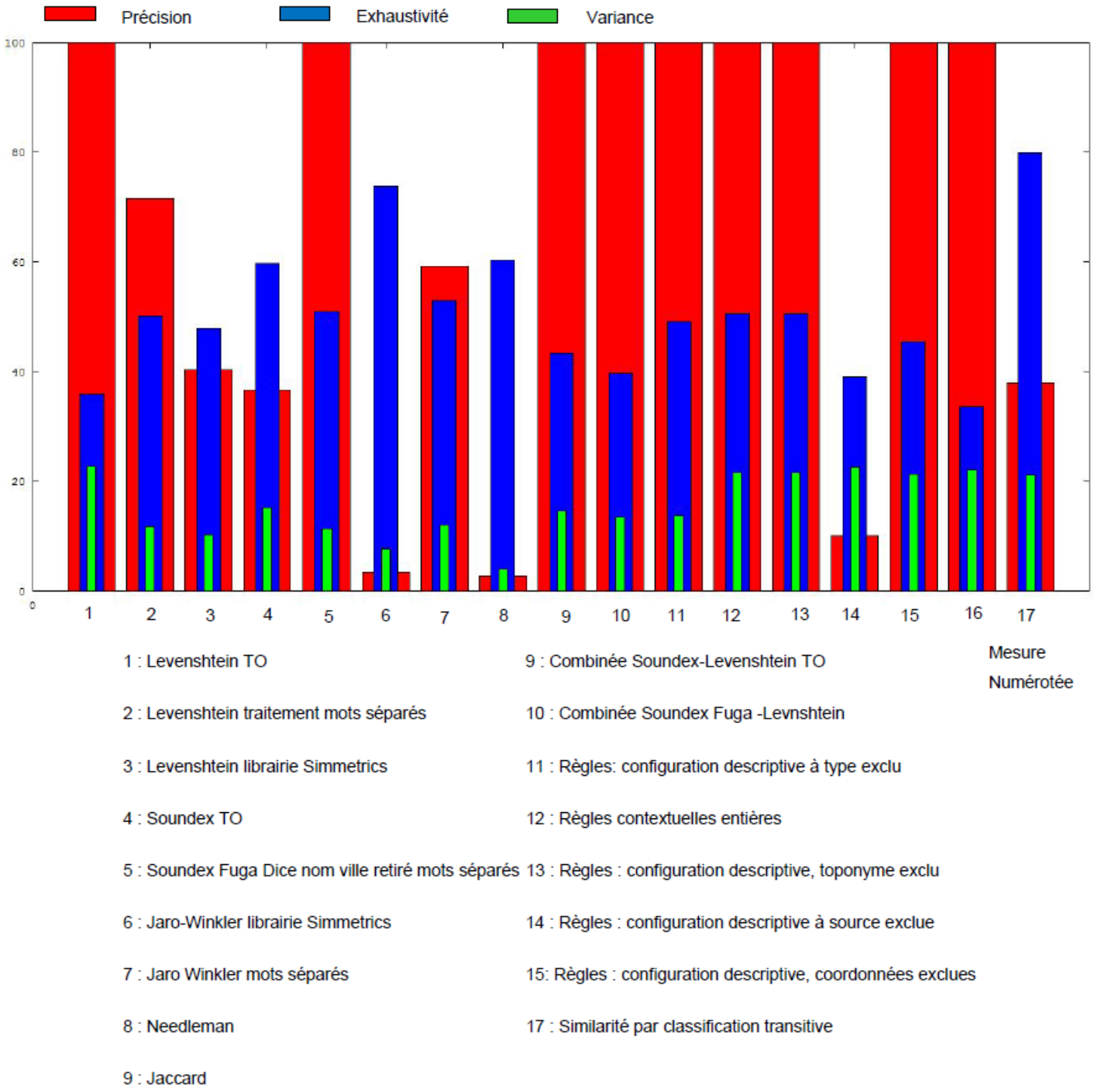


Figure 21 : Résultat de la comparaison des$(Mesures_k)k\epsilon 17$ mesures de similarité textuelles par Exhaustivité-Précision-Variance sur les échantillons de TO (Ontologie de Toponymes)

3.3) Etalonnage pour les mesures de similarité textuelle dans l'implémentation de l'AMR pour l'harmonisation des bases de données de TOTAL

Dans ce sous-chapitre on s'intéressera spécifiquement aux documents techniques, dans le cadre de l'application de l'AMR pour leur géoréférencement. Pour réaliser un étalonnage sur des mesures textuelles, ce type de donnée est très adapté car presque exclusivement textuel. Tous les champs attributaire, sauf des champs d'indices de référencement dans les bases de données ou des dates, sont des textes, parfois relativement longs.

On précise qu'il a été possible de géoréférencer les documents en cherchant dans les attributs de ceux-ci comme le titre ou un descriptif, des mots ou groupes de mots faisant référence à un nom de puits. En effet, les puits ayant des coordonnées, on peut alors rattacher les documents aux coordonnées des puits. Pour les documents ne faisant référence à aucun nom de puits, on cherche s'il est le doublon ou une version ou s'il a un lien suffisant avec un document pouvant, lui, contenir un nom de puits, donc être rattaché à des coordonnées.

Les titres de documents techniques sont des chaînes de caractères. Elles sont différentes des noms de puits ou des noms de lignes de navigation sismique, ou des noms d'auteurs car elles peuvent être composées de nombreux mots. Un mot se définit comme une sous-chaîne de caractères encadrée par des espaces.

Les phénomènes générateurs de vrais doublons dans les données analysées sont les suivants, dans l'ordre de gravité, traduisible en termes de coefficient de similarité à étalonner.

A l'échelle du mot

- L'insertion de caractères spéciaux parasites (comme «_ », « - », « / », « \ » etc.)
- La présence de zéros non significatifs dans les numérotations
- Les fautes de frappe comme
  - Les lettres doublées
  - La permutation entre deux caractères
  - La suppression d'un caractère
  - L'ajout d'un caractère
  - La substitution d'un caractère
- Les fautes d'orthographe ou différentes orthographes possibles, par exemple les « s » du pluriel en trop ou manquants… (difficile à programmer de manière générique, et polyglotte, donc la correction orthographique ne restera qu'approximative pour le moment. De toute façon les autres règles compensent pas mal cette règle)

A l'échelle du Titre (= ensemble de mots)

- La commutativité entre mots (cf. l'ordre des mots)
- Titres incomplets : il arrive qu'un titre A soit plus complet qu'un titre B, c'est-à-dire qu'il contient B, et le complète. Il peut contenir B en entier, et avec l'ordre originel des mots de B. Il peut contenir B en entier mais avec des permutations de mots. Il peut aussi contenir B de manière partielle. Il faudrait alors définir un coefficient de « proportion » afin de spécifier ce que l'on entend par « contenir ».

Les phénomènes générateurs de faux doublons :

Si l'on prend en compte les chiffres lorsqu'on a paramétré LAC avec un seuil de tolérance autorisant les permutations/substitutions/suppressions/ajout de caractères à l'échelle du mot.

⇨ Solution : discriminer chiffres et lettres, et ne prendre en compte que les lettres pour cette résolution de paramétrage.

L'étalonnage effectué :

Les mesures de similarité textuelle demandent d'analyser le texte sur deux aspects :

- Au niveau élémentaire du mot
- A l'échelle plus globale du titre.

Tableau 3 : règles de mesure de la similarité entre les mots, avec les valeurs de mesures qui leurs sont associées lors de l'étalonnage

Niveaux de similarité – Mots	Valeur du seuil en %
Egalité exacte entre tous les caractères des mots, dans l'ordre, sans tenir compte de la casse des caractères.	100
Egalité exacte entre tous les caractères des mots, dans l'ordre, et selon la casse des caractères, en enlevant les caractères spéciaux (_, #, -, @, /, \, ., ; , espace).	90
Egalité exacte entre tous les caractères des mots, dans l'ordre, et selon la casse des caractères, en enlevant les zéros non significatifs.	70
Egalité exacte entre tous les caractères des mots, dans l'ordre, et selon la casse des caractères, en enlevant les caractères spéciaux et en enlevant les zéros non significatifs.	60
Mesure de Jaccard : On enlève les caractères spéciaux, les zéros non significatifs, et les chiffres en général. Puis on étudie le rapport entre le nombre de caractères en commun et nombre total de caractères entre les deux mots.	50 à 0

Tableau 4 : règles de mesure de la similarité entre les titres, avec les valeurs de mesures qui leurs sont associées lors de l'étalonnage

Niveaux de similarité – Titres	Valeur du seuil en %
Egalité exacte entre tous les caractères des titres, dans l'ordre, sans tenir compte de la casse des caractères.	100
Egalité exacte entre tous les caractères des titres, dans l'ordre, et selon la casse des caractères, en enlevant les caractères spéciaux (_, #, -, @, /, \, ., ; , espace).	90
Egalité exacte entre tous les caractères des titres, dans l'ordre, et selon la casse des caractères, en enlevant les zéros non significatifs.	70
Egalité exacte entre tous les caractères des titres, dans l'ordre, et selon la casse des caractères, en enlevant les caractères spéciaux et en enlevant les zéros non significatifs.	60
Egalité complète en enlevant les caractères spéciaux, les zéros non significatifs, en tenant compte des redondances, mais quel que soit l'ordre des mots	50
Egalité complète en enlevant les caractères spéciaux, les zéros non significatifs, sans tenir compte des redondances, et quel que soit l'ordre des mots	40
Mesure de Jaccard entre les titres. Nous donne le rapport du nombre de mots communs (i.e. « similaires », selon la métrique de similarité définie dans le tableau des « Niveaux de similarité – Mots ») sur le nombre de mots total.	30 à 0

3.4) Similarité contextuelle

Dans le cadre des mesures de similarité, le contexte peut être défini comme l'ensemble de réalisations d'attributs ayant un impact suffisant sur la mesure effectuée, ou sur l'objet étudié pour en modifier l'état. En d'autres termes, il s'agit de l'ensemble des paramètres descriptifs des entités étudiées (toponyme, type, source, coordonnées) dont l'impact sur la mesure de similarité est suffisant.

Du point de vue relationnel, on pourrait considérer que même les mesures de similarité textuelle sont des paramètres descriptifs d'un lien, telle la distance géographique issue des deux couples de coordonnées du lien.

Connaître la fonction, le type, de l'entité géographique que l'on analyse est essentiel. Par exemple, les deux entités portant les toponymes « tennis court » et « tennis centre », éloignées l'une de l'autre de 15 mètres sont-elles équivalentes, sachant que les coordonnées dans les sources sont précises à plus de 10 mètres (positionnement par smart phone, donc positionnement en mode de navigation, dans la plupart des sources du Web) ?
La réponse serait qu'elles sont similaires car elles sont géographiquement proches et portent toutes les deux le mot clé « tennis ». Une autre réponse aurait pu considérer ces deux entités comme non équivalentes : cela dépend de l'échelle à laquelle on analyse les données, et du but de la classification. Si elle a pour but de cartographier l'ensemble des parkings d'une ville, alors il lui faut une mesure qui puisse discriminer les entités « Hilton Hotel » et « Hilton Hotel car park », même si l'une est la partie souterraine de l'autre. Les mêmes questions se posent pour les lignes de navigation sismiques.

Dans le cadre de la donnée géophysique, les attributs descriptifs sont en nombre bien plus élevé, ce qui donne un plus gros volume de données, en plus d'être plus hétérogènes.
Dans l'E&P, les questions permettant de dissocier les données semblables seront plutôt liées à des stratégies d'archivage et d'accès à la donnée mises en place, ou bien à des orientations choisies pour l'interprétation partant d'hypothèses initiales.

L’homme met en relation sa connaissance de l’objet réel, l’idée qu’il se fait de l’objet réel, et les informations que les sources de données lui apportent. Son avantage sur la machine est sa connaissance intuitive, ou empirique de l’objet en question. Ses méthodes de mesure de la similarité sont complexes, mettant en jeu cette intuition et expérience, mais surtout les caractères topologiques, géométriques, fonctionnels des objets qu’il compare ainsi que le dénombrement de leurs composantes, sans oublier la similarité sémantique qui reste difficile à implanter. La similarité sémantique se résume donc pour le moment à l’utilisation de la base de référence FINDER (ou plus généralement aux bases du Siège de TOTAL) comme dictionnaire.

3.5) Arbre de filtrage à tamis

L'AMR suit le courant des méthodes de mesure de similarité contextuelle. Comme on l'a expliqué dans le chapitre 2, les attributs sont regroupés par classes et hiérarchisés. On en forme un arbre de décision qui sera un « méta-comparateur » dans la classification automatisée. Ce méta-comparateur utilise des mesures de similarité textuelles si l'attribut est un texte ou un mot. Le type de mesure de similarité peut être choisi en fonction du type d'attribut textuel, par exemple on utilise des mesures différentes pour des noms de lignes de navigation sismiques et pour des titres de documents techniques.

On peut paramétrer simplement l'implantation de l'AMR, appelée LAC (Logiciel Automatique de Comparaisons) pour neutraliser un ou plusieurs des tamis ou critère(s) : le tamis, ou critère ne sera alors pas bloquant. De plus, deux types de classifications sont possibles : une classification globale suivant les règles de croisement et leur paramétrage, et une classification attribut par attribut. Par ailleurs, chacun des tamis peut contenir son propre arbre de décision interne.

Afin de définir un ensemble de filtres successifs, les critères de comparaison ont été classés en trois groupes :

- Une classe de critères « sémantiques », comprenant les critères permettant de nommer les lignes de navigation (exemple : le nom de la ligne)

- Une classe de critères « géométriques » comprenant les critères quantitatifs de géométrie caractérisant les lignes (par exemple coordonnées du premier SP)

- Une classe de critères relatifs à l'acquisition des lignes, regroupant les éléments de numérotation des critères géométriques (par exemple numéro de premier SP), ainsi que les paramètres d'acquisition dont on dispose comme l'intervalle moyen entre les traces, le rapport SP/CDP etc.

Dans chacune de ces classes, on attribue des valeurs de fiabilité mathématique, de fiabilité sémantique, de fiabilité de tolérance, et de fiabilité pour la comparaison.

La fiabilité sémantique concerne la justesse et la signification de la formulation mathématique dans le sens du « métier ». D'une certaine manière, il s'agit d'une justesse de modélisation du critère de comparaison. En effet, les formules de calcul peuvent être justes, mais le modèle inapproprié et inversement.

La fiabilité de facteur de tolérance permet d'évaluer en quelle mesure on est sûr des seuils que l'on utilise.

La fiabilité pour la comparaison permet de qualifier la capacité de discrimination du critère de comparaison pour le croisement des lignes sismiques.
Les trois tamis ont ensuite été hiérarchisés en fonction de ces fiabilités. D'après cette analyse, le tamis sémantique est le plus fiable pour le croisement, suivi du tamis géométrique et enfin du tamis d'acquisition.

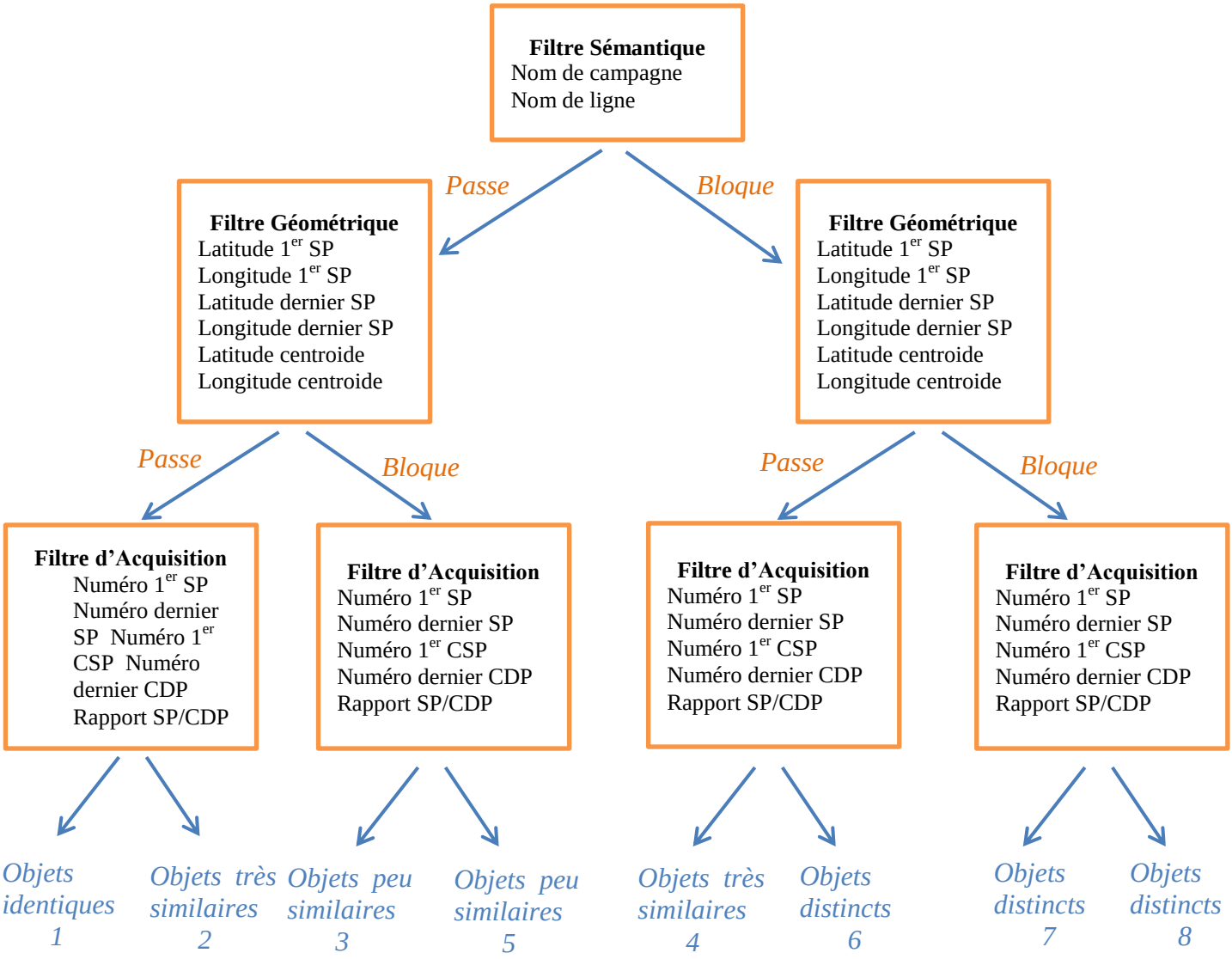


Figure 22 : Arbre de décision des règles de croisement multi critères pour les lignes de navigation 2D

Chaque chemin dans cet arbre est caractérisé par un numéro qui indique le degré de similarité des éléments comparés. Plus le degré de similarité d'un groupe de lignes, considérées similaires aux facteurs de tolérance près, est proche de 1 et plus le groupe est homogène, fiable. Plus il est proche de 8 et plus les lignes comparées sont distinctes.

Les degrés de similarité entre objets comparés permettent, lors du croisement global des lignes, de qualifier la qualité de chaque cluster en termes de similarité interne, entre les éléments qui le composent.

L'indicateur de qualité d'un cluster est calculé comme la moyenne des numéros des chemins des couples d'objets du cluster. Plus simplement, un cluster est composé d'un nombre N de lignes. Ces lignes forment 2 parmi N combinaisons possibles de couples. Chaque couple a subi les filtres de croisement pour être sélectionné dans le cluster. Donc chaque couple possède un chemin, avec un numéro dans l'arbre. On fait la moyenne des chemins de tous les couples du cluster.

3.6) Résolution et zone d'interfaçage

Comme expliqué précédemment, le choix entre deux lignes de navigation sismique dépend de la résolution à laquelle on regarde ces lignes. Il s'agit de la résolution de similarité. Selon la méthodologie AMR, 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 attributaires, 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 on compare 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, 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.

Il est intéressant d'aborder la question de la visualisation de 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.

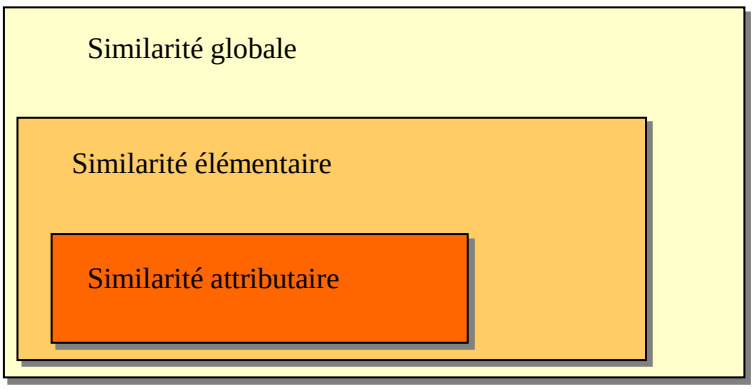


Figure 23 : Imbrication des différentes échelles de similarité

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 une sorte d'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 la faisabilité d'un lien entre une carte géographique construite par projection de coordonnées, et une telle carte de similarité, et qu'on développera au chapitre 6.

4) Regroupements

La construction de mécanismes de croisement de données segmenté constitue un procédé d'automatisation rendant possible le brassage de données volumineuses en peu de temps. Les algorithmes de classification automatique sont l'un des deux composants principaux de ces mécanismes.

Les méthodes actuelles de classification automatique sont soit déterministes soit non déterministes. Les méthodes déterministes nécessitent que l'on donne au préalable à la machine le nombre de groupes qu'elle devra former, telle la classification par nuées dynamiques. La classification est alors conditionnée par ce premier élément structurant des données. Par exemple, si l'on souhaite classer des arbres en un groupe d'arbres sains, un groupe d'arbres atteints d'une maladie et un groupe d'arbres secs potentiellement dangereux en cas de sécheresse, on envisagera d'avance qu'on ne souhaite voir que ces trois catégories. Il est aussi possible d'appliquer une méthode de classification automatique déterministe assistée : l'utilisateur se sert alors d'une visualisation spatiale de la donnée pour indiquer à la machine, en cliquant sur quelques représentants de groupes afin de déterminer visuellement (et humainement) le nombre final de classes.

D'un autre point de vue, si l'on ne souhaite pas pré-structurer les données, il existe des méthodes de classification non déterministes, où l'on ne sait pas d'avance combien de catégories seront formées. Selon les critères de catégorisation, on pourrait voir apparaître un groupe de jeunes arbres malades, un groupe de jeunes arbres sains, un groupe de vieux arbres sains, et aucun arbre sec. Ces méthodes permettent de constater l'émergence de groupes auxquels on ne s'attend pas.

Parmi les méthodes non déterministes figurent les classifications automatiques hiérarchiques ascendantes et descendantes. Dans l'un des cas on débute la classification en considérant que chaque donnée constitue un groupe : un singleton. Il s'agit alors de regrouper étape par étape les singletons voisins. Dans le second cas, on considère que toutes les données appartiennent au départ au même groupe, et on scinde ce groupe en plusieurs groupes étape par étape, en séparant les parties qui se distinguent les unes des autres. Lors d'une classification hiérarchique ascendante on constate un mouvement d'agrégation des données. On part des individus pour arriver à construire le groupe. Lors d'une classification hiérarchique descendante, on observe un phénomène d'« individuation » des données. On part du groupe pour aller vers les individus.

Le nombre de classes sera alors fonction du moment auquel on arrête la classification (le processus d'agrégation ou de désagrégation). Dans ce document, on nomme cela la résolution de la classification, et on choisit des méthodes de classification ascendantes hiérarchiques car elles se prêtent à une optimisation programmatique plus efficace. De plus, pour l'harmonisation des bases de données industrielles, les experts géophysiciens souhaitent, jusqu'à présent, examiner une donnée en retirant les doublons, ou pseudo-doublons de mauvaise qualité, si cela est possible. Il ne s'agit pas de faire une analyse des dynamiques de groupes, comme on aurait pu le faire en sociologie.

Pour faire le lien avec le chapitre précédent, lorsqu’on réalise une classification hiérarchique, le seuil de résolution correspond à l’indice du dendrogramme de la classification.

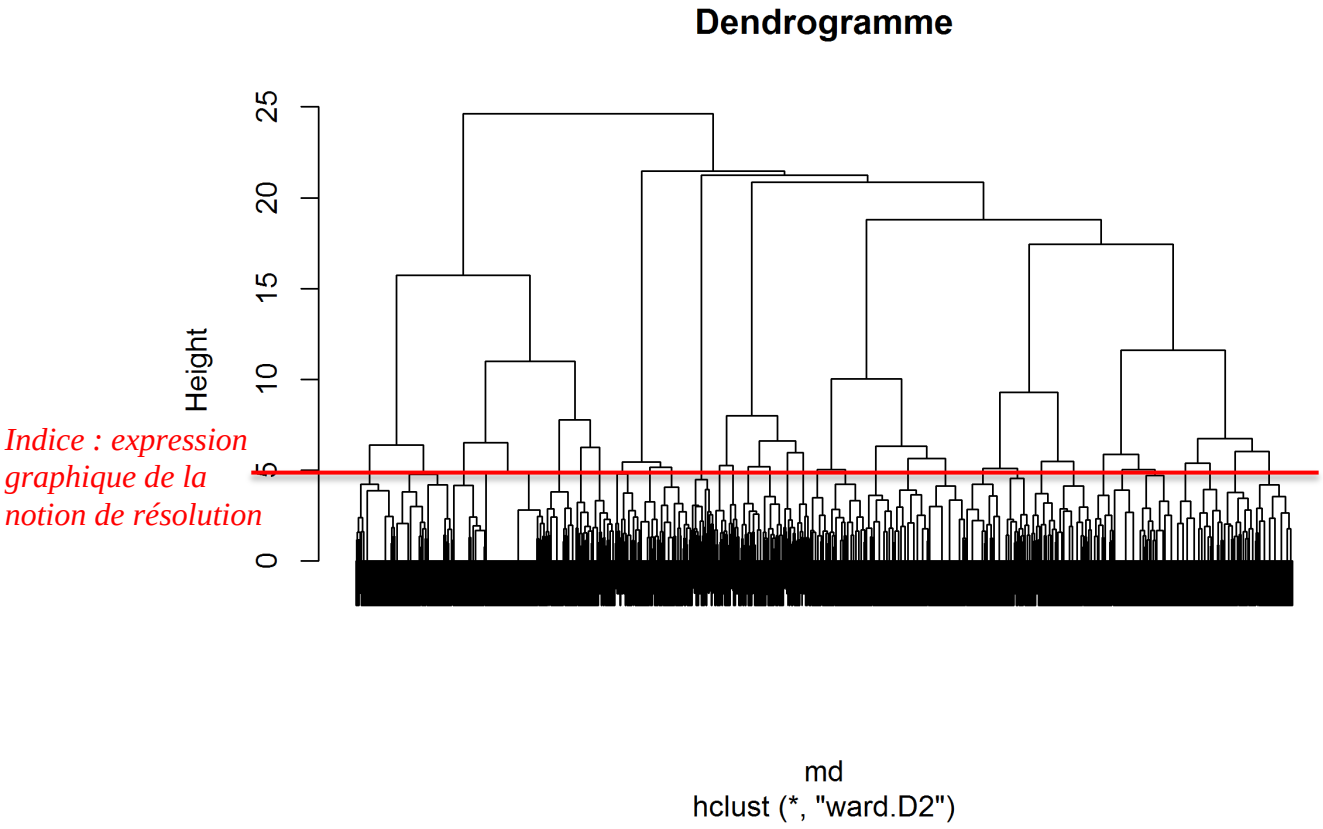


Figure 24 : Illustration d'un dendrogramme théorique représentant les différentes phases d'agrégation des données (les données étant placées sur l'axe horizontal).

Source : http://larmarange.github.io/analyse-R/classification-ascendante-hierarchique.html

Dans la figure 24, on peut visualiser d'une part le nombre d'étapes ayant été nécessaires pour aboutir à la classification. D'autre part, on peut y représenter graphiquement la notion de résolution de la classification telle qu'elle est définie dans l'AMR.

4.1) Différents types de classification automatique de l'AMR

4.1.1) 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 correspond à une situation spécifique de Data Management. L'approche algorithmique sera adaptative.

Dans certains types de bases de données, les informations peuvent être lacunaires, c'est-à-dire que tous les attributs caractérisant une donnée ne sont pas renseignés. 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. En effet, en prenant une classification par clusters, on obtiendrait toutes les données exactement similaires à l'attribut vide s'il correspond à un type d'attribut hiérarchiquement prépondérant, et cela conduirait à obtenir un unique cluster contenant toutes les données.

De l'autre côté, si on considère l'attribut vide comme exactement distinct de tout autre attribut, et s'il est placé dans un tamis prépondérant, on risquerait de ne pas pouvoir rattacher des données entre deux bases où pour la même donnée, différents attributs seraient renseignés dans chacune des bases. 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.

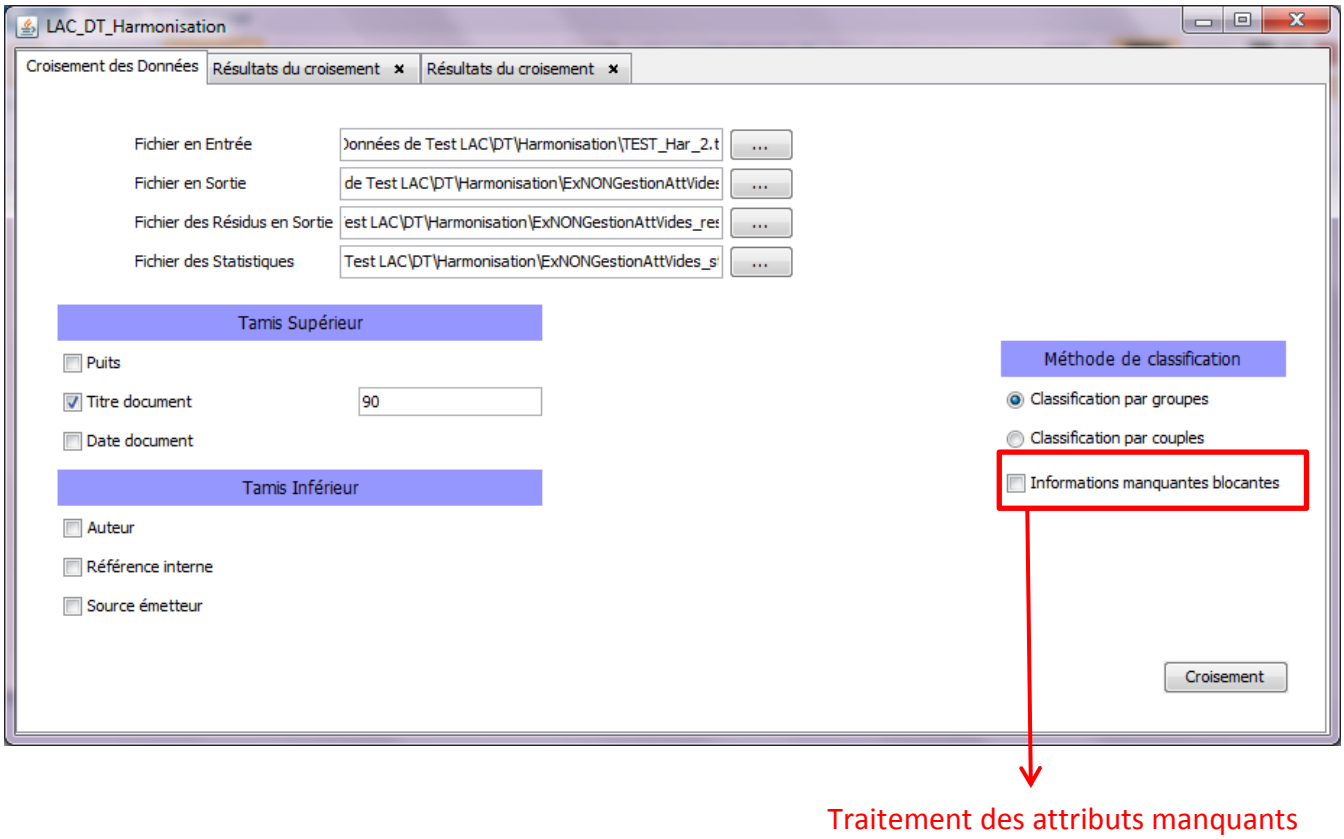


Figure 25 : Interface homme-machine du logiciel de comparaisons automatiques (LAC). Cette interface permet de paramétrer les informations nécessaires aux comparaisons et de gérer les attributs lacunaires.

Dans les exemples des figures 25, 26 et 27, on traite via le logiciel LACun ensemble de données où sont présents des attributs non renseignés. Il s'agit de comparer les résultats de la classification par mesure de ressemblance, lorsqu'on réalise une gestion de ces attributs, à une comparaison où l'on ne gère pas les attributs manquants.

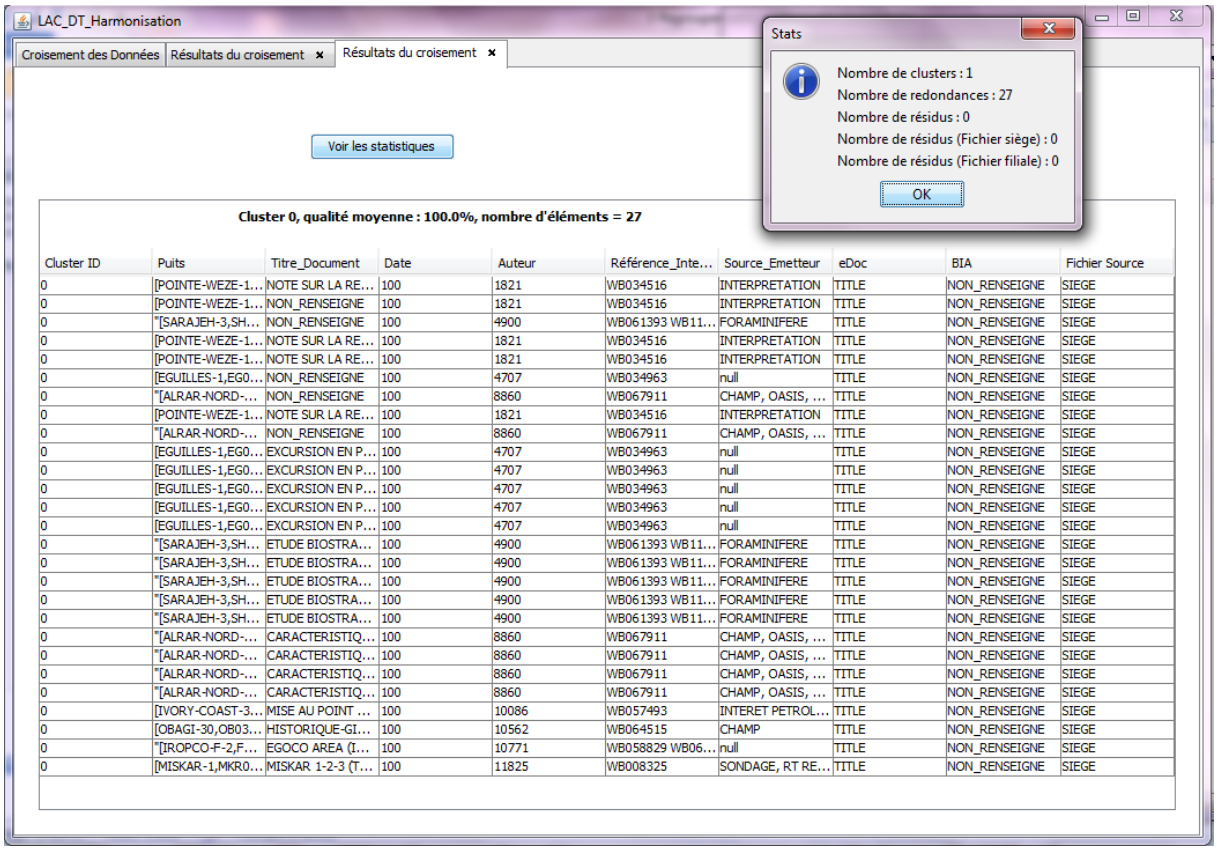


Figure 26 : Résultats de comparaison de données concernant la documentation technique associée à des puits de forage, sans gestion des attributs lacunaires (mention « NON RENSEIGNE »)

Dans le premier cas, le fait de ne pas gérer ces attributs, dans une classification par groupes, produit un résultat où il n'y a pas de distinction de différents groupes : pas de différentiation possible. Il est donc également impossible de repérer des doublons potentiels à harmoniser.

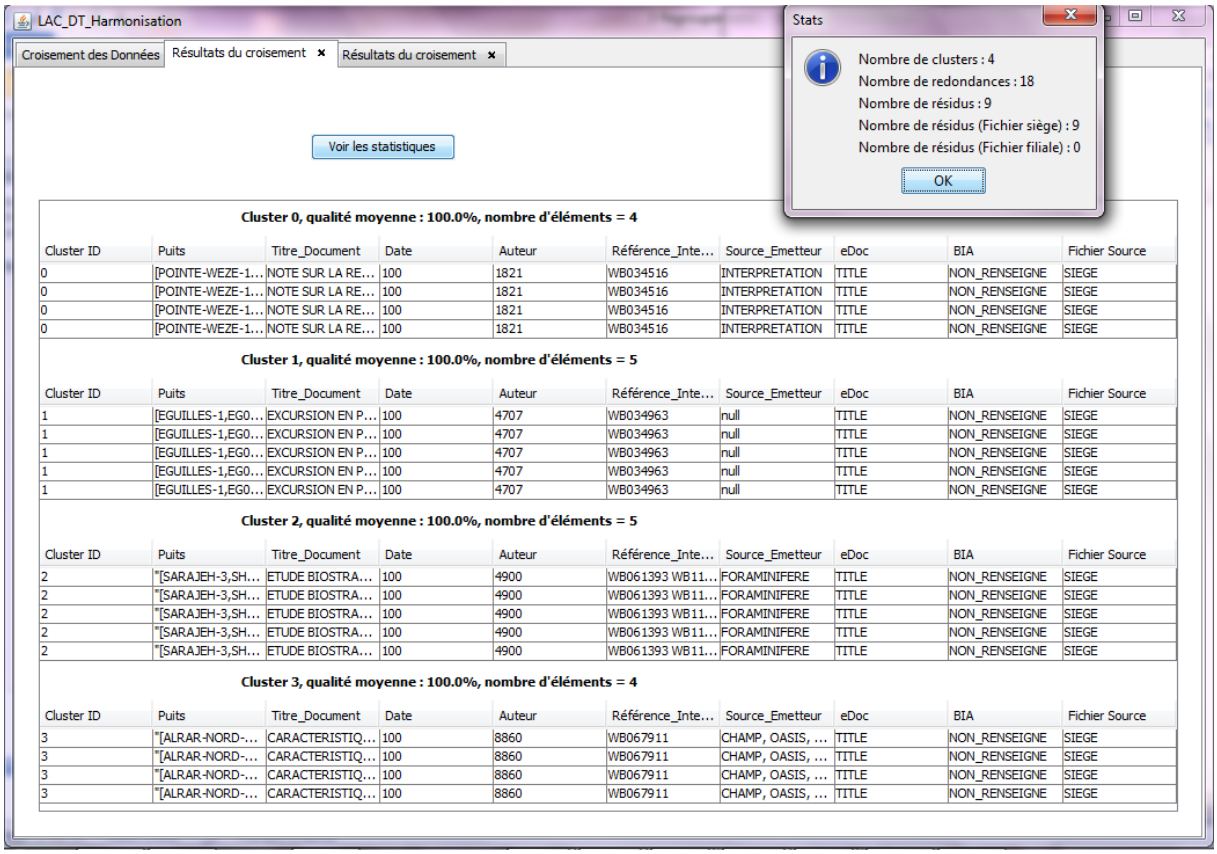


Figure 27 : Résultats de comparaison de données concernant la documentation technique associée à des puits de forage, avec gestion des attributs lacunaires (mention « NON RENSEIGNE »)

Par contre, dans le cas où le logiciel gère la présence des attributs vides, pour les mêmes critères de comparaison, et les mêmes seuils de résolution, quatre groupes de similarité apparaissent. Il est donc possible d'harmoniser les données comparées car la différentiation a eu lieu.

Il est méthodologiquement important de remarquer que la gestion par AMR des attributs manquants ne s'effectue pas de manière uniforme sur tous les attributs. Ne seront gérés que ceux des attributs non renseignés faisant partie des attributs sélectionnés comme critères de comparaison. Les autres attributs vides ne seront pas bloquants ou discriminants.

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 les suivantes :

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

------Begin Header----------
Export of geometry info for 2D line= LAP-12
Datum: WGS 84
Reference Meridian: Greenwich
Projection System: UTM zone 20N 63W
Unit: m
Average trace interval: 51.04752947062986
------End Header----------

survey_name	line_name	SP	CDP	X	Y	latitude	longitude	average_trace_interval
LAP	LAP-12	-46.0	1.0	1	1	10.013952136650445	-60.897772274730876	51.04752947062986
LAP	LAP-12	-44.0	2.0	2	2	10.013604875964432	-60.89746684753428	51.04752947062986
TT_80_WG	80-279B	3530.0	356.0	1	1	11.098401966631377	-60.139342646646874	12.736168862574477
TT_80_WG	80-279B	3531.0	356.0	2	2	11.098401966631377	-60.139342646646874	12.736168862574477
TT_80_WG	80-279B	3532.0	365.0	3	3	11.097638942789585	-60.13865992703093	12.736168862574477
TT_80_WG	80-279B	3533.0	365.0	4	4	11.097638942789585	-60.13865992703093	12.736168862574477
TT_80_WG	80-279B	3534.0	365.0	5	5	11.097638942789585	-60.13865992703093	12.736168862574477
TT_80_WG	80-279B	3535.0	365.0	6	6	11.097638942789585	-60.13865992703093	12.736168862574477
TT_80_WG	80-279B	3536.0	365.0	7	7	11.097638942789585	-60.13865992703093	12.736168862574477
TT_80_WG	80-279B	3537.0	365.0	8	8	11.097638942789585	-60.13865992703093	12.736168862574477
TT_80_WG	80-279B	3538.0	365.0	9	9	11.097638942789585	-60.13865992703093	12.736168862574477
LINE_WITH_NO_XYs	80-380B	2174.0	1934.0	1	1	10.042248164129722	-59.84822561251935	12.71558919897688
LINE_WITH_NO_XYs	80-380B	2170.0	1926.0	2	2	10.042889203078307	-59.84752492659772	12.71558919897688
LAP	LAP-12	-46.0	1.0	1	1	10.013952136650445	-60.897772274730876	51.04752947062986
LAP	LAP-12	-44.0	2.0	2	2	10.013604875964432	-60.89746684753428	51.04752947062986
TT_80_WG	80-279B	3537.0	365.0	8	8	11.097638942789585	-60.13865992703093	12.736168862574477
TT_80_WG	80-279B	3538.0	365.0	9	9	11.097638942789585	-60.13865992703093	12.73616886257447
LAP	LAP-12	-46.0	1.0	1	1	10.013952136650445	-60.897772274730876	51.04752947062986
LAP	LAP-12	-44.0	2.0	2	2	10.013604875964432	-60.89746684753428	51.04752947062986
LINE_WITH_NO_XYs	80-380B	2174.0	1934.0	1	1	10.042248164129722	-59.84822561251935	12.71558919897688
LINE_WITH_NO_XYs	80-380B	2170.0	1926.0	2	2	10.042889203078307	-59.84752492659772	12.71558919897688

Figure 28 : Exemple de données à harmoniser. Les lignes surlignées d'une même couleur sont des exemples de doublons.

- la réconciliation d'informations et de différents supports (par exemple entre une base de données comportant les données de navigation d'une campagne sismique, et une base de données contenant les traces sismiques de cette même acquisition sismique.)

- la reconstitution et le rattachement documentaire

Titre	Mot libre	Doc_ID	Puits:Alias puits
ETUDE GEOPHYSIQUE PAR LA METHODE SISMIQUE - BASSIN SEDIMENTAIRE DE LA COTE D IVOIRE - DECEMBRE 1953 - AVRIL 1954 - TEXTE – PLANCHES	NON_RENSEIGNE	96387	BAOBAB-I-3
GEOLOGIE - COTE DIVOIRE /MISSION DE PRERECONNAISSANCE PETROLIERE EN A.O.F. - BASSIN SEDIMENTAIRE DE COTE D IVOIRE - RAPPORT TECHNIQUE D ACTIVITE (DECEMBRE 1953 - MARS 1954)	NON_RENSEIGNE	8515	BAOBAB-P-1
NOTE PRELIMINAIRE SUR LES POSSIBILITES PETROLIERES DU BASSIN DE LA COTE D IVOIRE	NON_RENSEIGNE	427064	BAOBAB-P-2
NOTE PRELIMINAIRE SUR LES POSSIBILITES PETROLIERES DU BASSIN DE LA COTE D'IVOIRE	NON_RENSEIGNE	453434	BAOBAB-P-3
PETROLES DE LA COTE D IVOIRE	NON_RENSEIGNE	438205	BAOBAB-P-4
SENEGAL ET COTE D IVOIRE	NON_RENSEIGNE	427086	BAOBAB-P-5
NOTE POUR MONSIEUR BENEZIT (PROPECTIONS GEOPHYSIQUES) PLUS 6 PLANCHES	NON_RENSEIGNE	453440	BAOBAB-P-6
GEOLOGIE - COTE D IVOIRE /GEOLOGIE DU BASSIN SEDIMENTAIRE DE LA COTE D IVOIRE	NON_RENSEIGNE	7034	BAOBAB-P-7
ETAT ACTUEL DE L EXPLORATION ET PERSPECTIVES DU SAHARA ALGERIEN ET DE LA COTE D IVOIRE	INTERET PETROLIER	28539	BAOBAB-P-8
PROBLEME PETROLIER EN BASSE COTE D IVOIRE	NON_RENSEIGNE	427610	BAOBAB-P-9
PROBLEME PETROLIER EN BASSE COTE D IVOIRE, DONNEES GEOLOGIQUES RECHERCHES EFFECTUEES PROGRAMME A ENVISAGER	NON_RENSEIGNE	429339	BAOBAB-P-10
COTE D'IVOIRE ET GOLD COAST	NON_RENSEIGNE	521548	BAOBAB-P-13
NOTE SUR LE BASSIN DE LA COTE D IVOIRE	NON_RENSEIGNE	2144	BAOBAB-P-13R
MESURES GRAVIMETRIQUES ET MAGNETIQUES DANS LA PARTIE CENTRALE DE L AOF - MALI - COTE D IVOIRE - HAUTE VOLTA - NIGER - TOGO – BENIN	INTERPRETATION	22025	BAOBAB-P-14A
MISE AU POINT DE NOS CONNAISSANCES ACTUELLES SUR LE BASSIN DE COTE D IVOIRE ET PROJET DE PROGRAMME DES TRAVAUX (PERMIS COTE D IVOIRE)	NON_RENSEIGNE	2143	BAOBAB-WP-11
FORAGE DE GRAND-BASSAM /	NON_RENSEIGNE	2148	B-3 ; B-3X ; B003 ; BASSAM-3X
ETUDE PHOTOGEOLOGIQUE DU BASSIN DE LA COTE D IVOIRE	NON_RENSEIGNE	2138	BELIER-AO-1 ; BLR-01
PROJET D EMPLACEMENT DE SONDAGES GEOLOGIQUES PETIT BASSAM A ET B PB.A , PB.B REGION D ABIDJAN (COTE D IVOIRE)	NON_RENSEIGNE	2139	BELIER-AO-2 ; BLR-02
PROJET D EMPLACEMENT DE SONDAGES GEOLOGIQUES, PETIT BASSAM A ET B, REGION D ABIDJAN, COTE IVOIRE	PROJET	27982	BELIER-AO-3 ; BLR-03
NON_RENSEIGNE	NON_RENSEIGNE	NON_RENSEIGNE	BELIER-AO-4 ; BLR-04
NON_RENSEIGNE	NON_RENSEIGNE	NON_RENSEIGNE	BELIER-AO-5 ; BLR-05
NON_RENSEIGNE	NON_RENSEIGNE	NON_RENSEIGNE	BELIER AO-6 ; BELIER-AO-6 ; BLR-06

NON_RENSEIGNE	NON_RENSEIGNE	NON_RENSEIGNE	BELIER AO-7 ; BELIER-AO-7 ; BLR-07
NON_RENSEIGNE	NON_RENSEIGNE	NON_RENSEIGNE	BELIER AO-8 ; BELIER-AO-8 ; BLR-08
NON_RENSEIGNE	NON_RENSEIGNE	NON_RENSEIGNE	BELIER AO-9 ; BELIER-AO-9 ; BLR-09
NON_RENSEIGNE	NON_RENSEIGNE	NON_RENSEIGNE	BELIER A0-10 ; BELIER-AO-10 ; BLR-10
NON_RENSEIGNE	NON_RENSEIGNE	NON_RENSEIGNE	BELIER-AO-11 ; BAOBAB-WP-11
NON_RENSEIGNE	NON_RENSEIGNE	NON_RENSEIGNE	BELIER-AO-15 ; BLR-15
NON_RENSEIGNE	NON_RENSEIGNE	NON_RENSEIGNE	BELIER-AO-16 ; BLR-16
NON_RENSEIGNE	NON_RENSEIGNE	NON_RENSEIGNE	BELIER-AO-17 ; BLR017
NON_RENSEIGNE	NON_RENSEIGNE	NON_RENSEIGNE	BELIER-AO-18 ; BLR018
NON_RENSEIGNE	NON_RENSEIGNE	NON_RENSEIGNE	BELIER-AO-19 ; BLR019
NON_RENSEIGNE	NON_RENSEIGNE	NON_RENSEIGNE	BELIER-AO-20 ; BLR020
NON_RENSEIGNE	NON_RENSEIGNE	NON_RENSEIGNE	BELIER-AO-21 ; BLR021
NON_RENSEIGNE	NON_RENSEIGNE	NON_RENSEIGNE	BELIER-AO-22 ; BLR022
NON_RENSEIGNE	NON_RENSEIGNE	NON_RENSEIGNE	BELIER-AO-23 ; BLR023
NON_RENSEIGNE	NON_RENSEIGNE	NON_RENSEIGNE	BELIER-AO-24 ; BLR024

Figure 29 : Exemple de données à rattacher, sur deux sources différentes, l'une correspondant aux données du Siège et l'autre aux données d'une filiale. Le document surligné en violet sera rattaché aux deux puits BAOBAP-WP-11 et BELIER-AO-11.

- le géo-référencement
- le croisement multicritère pour l'analyse et l'interprétation des phénomènes

4.1.2) 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 de données ». 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 comparer les données d'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. Dans une même région, des campagnes d'acquisition sismique peuvent avoir été faites par différentes technologies, ou méthodologies. De plus différents traitements par exemple d'analyse du signal peuvent avoir été appliqués aux 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 de perte de données lors de migrations de bases ou de changements de support de stockage de l'information.

L'algorithme de regroupement par couples est présenté en annexe.

4.1.3) 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 de prendre comme repère 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. L'algorithme est en annexe.

4.1.4) 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 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.

Cet algorithme est adapté à l'harmonisation de données provenant de la même source, ou de données de sources distinctes à condition de ne pas avoir besoin de différentier ces données selon leurs sources.

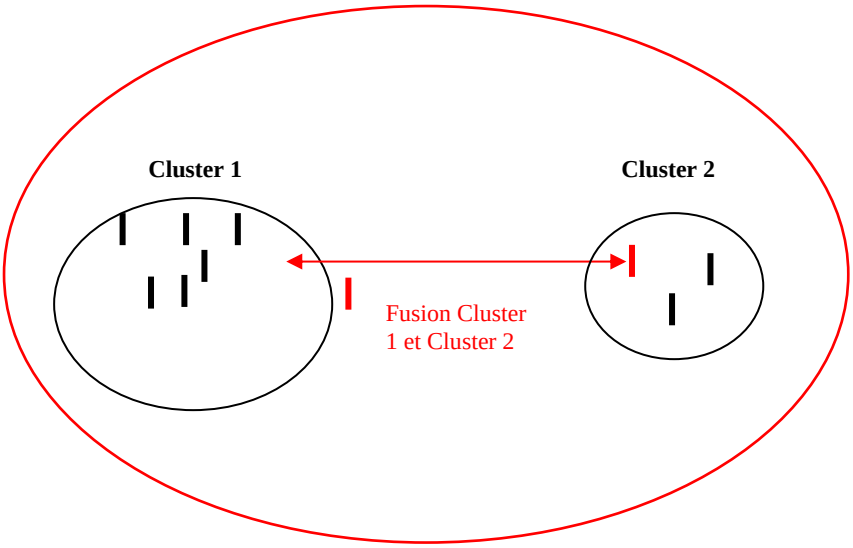


Figure 30 : Exemple de fusion entre deux clusters. Phénomène de contagion.

L'algorithme de classification automatique par groupes, couplée avec le système de filtrage est en annexe.

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 aux différentes natures de critères. La classification est automatique, ainsi que l'application du système de filtrage et des mesures.

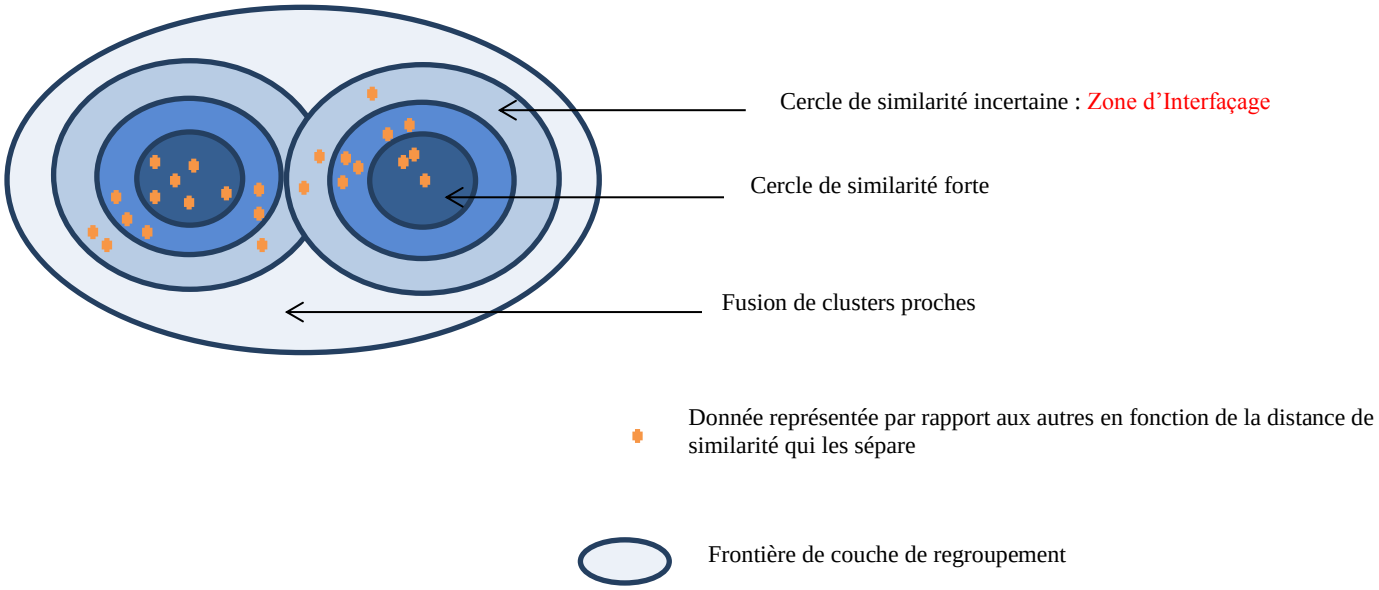


Figure 31 : Représentation de données dans un graphique de similarité, avec les contours de clusters formés selon différents vecteurs de similarité.

4.1.5) Les différences entre les trois algorithmes de classification et leur combinaison

Les couples

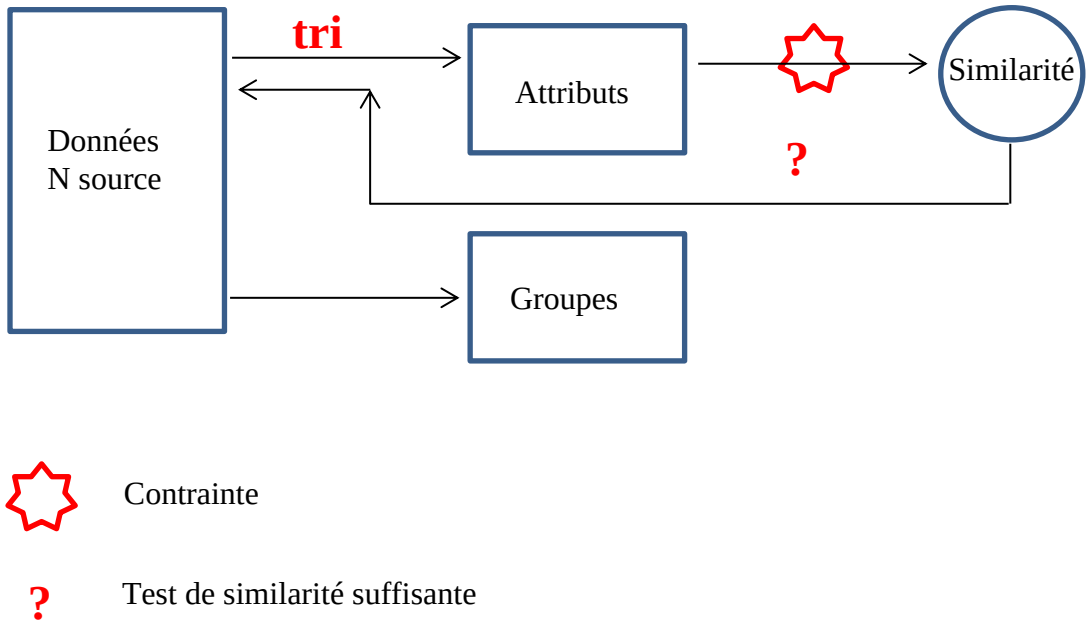


Figure 32 : Schématisation de mesure automatisée de ressemblance avec une classification en structure de couples.

Dans la classification par couples, on distingue trois mécanismes spécifiques : un tri attributaire, une contrainte constituée par une conversion des attributs en un chiffre représentant leur degré de ressemblance, et un système de seuillage.

Dans cette classification, on considère que les données se trouvent toutes à un même niveau de fiabilité, et aucune n'est un point de comparaison privilégié pour les autres.

Les groupes asymétriques

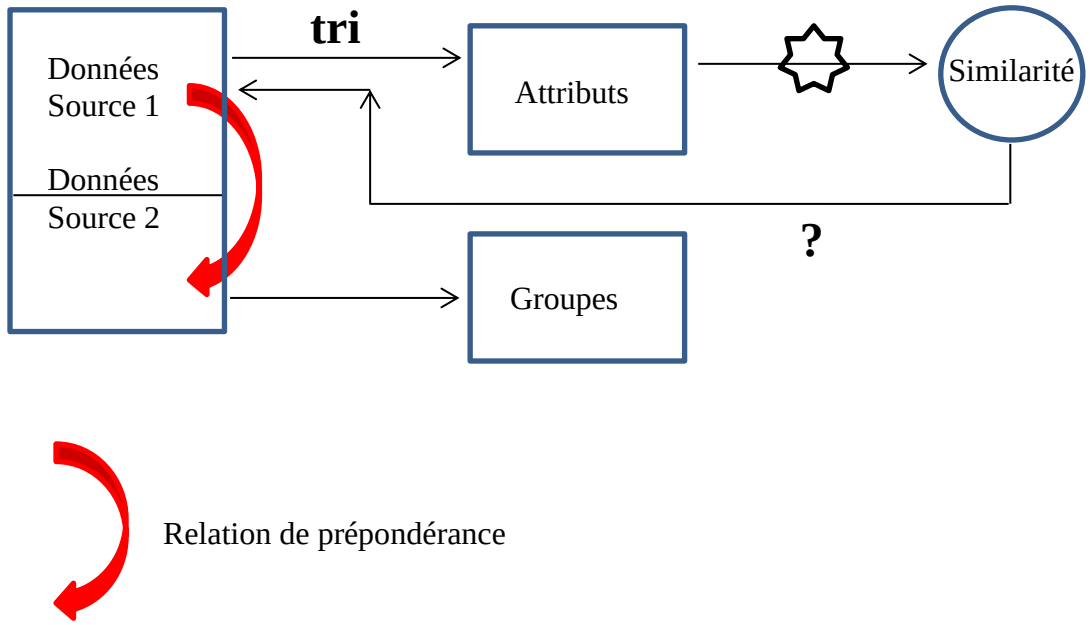


Figure 33 : Schématisation de mesure automatisée de ressemblance avec une classification en structure de groupes asymétriques

Cette méthode de classification contient tous les mécanismes décrits pour les couples. Elle se distingue par une asymétrie placée sur un attribut des données, comme la base de données de provenance. Les données provenant de la source n°1 seront prépondérantes par rapport aux données provenant de la source n°2, c'est-à-dire qu'il y aura une relation de classification de 1-n. Une donnée de la source n°1 pourra être reliée à plusieurs données de la source N°2, mais pas l'inverse. De plus, les données de la source N°1 ne seront pas comparées entre elles. Une application de cette méthode sera utile lorsqu'on dispose déjà d'une base de données de référence sur laquelle on réalise un contrôle qualité régulier.

Les clusters

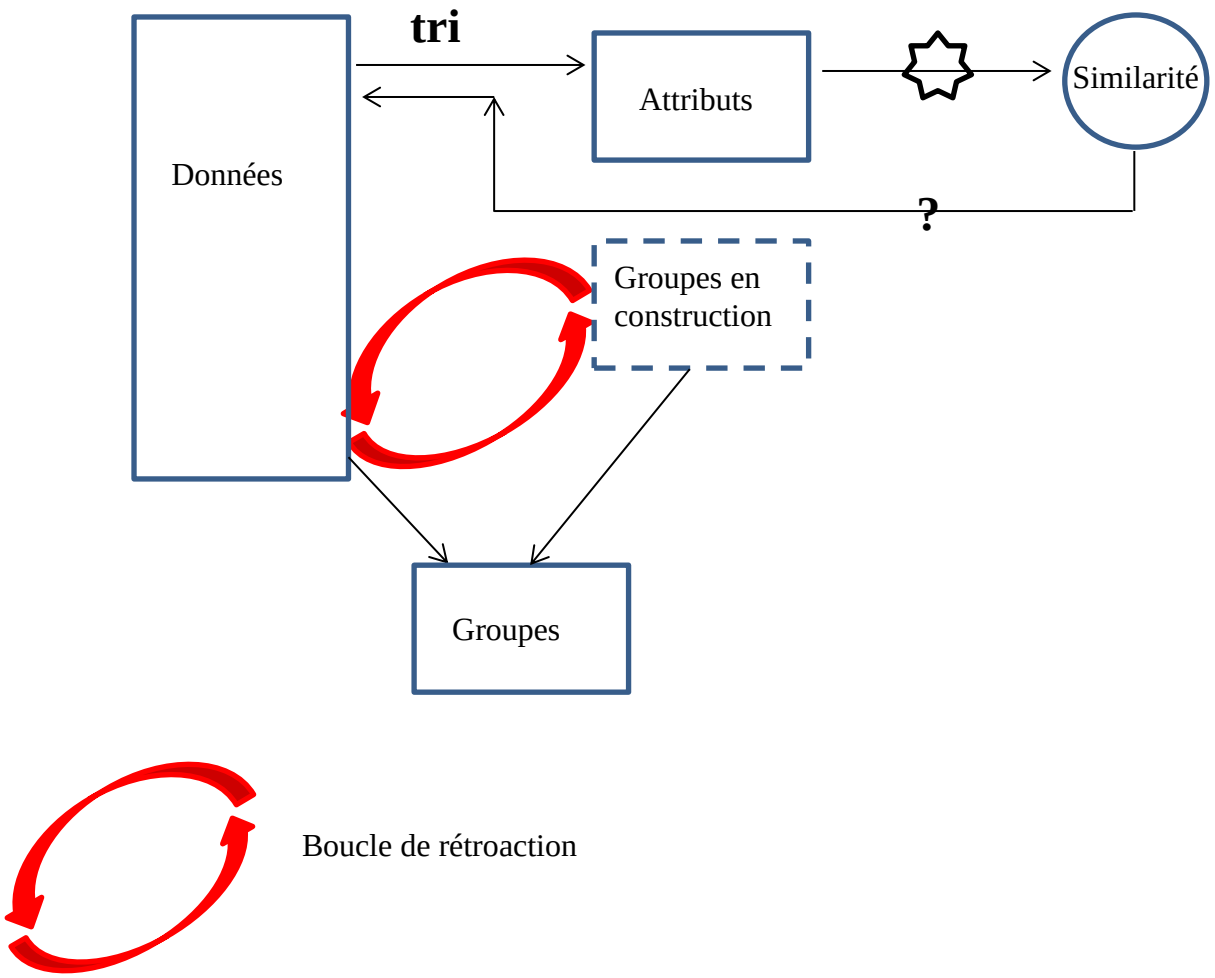


Figure 34 : Schématisation de mesure automatisée de ressemblance avec une classification en structure de clusters

La classification en clusters proposée dans l'AMR suit quant à elle un principe de boucle de rétroaction.

Elle représente les mécanismes de fusion de clusters, et de construction progressive des classes. Cette rétroaction a l'avantage d'accélérer le processus de regroupement grâce à une indexation des données selon le numéro du cluster temporaire d'appartenance.

3.1.5) Synthèse adaptative

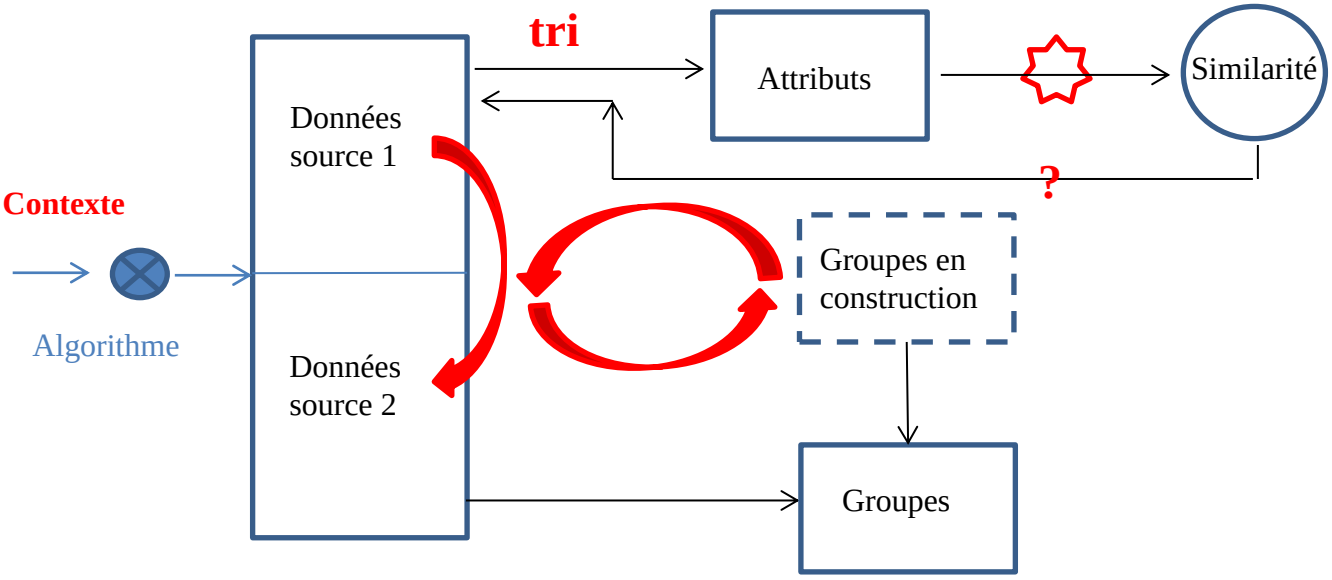


Figure 35 : Schématisation de mesure automatisée de ressemblance avec une classification en structure adaptative

Si on superpose les trois algorithmes précédents, on obtient un algorithme très synthétique permettant de s'adapter aux différentes situations « data management ». Une situation « data management » est une configuration de la base de données ou des bases de données associée à un besoin tel que le dé-doublonnage, ou le rattachement de données partiellement renseignées les unes aux autres (cette configuration arrive par exemple dans une base de données étant le résultat de plusieurs fusions historiques antérieures). L'algorithme global de classification ne doit pas seulement s'adapter aux situations « data mangement » mais aussi à la résolution à laquelle on souhaite analyser les données, c'est-à-dire à la tolérance que l'on souhaite avoir pour les données. Dans toutes les classifications proposées pour l'AMR, ici ou toute autre classification que l'on souhaiterait utiliser à la place, il est indispensable d'avoir un algorithme non déterministe concernant le nombre de classes. Remarquons que le workfow de réconciliation de documents techniques a suivi un processus utilisant deux types de classifications : les clusters pour harmoniser les données de chacune des sources et la classification par couples pour réconcilier les deux sources ensuite.

4.2) Tests : Exemple de résultats obtenus sur le Brésil

Pour comparer les données issues du traitement automatique à celles comparées manuellement, il est nécessaire d'appliquer les mêmes facteurs de tolérance et reproduire les mêmes conditions de test. L'une des motivations de ces tests a été la comparaison en précision des résultats et en performances entre la méthode manuelle et LAC, puis entre l'outil InnerLogix et LAC afin de valider l'AMR et l'une de ses implémentations pour ces cas d'utilisation communs. La comparaison manuelle consiste suit cinq phases :

- **Première étape** : extraction de tous les doublons suivant le critère de stricte égalité du nom de ligne

- **Deuxième étape** extraction de tous les doublons suivant le critère de stricte égalité du centroide

- **Troisième étape** : extraction de tous les doublons selon l'égalité exacte de la longueur linéaire

- **Quatrième étape** : extraction de tous les doublons selon l'égalité du centroide et de la longueur linéaire à 100m près

On peut préciser que les tests à tolérances plus grandes n'auraient pas de sens au niveau de l'application métier. De plus, on a choisi ces trois attributs que sont le nom de ligne, le centroide et la longueur linéaire parce que ce sont les seuls utilisés par la méthode manuelle.

Les résultats des tests sont les suivants :

4.2.1) Etape 1 : Egalité exacte entre les noms de lignes

75 clusters trouvés par la méthode manuelle, et 77 sont trouvés par LAC. Les 75 clusters manuels sont aussi trouvés par LAC. 2 clusters supplémentaires justes sont trouvés par LAC. Ici LAC est plus exhaustif que la méthode manuelle, mais tout aussi précis.

Tableau 3 : Résultats des tests pour la phase 1

Nb vrai clusters communs Lac et Workflow manuel	75
Nb vrai clusters manqués par erreur par LAC	0
Nb vrai clusters correctement manqués par LAC	0
Nb vrai clusters trouvés correctement en plus par LAC	2
Nb faux clusters trouvés par LAC	0
Nb vrai faux clusters trouvés par la méthode manuelle	0

4.2.2) Etape 2 : Egalité exacte entre centroides

76 clusters trouvés dans la méthode manuelle, dont 11 qui sont marqués comme étant de faux doublons. Donc 65 cluster de vrais doublons.

65 clusters trouvés par LAC. 64/65 = 98% des clusters de doublons vrais trouvés par la méthode manuelle ont été trouvés par LAC. Seul un cluster manque. Et un cluster a été trouvé par LAC mais pas par la méthode manuelle.

Dans ce test, LAC est à la fois plus précis et plus exhaustif que la méthode manuelle.

Ce cluster est un vrai doublon après vérification. Par ailleurs, on peut noter que LAC n'a pas trouvé de faux doublons. Ce point nous a fait réaliser que le workflow manuel (utilisant ArcMap et ProSource via des exports SHP de SISMAGE) n'était pas tout à fait fiable dans le calcul de la longueur linéaire et des centroïdes. Ces tests sont donc à lire avec un certain recul pour les données quantitatives.

Tableau 4 : Résultats des tests pour la phase 2

Nb vrai clusters communs Lac et Workflow manuel	64
Nb vrai clusters manqués par erreur par LAC	1
Nb vrai clusters correctement manqués par LAC	11
Nb vrai clusters trouvés correctement en plus par LAC	1
Nb faux clusters trouvés par LAC	0
Nb vrai faux clusters trouvés par la méthode manuelle	11

4.2.3) Etape 3 : Egalité exacte entre les longueurs linéaires

87 clusters trouvés dans la méthode manuelle, dont 11 qui sont marqués comme étant de faux doublons. Donc 77 clusters de vrais doublons.

79 clusters trouvés par LAC. 100% des clusters de doublons vrais trouvés par la méthode manuelle ont été trouvés par LAC. Et deux clusters ont été trouvés par LAC mais pas par la méthode manuelle. Ces clusters sont de vrais doublons après vérification. Par ailleurs, on peut noter que LAC n'a pas trouvé de faux doublons.

Tableau 5 : Résultats des tests pour la phase 3

Nb vrai clusters communs Lac et Workflow manuel	77
Nb vrai clusters manqués par erreur par LAC	0
Nb vrai clusters correctement manqués par LAC	11
Nb vrai clusters trouvés correctement en plus par LAC	2
Nb faux clusters trouvés par LAC	0
Nb faux clusters trouvés par la méthode manuelle	11

4.2.4) Etape 4 : Egalité exacte entre les longueurs linéaires et les centroïdes tolérance 100

117 clusters trouvés dans la méthode manuelle, dont 84 sont aussi trouvés par LAC. 2 vrais clusters sont trouvés par LAC et manqués par la méthode manuelle. 90 clusters sont trouvés par LAC, donc 73% de clusters communs avec la méthode manuelle. LAC ayant trouvé 9 vrais clusters supplémentaires, il est plus exhaustif que la méthode manuelle pour ce test.

Tableau 6 : Résultats des tests pour la phase 4

Nb vrai clusters communs Lac et Workflow manuel	84
Nb vrai clusters manqués par erreur par LAC	9
Nb vrai clusters correctement manqués par LAC	36
Nb vrai clusters trouvés en plus par LAC	6
Nb faux clusters trouvés par LAC	0
Nb faux clusters trouvés par la méthode manuelle	36

Ces tests nous montrent que la méthode automatique de recherche de doublons dans les données du Brésil donne des résultats meilleurs que la méthode manuelle en termes de précision et d'exhaustivité. La méthode manuelle nécessite trois semaines de travail pour un géophysicien pour des données contenant 4400 lignes de navigation sismique. Cela explique le fait que nous n'ayons pas pu faire de plus nombreux tests pour obtenir un échantillon à étudier avec des indicateurs statistiques.

Sur ces données, LAC et ILX donnent les mêmes résultats, à peu de lignes près. On présume que les différences constatées viennent du fait que LAC utilise directement les coordonnées cartographiques du fichier originel pour calculer la longueur linéaire et le centroïde, tandis que ILX utilise les coordonnées géographiques du fichier originel qu'il reconvertit en coordonnées cartographiques. On précise que ILX ne peut pas donner des mesures de similarité chiffrées, ce qui ne nous permet pas d'effectuer la comparaison au-delà de l'étude par rapport au référent commun qu'est la méthode manuelle et des performances temporelles, présentées ci-après.

4.2.5) Performances temporelles pour les données Brésil (4411 lignes, et 1381137 SP)

Tableau 7 : Performances temporelles de LAC

Tâche	Temps d'exécution LAC
Export des données depuis SISMAGE en ASCII	~1h
Réorganisation des données et calcul des attributs	1h50 (sur réseau TOTAL), 30min (en local)
Classification Globale des données (tous les attributs, ie 20 attributs pris en compte)	5 min
Classification des données par attribut (16 attributs)	1h20 i.e. 16*5 min

On considère chaque coordonnée comme un attribut. En interface, on ne manipule que des couples de coordonnées, ce qui fait 14 critères à sélectionner car 6 d'entre eux sont des couples.

Tableau 8 : Performances temporelles de ILX

Tâche	Temps d'exécution ILX
Export des données depuis SISMAGE en ASCII	~1h
Chargement des données dans OW	2h
Classification Globale des données (3règles Unicity et 11 Completeness)	20 min

5) Dimension système expert – automatisation

Le premier harmonisé contenait 4411 lignes sismiques, avec un nombre total de 1 381 137 points de tir, dans un fichier .dat de 1,6 Go. On peut considérer qu'on commencer à entrer dans le cadre de données volumineuses.

L'augmentation des flux de données ne s'opère pas uniquement dans le domaine industriel des géosciences mais un peu partout avec les technologies de l'information et de la communication qui augmentent en performances. En effet, le prix des technologies de stockage de l'information et des ordinateurs baisse depuis des années, la quantité de données produites et mesurées par des systèmes numériques et électroniques augmente. La technologie est de plus en plus sophistiquée pour traiter la donnée, notamment en temps réel, ou du moins aussi vite qu'elle est produite. Ces phénomènes ont pour conséquence un intérêt accru pour le développement de systèmes cognitifs qui permettraient par exemple de meilleures gestions des connaissances et compétences dans les entreprises et universités, d'accéder à des outils plus efficaces pour la prévention et la gestion de crises aussi bien économiques que relevant des risques naturels ou sociétaux. Ainsi, le travail d'harmonisation des bases de données concerne-t-il, plus généralement, la gestion de l'information et de la donnée : information sous forme stockable et représentable.

On distingue deux types de traitements des données : les traitements statiques et les traitements dynamiques. Les traitements statiques concernent une utilisation prédéterminée de la donnée, l'automatisation de tâches est faisable, la logique est entièrement implémentée et les algorithmes échouent devant le changement et l'incertitude. Ce sont des types de traitement qui fonctionnent bien pour des données structurées et peu transformables. Les traitements dynamiques quant à eux s'adaptent aux changements dans les données, ils pourraient être munis d'une capacité d'apprentissage issue de l'expérience sans être reprogrammés.

Certains mécanismes sont spécifiques aux systèmes cognitifs et systèmes experts, comme le fait de traiter des données de l'intérieur du système et ainsi que des données provenant de l'extérieur du système, d'identifier des patterns (schémas de comportements et situations déjà répertoriés), évaluer une donnée dans un contexte et aider à la prise de décision.

Dans l'AMR on exerce une démarche d'évaluation contextuelle des données car on fait appel à des mesures de similarité composées et comme encapsulées dans une méthode finale contextuelle : l'un des rôles du filtre hiérarchique, type d'arbre de décision contextuel et pondéré utilisé sur des bases de connaissances qui sont la cible même de notre analyse et harmonisation.

A un premier degré d'utilisation, l'AMR permet de traiter l'information qui entre et sort des bases, c'est un outil pour donner des résultats. Mais on peut préciser que si un système expert est constitué de bases de connaissances, de faits, et de programmes, à un second degré d'utilisation, l'AMR devient un outil de régulation du système cognitif même.

5.1) Graphe d'appel de LAC et enchaînement des mécanismes de l'AMR

A partir de la méthodologie d'AMR, un ensemble de programmes a été mis en place, regroupés sous le nom LAC (Logiciel Automatique de Croisement).

Le logiciel LAC est un outil proche du système expert, doté d'une interface permettant :

- De restructurer le fichier originel d'export contenant les lignes de navigation sismique par SP, et calculer les critères de comparaison complémentaires (SP/CDP, longueur linéaire, centroïde, densité de SP par ligne pour le moment. D'autres critères peuvent être ajoutés).
- De regrouper sous forme de clusters les lignes de navigation selon les critères et seuils de résolution sélectionnés.
- De mesurer la similarité entre les lignes de navigation sismiques, et de mesurer la qualité de la classification

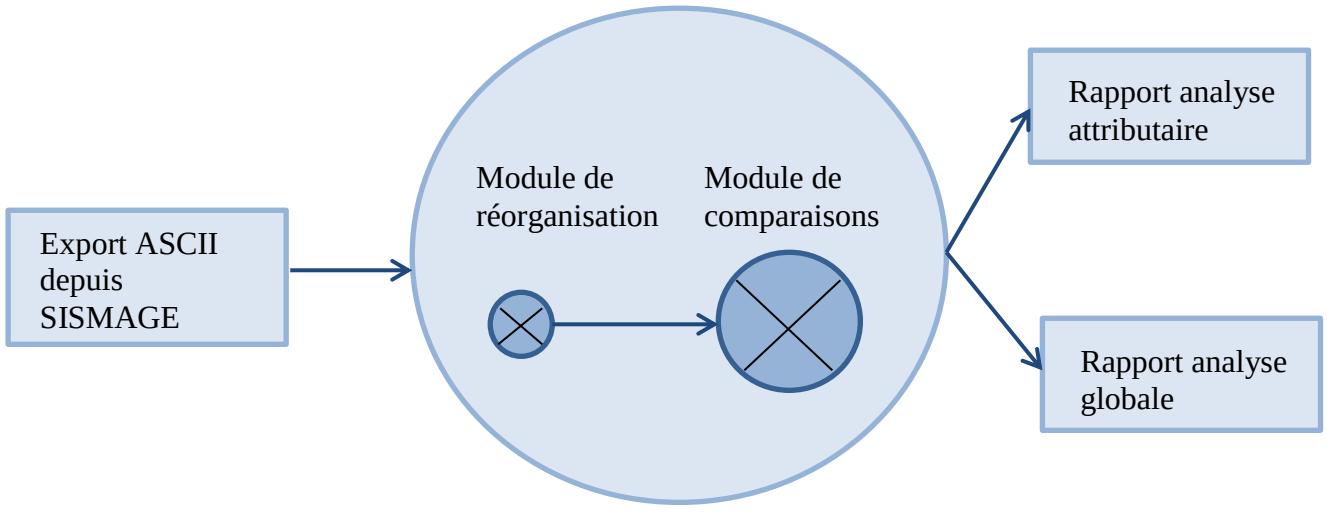


Figure 36 : Principe du workflow – Etape 1, un outil de diagnostic

LAC permet, en outre, de corriger les données à partir de l'interface, c'est-à-dire qu'il peut générer sur demande un fichier dans le format de l'export initial ne contenant que les lignes de navigation que l'utilisateur a choisi de garder. Ce fichier peut être réimporté dans la base de données.

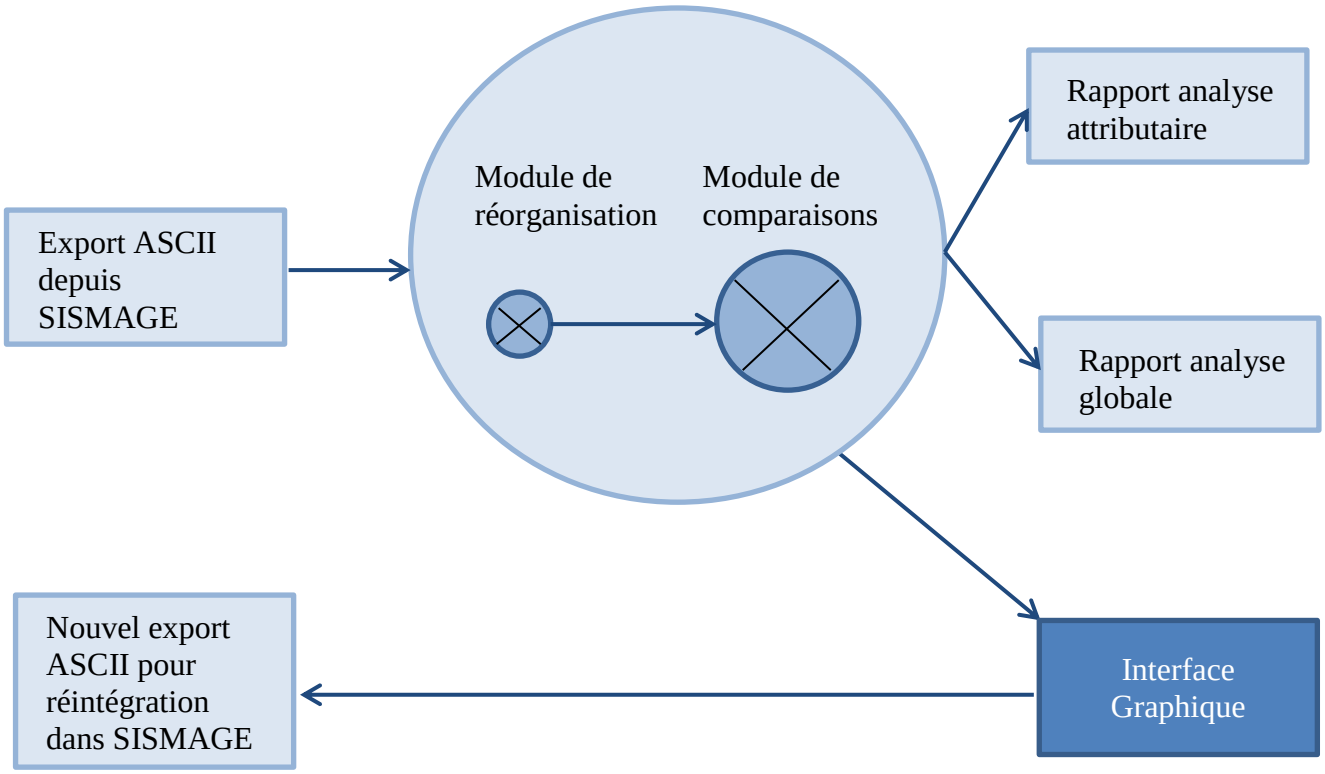


Figure 37 : Principe du workflow – Etape 2, un outil de correction

Le module de comparaisons contient quatre sous-modules, pour la gestion des mécanismes de lecture-écriture des fichiers, pour les mécanismes de l'arbre de décision utilisé dans le cadre du filtrage à tamis, pour la gestion des métriques de similarité, et enfin pour les mécanismes de classification automatique.

Le moteur de classification va centraliser tous les autres sous-modules par un systèmes d'engrenages où la classification appelle le sous-module de filtrage, en lui envoyant les seuils de tolérance, qui appellera quant à lui les métriques de similarité.

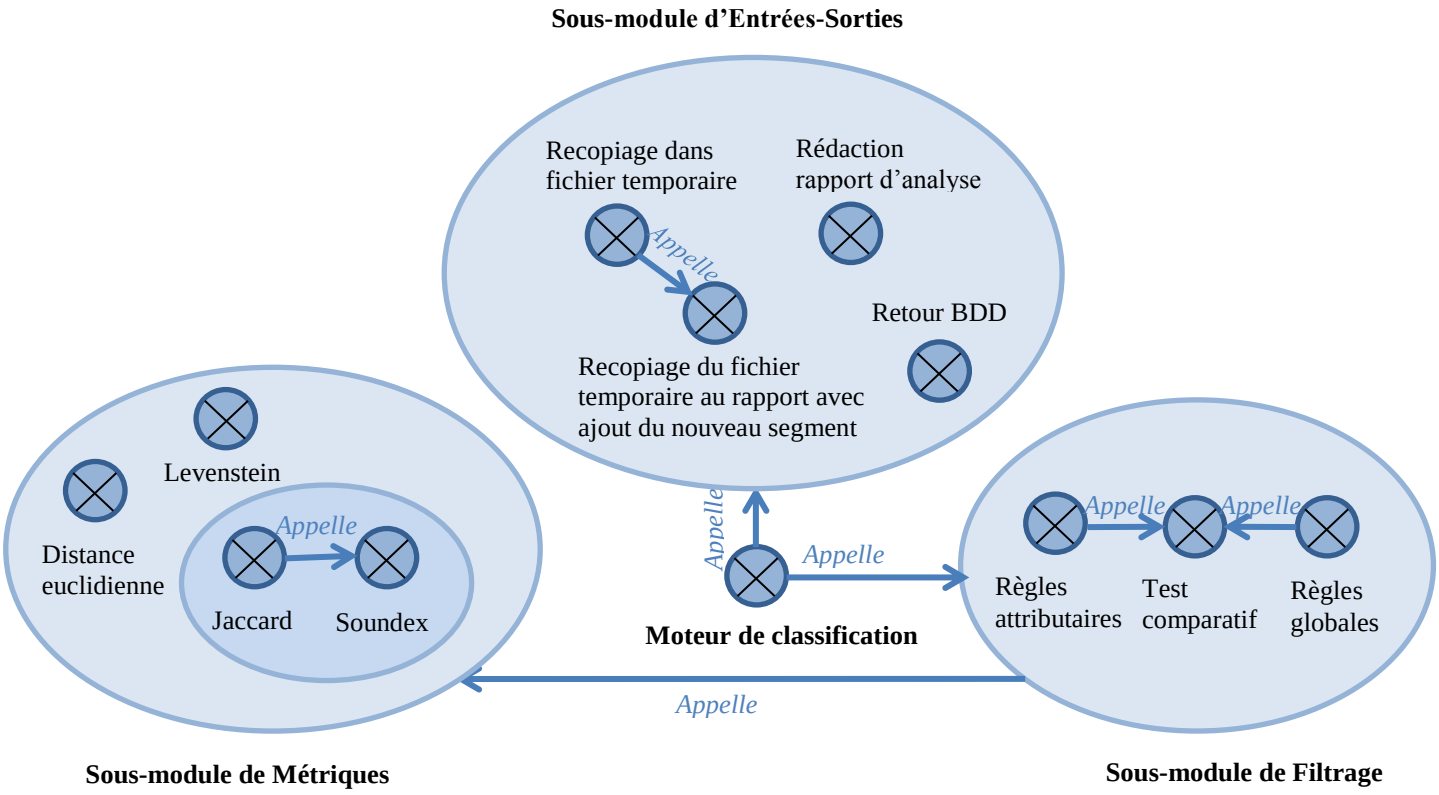


Figure 38 : Schéma d'articulation des modules Réorganisation et Comparaisons de LAC

Chaque métrique de similarité correspond à un type d'attribut. La sélection de la bonne métrique pour le bon attribut est réalisée dans le sous-module de filtrage par les règles de comparaison attributaires. Ensuite les règles de comparaison globales permettront de réaliser une comparaison prenant en compte la combinaison attributaire. Ensuite, le test final de filtrage permet de passer dans un arbre de décision seuillé par la famille de seuils de tolérance passée en argument : on obtient la mesure de similarité finale. On obtient aussi le verdict sur la question de savoir si les données sont suffisamment similaires pour être considérées comme doublons pour la résolution à laquelle on les regarde.

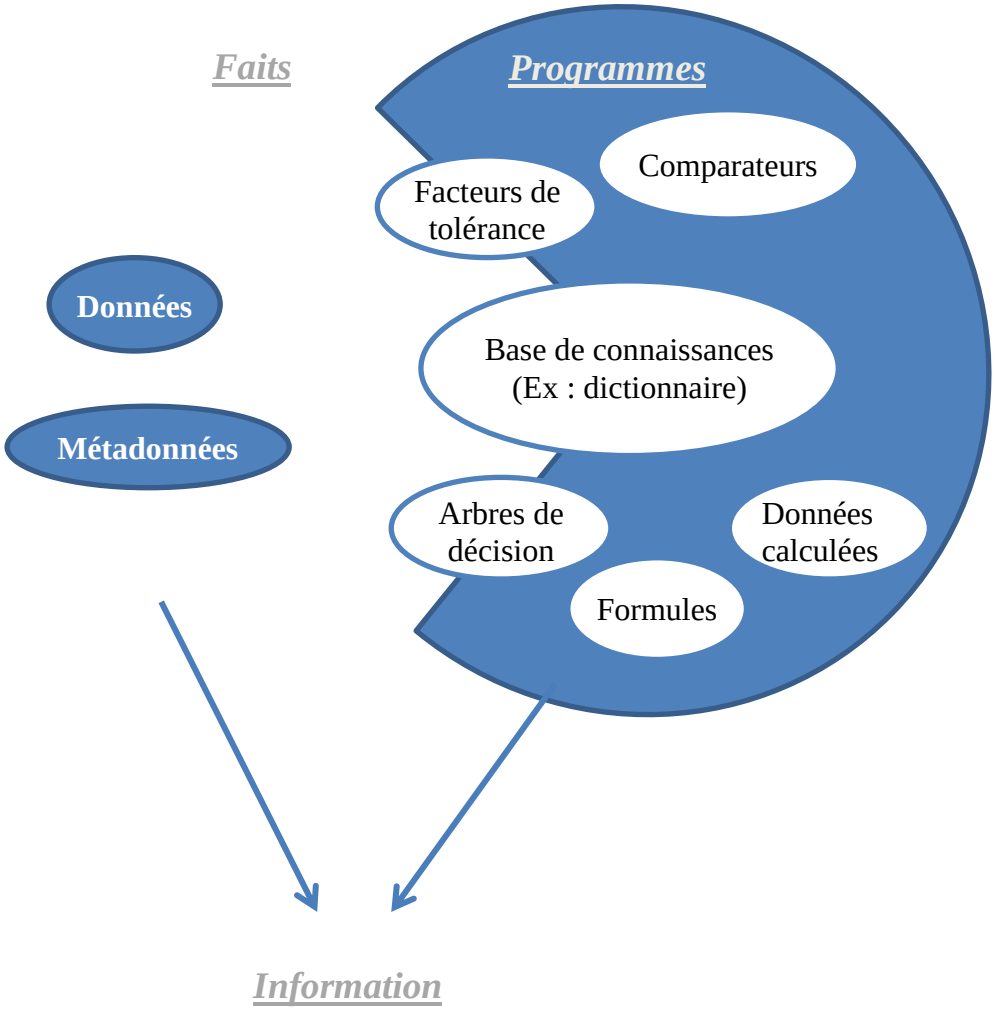
Comme expliqué dans les parties d'algorithmique, les optimisations faites pour segmenter le traitement du fichier initial fonctionnent avec un système à bascule entre le fichier de traitement et un fichier temporaire. La rédaction d'un rapport d'analyse à la fin de la classification permet d'indiquer les doublons identifiés pour la résolution souhaitée, la qualité globale de chaque groupe, les résidus (éléments n'appartenant à aucun groupe), ect.

5.2) L’apport des mécanismes d’intelligence artificielle au fonctionnement de LAC

Tout en tenant compte des contraintes liées aux moyens technologiques accessibles pour le déploiement dans le cadre industriel, l'AMR a été implantée au plus proche possible d'un système intelligent.

En effet, LAC contient des composantes allant d’une partie stable et à une partie variable avec quelques intermédiaires. La partie stable correspond à la bibliothèque de métriques de similarité, ou encore aux données portant des attributs enregistrés sur le terrain qui sont des faits invariants dans la base de données après enregistrement. Ensuite l’ensemble des éléments de modélisation et de formats que l’utilisateur met en place lors de la phase méthodologique constitue une partie stable mais paramétrable.

Il est possible de faire l’étude de l’impact de réorganisations hiérarchiques des critères de comparaison sur la classification de données obtenues, ou simplement de rendre un critère prépondérant par rapport à d’autres critères via le paramétrage par les facteurs de tolérance. Ces facteurs de tolérance rendent le système de filtrage entièrement flexible : c’est une partie variable.



Information	**Variable**	**Stable paramétrable**	**Stable**
Interne	Seuils de tolérance inférés	Stratégie de classification contextuelle	Règles de comparaison (comparateurs/métriques de similarité/formules)
Interface/Frontalière	Base de connaissances	Arbre de décision Métriques de similarité alimentées	Attributs calculés
Externe	Hiérarchie attributaire	Seuils de tolérance paramétrés dans l'IHM	Données, Métadonnées

Figure 39 : Structure de l'information depuis un point de vue système expert

On retrouve un certain nombre de ces éléments dans l'AMR et dans LAC dans la génération d'attributs de synthèse, ainsi que dans les mécanismes de classification permettant d'identifier des groupes ou des comportements des données dans les groupes ayant pu varier dans le temps, ou selon les technologies d'acquisition. Ces détection restent entièrement dépendantes des bases de connaissances, car de fortes lacunes dans le renseignement des dates par exemple ne permettrait pas des détections pertinentes des changements. Il est important de préciser qu'au cours de l'identification des groupes, les meilleurs représentants sont aussi calculés.

L'autonomie relative du système expert est aussi un critère important d'optimisation pour l'analyse de données volumineuses et complexes, d'où l'importance de pouvoir déterminer de manière automatique ou assistée les seuils de tolérance des analyses multicritères. Une solution d'ajustement de ces seuils sera donnée dans le chapitre des visualisations.

Les algorithmes de LAC sont bâtis au mieux pour pouvoir être parallélises et vectorisés, mais le cap n'est pas encore franchi en termes de programmation. Les mécanismes de ce type, qu'utilisent aussi les technologies Big Data en vogue aujourd'hui, permettraient de faire un pas décisif pour faire avec LAC de l'apprentissage automatique. Il est nécessaire d'ajouter à ce modèle une technologie de stockage : une base de données relationnelle, ou préférentiellement un système indexé de fichiers permettraient de constituer la mémoire persistante du système, donc l'acquisition d'expérience et la constitution de références. En effet, les lignes de navigation sismiques ayant une caractéristique géométrique de linéarité, la linéarité de mécanique des systèmes de fichiers indexés permettrait un stockage très adapté.

Comme on le verra dans le chapitre prochain, il sera aussi possible de dresser un outil de cartographie et d'analyse de bases de données, à condition d'avoir l'autorisation de se connecter aux bases et de développer les convertisseurs de formats pour en extraire des modèles attributaires des données.

Mécanismes IA	*Mécanismes LAC*
Données externes	Données provenant d'acquisitions sismiques
Données internes	Attributs de synthèse
Programmes statiques	Utilisation de LAC sur fichiers pour harmoniser
Données dynamiques	Utilisation de LAC tournant en permanence sur des bases de données pour en suivre l'évolution sous forme d'image
Aide à la prise de décision	Chiffrage de la similarité sur 100% comme indicateur de fiabilité des clusters pour décider de dédoublonner
Evaluation d'une donnée dans son contexte	Métriques de similarité contextuelle et arbres de décisions
Identifier des patterns	Classifications et calculs des meilleurs représentants

Tableau 9 : Les mécanismes liés aux pratiques d'intelligence artificielle dans LAC

Afin d'introduire certaines méthodes de mesure de la qualité des groupes formés, il est nécessaire de redéfinir de manière mathématique la notion de similarité (notée s) et la notion de dissimilartié (d) qui sont des métriques telles que, pour deux données i et j, et nb le nombre de groupes issus de la classification :

$$s(i,j) = s(j,i) \qquad d(i,j) = d(j,i)$$
$$s(i,i) \geq s(i,j) \qquad d(i,j) \geq 0$$
$$s(i,j) \geq 0 \qquad d(i,i) = 0$$

On définit alors $(gi)_{iC[\ 1\ ..\ nb]}$ les centres de gravités des données. On note $(Nbi)_{iC[\ 1\ ..\ nb]}$ leur pondération qui est le nombre d'éléments de chaque groupe. Ceux-ci sont calculés comme étant la donnée représentant le mieux le groupe. Lorsqu'on travaille avec des données uniquement quantitatives, il est facile de faire une moyenne pour définir un centre de gravité. Mais lorsqu'il s'agit de données portant des attributs aussi hétérogènes que les lignes de navigation sismique, trouver un représentant juste devient plus difficile.
En reprenant la classification attributaire, on choisira alors pour l'AMR une manière de définir un représentant par groupe d'attributs. Nous avons défini un filtre sémantique, un filtre géométrique et un filtre d'acquisition.

Pour le filtre géométrique, on fera une moyenne classique de chaque attribut pour l'ensemble des points d'un groupe. Les attributs sémantiques et d'acquisition quant à eux

seront choisis comme étant ceux appartenant à la donnée qui est la plus proche en termes de similarité de toutes les autres données du groupe.

Chaque centre de gravité est pondéré par l'inverse du nombre d'éléments que contient le groupe qu'il représente.
On appelle g le centre de gravité de l'ensemble des données.

Ensuite on définit l'inertie intra-groupe $(Ii)_{iC[\ 1\ ..\ nb]}$, l'inertie intra-groupe globale (*Iin*) et l'inertie inter-groupes (*Iex*) ainsi :

$$pour\ i\ dans\ [1\ ..nb], Ii = \sum s(gi, g)$$
$$Iin = \sum Nbi * Ii$$
$$Iex = \sum Nbi * s(gi, g)$$

On appelle inertie globale $Ig = Iin + Iex$. Cette notion sera utilisée dans le dernier chapitre où est fait un travail sur les méthodes de visualisation, notamment combinées à des outils statistiques.

Dans le cas où l'on compare des classifications à nombre égal de classes sur un même jeu de données, alors la meilleure classification est celle ayant l'inertie intra-groupe la plus faible et l'inertie inter-groupes la plus forte. En d'autres termes, la meilleure classification est celle où l'attraction entre les données d'un même groupe, globalement pour tous les groupes est plus grande que l'attraction des groupes entre eux.
Peut-être peut-on à aller jusqu'à utiliser l'image de la pression : la pression au sein des groupes est suffisamment faible par rapport à la pression à l'extérieur des groupes, pour que les données restent agrégées.

Si l'on compare des classifications à nombre différent de classes, la classification présentant la valeur maximale de (Iex-Iin), donc le différentiel d'inertie (de pression ?) entre l'environnement extérieur et l'environnement intérieur des groupes est le plus grand, assurant la configuration la plus « solide » d'agrégation.

Ainsi, en calculant ces indicateurs sur des classifications répétées, avec variation des seuils de résolution, on pourra trouver la meilleure classification, mais aussi la meilleure combinaison de seuils de résolution pour comparer les données concernées. Plus cette

combinaison contiendra des seuils exigeants, et plus fine sera la résolution à laquelle on regardera les données, et plus fiable sera l'analyse.

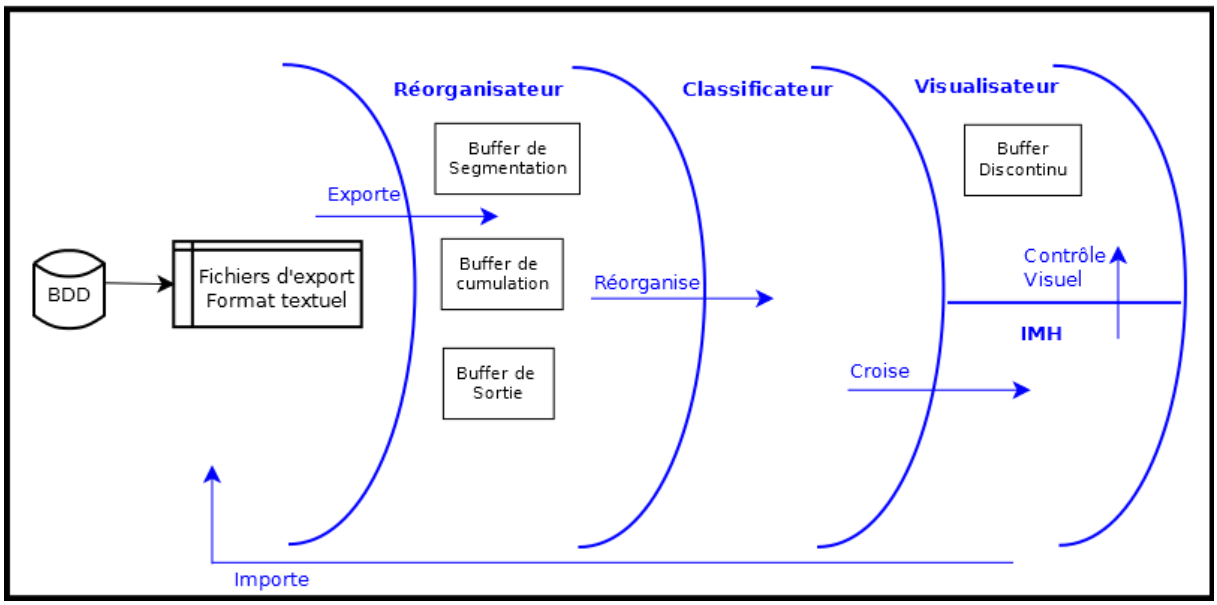


Figure 40 : Architecture de LAC par abstractions concentriques et processus engendrés

Les méthodes d'intelligence artificielle permettent de gagner un temps de traitement considérable du fait de l'automatisation du raisonnement pour la comparaison des profils de navigation sismique, et non pas uniquement du fait de l'automatisation des traitements servant au raisonnement.

LAC en tant que système expert pourrait a été très rapidement adapté à des données puits, il pourrait aussi être utilisé à des buts d'automatisation de la mise en œuvre de standardisation au niveau des identifiants des données, ou à la standardisation des noms, voire des fichiers des bases de données de TOTAL.

6) Eléments de visualisations

Le problème de la visualisation de données industrielles complexes à attributs hétérogènes et multiples ressemble au problème de visualisation des images multi-bandes en télédétection. On utilisera donc les méthodes statistiques et représentatives appliquées dans les travaux de télédétection comme l'analyse en composantes principales pour visualiser les données des bases industrielles. On comparera ces représentations à des représentations géographiques lorsque les données sont géo-référencées et à une méthode développée ici de visualisation par un algorithme de « similarité gravitationnelle ».

Une question récurrente dans les différentes méthodes de visualisation est celle de la représentation de données comportant non seulement des attributs quantitatifs, mais aussi des attributs qualitatifs tels que les noms ou textes.

De plus, comme expliqué dans les premiers chapitres de ce travail, les méthodes manuelles d'harmonisation de bases de données à partir de visualisation des données et de leurs attributs ne sont plus envisageables en termes de temps de traitement et de volume de données à traiter. L'objectif de ce chapitre est d'orienter la visualisation vers la mise en évidence de clusters, ou de configurations structurelles analysables des données dans les bases.

6.1) Visualisation géographique

Figure 41 : Représentation cartographique des données du Brésil sur lesquelles ont été faits les tests de performance de LAC vs ILX.

La figure 41 a été réalisée avec l'outil ArcGis.

Les lignes de navigation sismiques sont ici cartographiées en mettant en évidence leur position géographique et leur nom de campagne : couleur du point.
On remarque qu'il y a des recouvrements entre les différentes campagnes. Or, comme la géométrie d'acquisition sismique est déterminée en fonction de la carte de couverture souhaitée et en fonction des contraintes territoriales d'acquisition, on peut envisager qu'il y ait non seulement redondance d'information dans cette zone, mais aussi redondance en termes de navigation, donc de géométrie des lignes de navigation. Cependant, comme il s'agit d'acquisition marine, en eau peu profonde, la probabilité de redondance géométrique est moins grande que s'il s'agissait d'acquisition terrestre, où les contraintes territoriales sont plus fortes (zones habitées, routes, pipelines etc.).

On insiste sur le fait que cette représentation géographique permet une première analyse globale des données permettant de supposer que la probabilité qu'il y ait de la redondance est non négligeable, mais qu'elle ne permet pas l'identification de doublons.

6.2) Visualisation par analyse en composantes principales (ACP)

L'analyse en composantes principales est une méthode d'analyse de données visant à déterminer la contribution de chaque attribut à la caractérisation de l'ensemble des données.

On peut alors définir les deux axes de la représentation plane la plus complète possible des données multi-attributaires. Il ne s'agit pas de projection stricto sensu, mais plutôt de synthèse d'information. Cette représentation, comme nous le verrons, ne donne pas satisfaction pour la visualisation des groupes de données similaires pour des données volumineuses. Cependant, cette méthode mathématique va nous permettre de valider l'hypothèse méthodologique de l'AMR vue au premier chapitre : la classification attributaire hiérarchisée.

Par un souci de temps, ce travail n'a pas été implémenté dans LAC (il pourrait l'être, notamment pour automatiser la phase de modélisation de l'AMR) : j'ai utilisé le logiciel gratuit d'analyse de données R pour réaliser les calculs statistiques et afficher les graphiques associés.

6.2.1) Analyse pour l'ensemble des attributs numériques

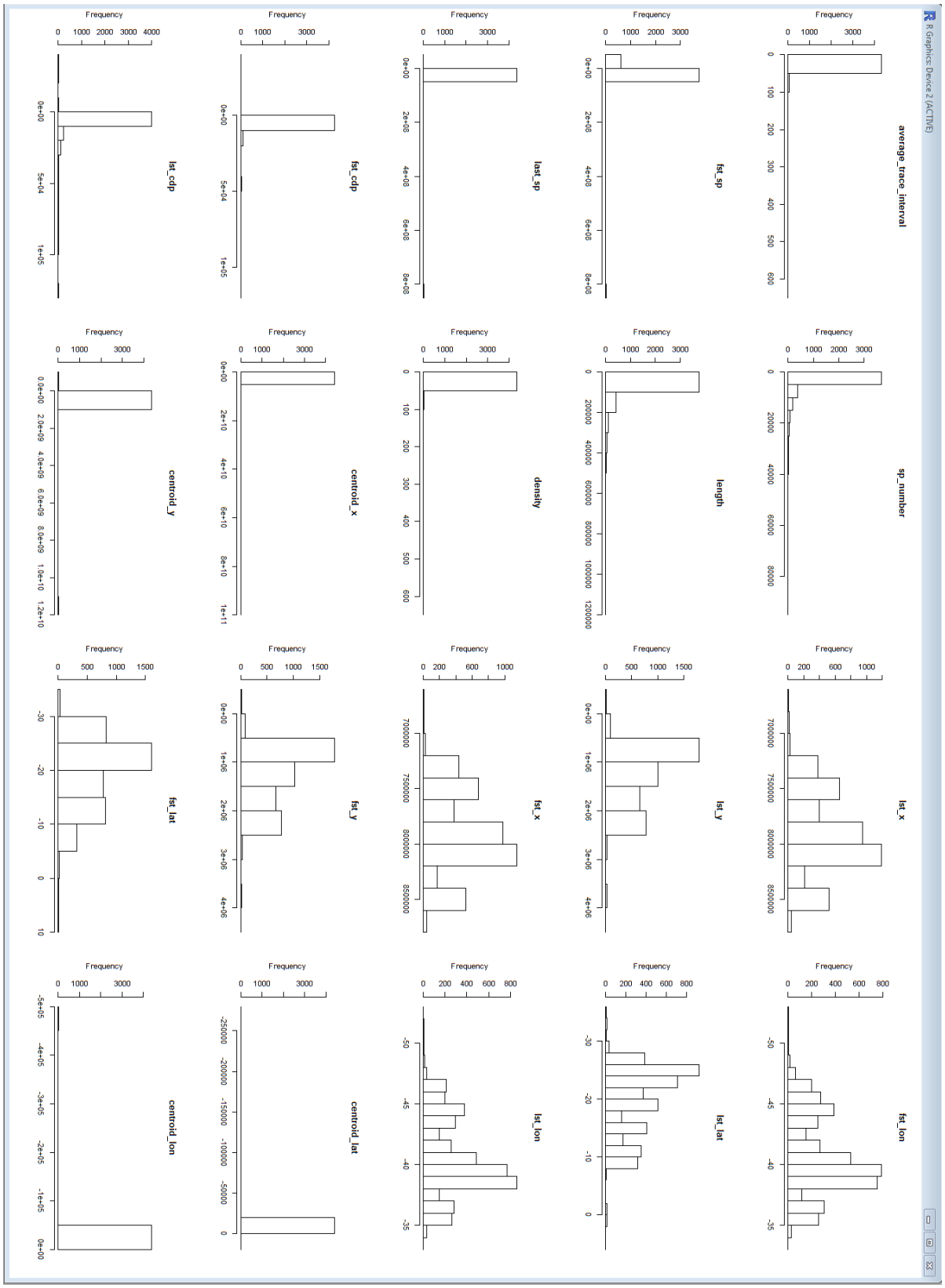


Figure 42: Histogramme des attributs numériques pour l'ensemble des données du Brésil

Les histogrammes des attributs numériques des données se présentent sous la forme de deux familles : les attributs à histogrammes resserrés et les attributs à histogrammes étalés. Les histogrammes étalés permettent une meilleure lisibilité de l'agencement des données. On pourrait envisager de faire un traitement supplémentaire pour étaler les histogrammes trop resserrés.

Cet ensemble d'histogrammes montre que sans autre traitement et analyse des attributs, les attributs géographiques tels que les longitudes et latitudes, les coordonnées cartésiennes permettent la meilleure différentiation des données, suivis par la longueur linéaire et l'attribut d'acquisition qu'est le SP number.

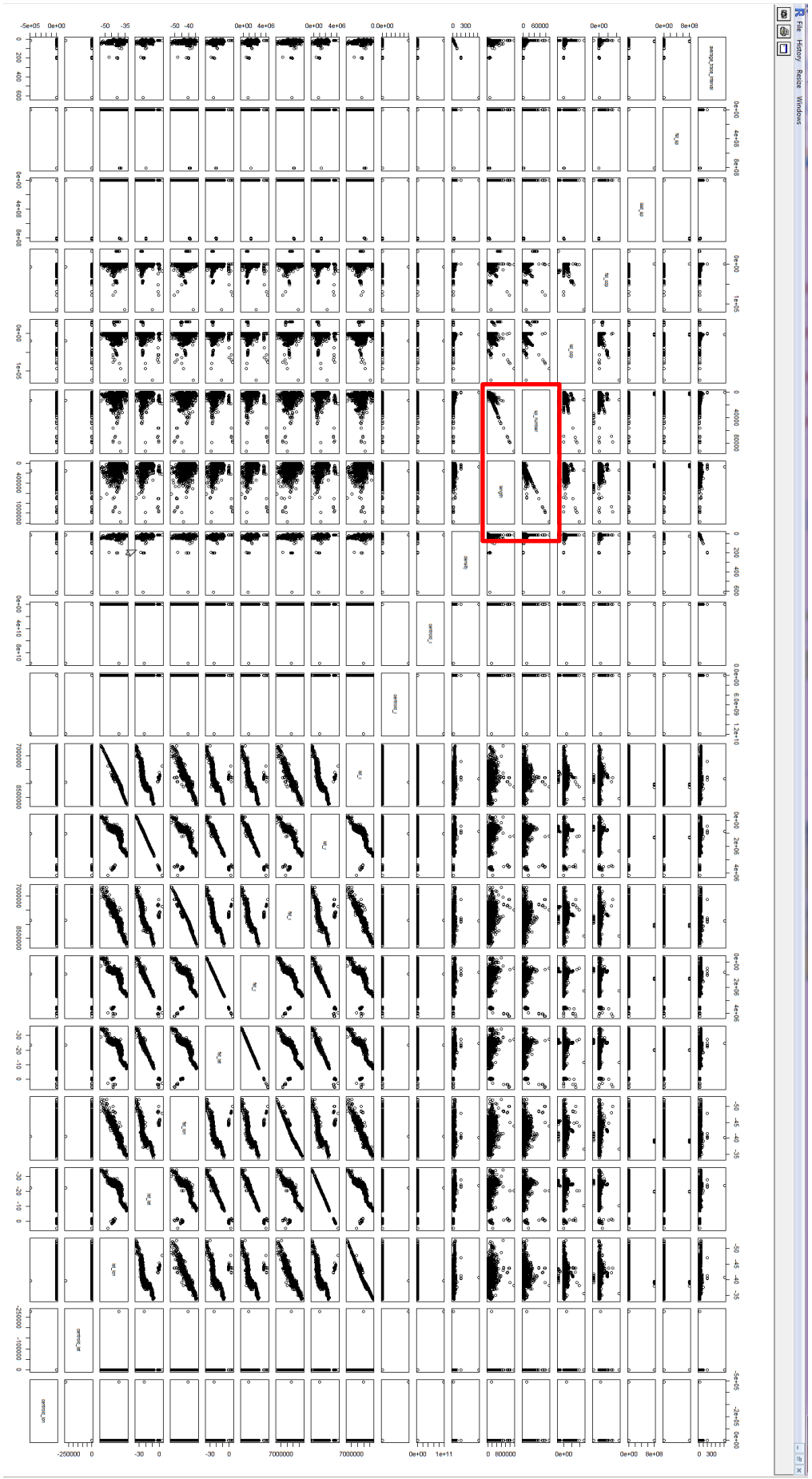


Figure 43 : Relation des attributs numériques entre eux (corrélation attributaire deux à deux), pour l'ensemble des données du Brésil

Les graphes de corrélation montrent que les coordonnées géographiques et les coordonnées cartésiennes sont liées car le semi de points de leur graphe de corrélation est proche de former une droite : en effet, les coordonnées cartésiennes sont issues de la transformation des coordonnées géographiques mesurées sur le terrain par GPS.

Les graphes de corrélation nous montrent aussi une tendance pouvant refléter un lien entre la longueur linéaire et le SP number (encadrement rouge sur la figure). Or, il est possible que le système numérotation en SP de chaque ligne (notamment d'une campagne à l'autre) varie, notamment concernant l'origine de la numérotation. Le système de numérotation serait alors assez caractéristique de la ligne. De même, la longueur linéaire d'une ligne, si elle est calculée avec une précision suffisante, peut aussi être très représentative d'une ligne. Ces deux attributs pourraient donc être liés par la relation logique qu'est l'appartenance à une même ligne. Dans le jeu de données du Brésil, il est probable que ces deux attributs soient calculés avec plus de précision que les autres attributs, excepté les coordonnées. Notons que ce sont des attributs dont l'un appartient au filtre géométrique et le second au filtre des paramètres d'acquisition. On pourrait en déduire une concordance ente les paramètres physiques (géométriques) et les paramètres déterminant les conditions d'acquisition telles que le système de numérotation. Il est donc pertinent d'inclure ces deux natures différentes d'attributs pour caractériser une ligne de navigation sismique.

Cela valide graphiquement l'une de nos hypothèses implicites de modélisation : la prise en compte de données d'acquisition et de données géométriques parmi l'ensemble des informations disponibles pour caractériser ces données-ci.

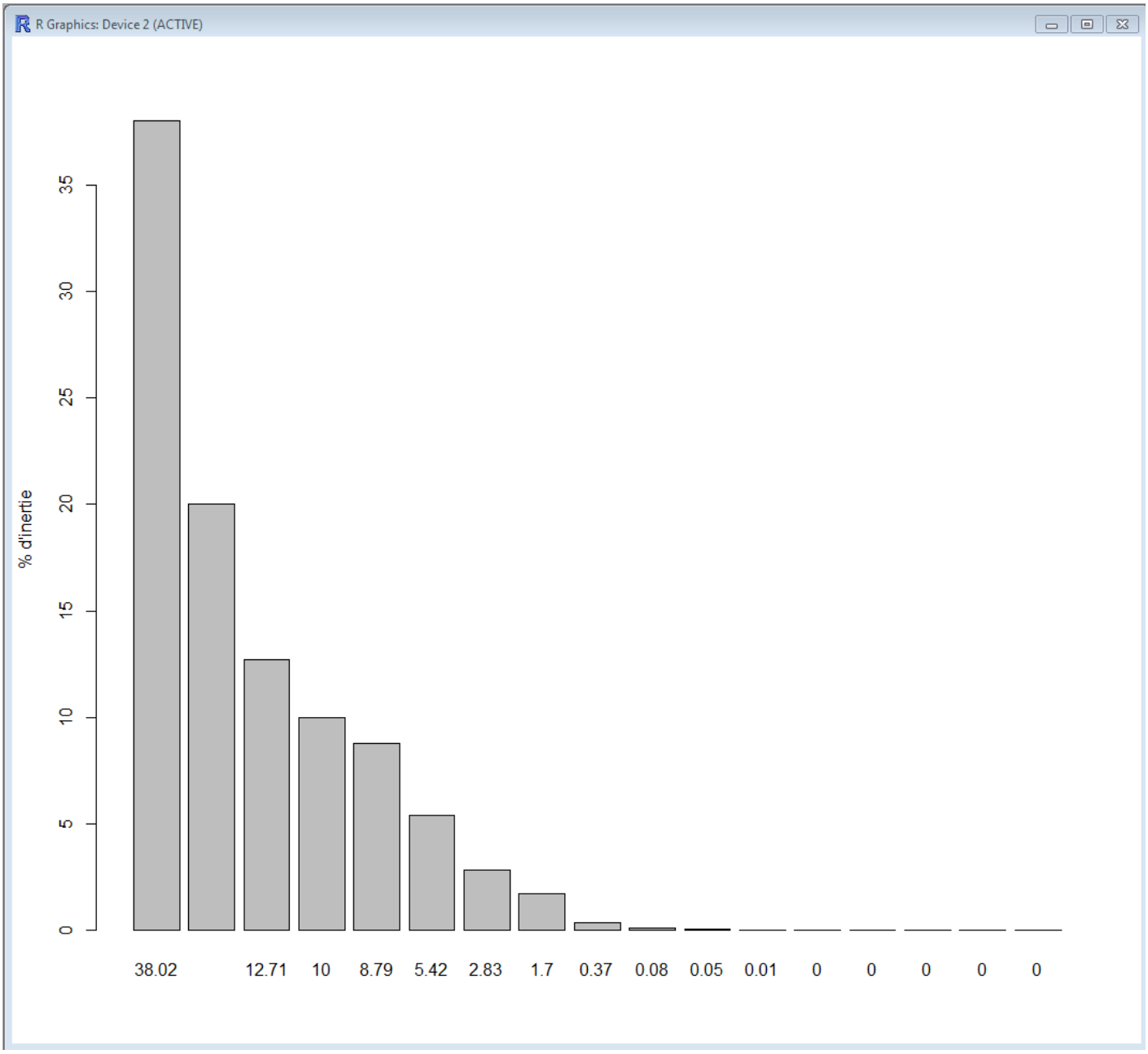


Figure 44 : Valeurs propres en pourcentage d'inertie

	Comp1	Comp2
average_trace_interval	0	3
fst_sp	1	0
last_sp	1	0
fst_cdp	48	4
lst_cdp	73	11
sp_number	108	9
length	156	6
density	0	3
centroid_x	0	2490
centroid_y	0	2490
lst_x	12343	1
lst_y	1196	0
fst_x	1250	0
fst_y	1189	0
fst_lat	1204	0
fst_lon	1177	1
lst_lat	1210	0
lst_lon	1152	1
centroid_lat	0	2490
centroid_lon	0	2490

Tableau 10 : Contribution de chaque attribut à la construction de deux axes de composantes principales

	Comp1	Filtre
lst_x	12343	Géométrie
fst_x	1250	
lst_lat	1210	
fst_lat	1204	
lst_y	1196	
fst_y	1189	
fst_lon	1177	
lst_lon	1152	
length	156	
sp_number	108	Acquisition
lst_cdp	73	
fst_cdp	48	

Tableau 11 : Contribution de chaque attribut au premier axe de représentation, tri en contribution croissante

	Comp2	Filtre
centroid_x	2490	Géométrie
centroid_y	2490	
centroid_lat	2490	
centroid_lon	2490	
lst_cdp	11	Acquisition
sp_number	9	
length	6	
fst_cdp	4	
average_trace_interval	3	
density	3	

Tableau 12 : Contribution de chaque attribut au second axe de représentation, tri en contribution croissante

6.2.2) Analyse pour l'ensemble des attributs numériques, en retirant les coordonnées cartésiennes

	Comp1	Comp2
average_trace_interval	24	1971
fst_sp	5	20
last_sp	5	20
fst_cdp	236	637
lst_cdp	415	1356
sp_number	583	1580
length	677	866
density	24	1972
fst_lat	1962	416
fst_lon	2120	212
lst_lat	1987	384
lst_lon	1959	408
centroid_lat	2	78
centroid_lon	2	79

Tableau 13 : Contribution de chaque attribut à la construction de deux axes de composantes principales

	Comp1	Filtre
fst_lon	2120	Géométrie
lst_lat	1987	
fst_lat	1962	
lst_lon	1959	
length	677	
sp_number	583	Acquisition
lst_cdp	415	
fst_cdp	236	
average_trace_interval	24	
density	24	
fst_sp	5	
last_sp	5	

Tableau 14 : Contribution de chaque attribut au premier axe de représentation, tri en contribution croissante

	Comp2	Filtre
density	1972	Acquisition
average_trace_interval	1971	
sp_number	1580	
lst_cdp	1356	
length	866	Géométrie
fst_cdp	637	
fst_lat	416	
lst_lon	408	
lst_lat	384	
fst_lon	212	
centroid_lon	79	
centroid_lat	78	

Tableau 15 : Contribution de chaque attribut au second axe de représentation, tri en contribution croissante

Par rapport à la modélisation attributaire, on remarque que le premier axe est fortement représenté par les attributs géométriques que sont les coordonnées des premiers et derniers SP.

Le second axe est fortement représenté par les attributs d'acquisition que sont l'intervalle inter-traces, la densité de SP par ligne, le nombre de SP par ligne, et le dernier CDP, puis en moindres proportions par le premier CDP, les premier et dernier numéros de SP.

Cependant il est important de noter que les attributs géométriques que sont le centroïde et la longueur linéaire ne s'avèrent pas très représentatifs des données. De plus, ils participent plus au second axe qu'au premier, caractérisé majoritairement par les attributs géométriques.

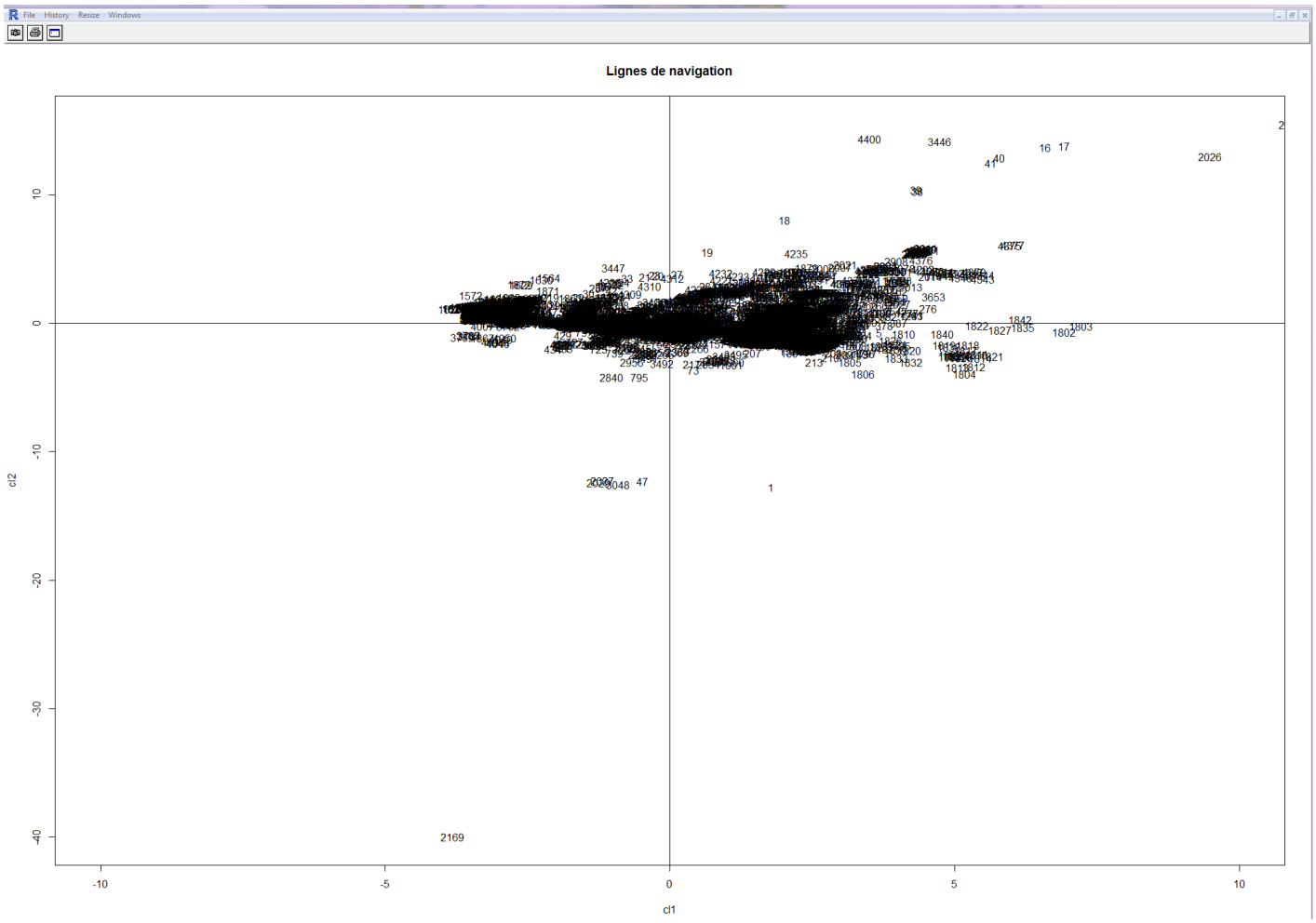


Figure 45 : Représentation complète des données du Brésil, sur les deux axes de synthèse d'information.

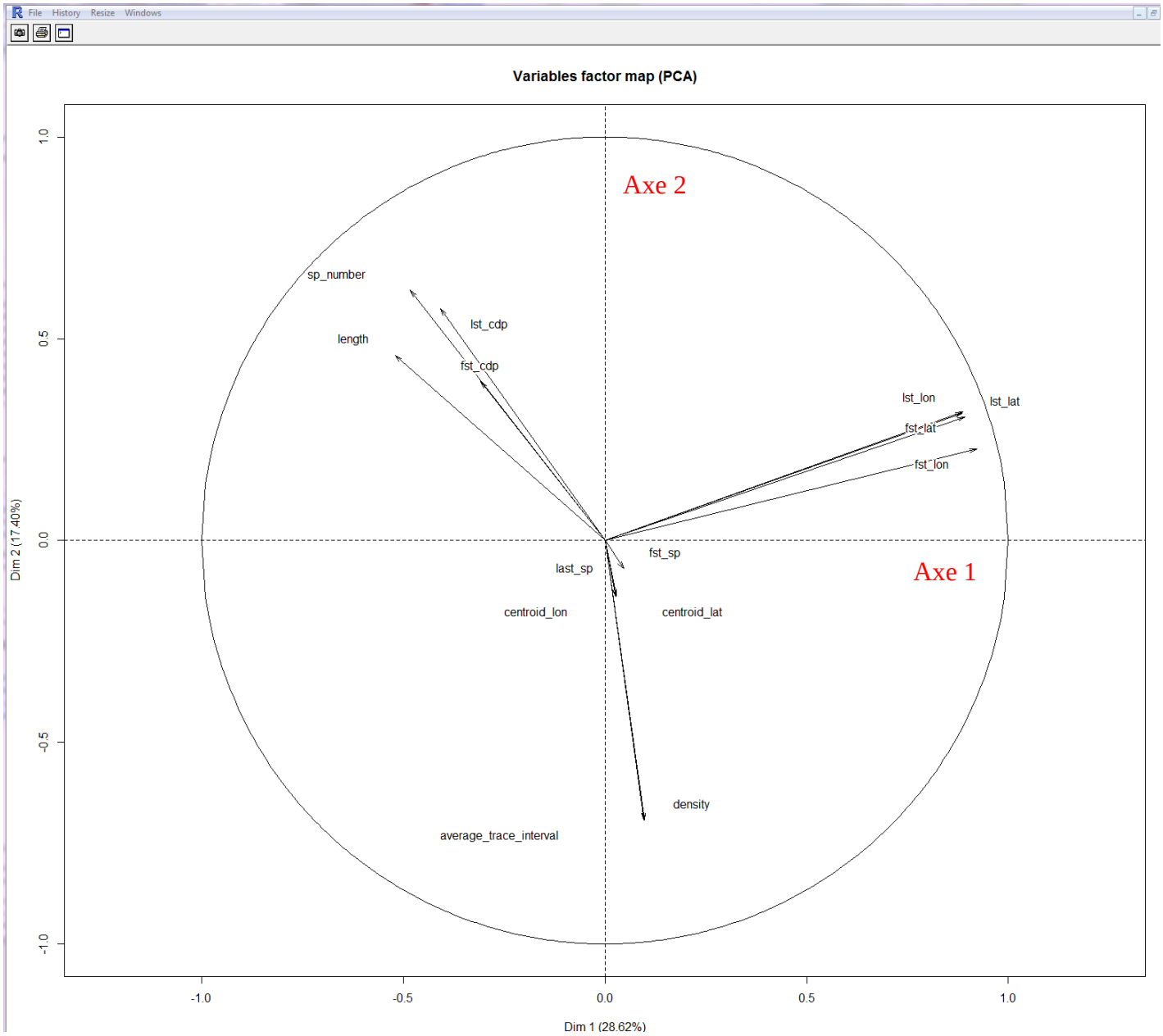


Figure 46 : Représentation des vecteurs attributaires constituant des deux axes principaux de représentation.

Une autre méthode statistique de visualisation telle que l'analyse factorielle de correspondance (AFC) peut être utilisée. Une AFC permet de voir les dépendances (ainsi que les indépendances et oppositions) entre les données, entre les attributs, et entre données et attributs. Cependant, après étude, elle aboutit dans notre cas aux mêmes conclusions que l'ACP : la méthode valide la hiérarchie attributaire mais la visualisation des données et attributs sur la carte ne permet aucune lisibilité car les données sont trop volumineuses et les attributs trop nombreux.

6.3) Visualisation « gravitationnelle » par mesures de similarité

Dans ce chapitre, on présentera l'apport d'une méthode nouvelle de représentation pour les bases de données, menant à l'analyse par imagerie.

6.3.1) Algorithme de visualisation en graphe éclaté – Etape 1

Pour cette représentation, on commence le travail à partir de l'ensemble des données en tant que semis d'objets géo-scientifiques auxquels des numéros de cluster ont été attribués via LAC. LAC nous a servi à segmenter les données avant de constituer l'image.

Pour la représentation des centres des clusters, on définit un repère polaire de centre G, centre de gravité du semi d'objets. Et de coordonnées (ρ, θ) où ρ sera la similarité en pourcentage entre le centre à représenter et G. On note θ la similarité entre le centre à représenter et le centre du cluster suivant à représenter.

On calcule les distances attributaires maximales entre les objets, notées :

$$(Smax_k)_{k\mathbb{C}[1..nbatt]}$$

On calcule (gi) les centres de gravités des clusters : ce seront les représentants des groupes.

Pour tout cluster i, pour tous objets j et j' du cluster i, et pour tout attribut k de cet objet, soit Similarité$\left(obj_{i,j,k} , obj_{i,j',k} \right) = S_{j,j'}^{i,k}$ la similarité entre l'objet j et l'objet j' pour l'attribut k.

Alors on note $\left(S_{j,j'}^{i,k}\right)\% = \frac{S_{j,j'}^{i,k} * 100}{Smax_k}$ cette même mesure de similarité en pourcentage.

On obtient par somme, la similarité sur tous les attributs entre l'objet j et l'objet j' :

$$S_{i,j,j'} = \frac{1}{nbatt} \sum_{k=1}^{nbatt} \left(S_{j,j'}^{i,k}\right)\%$$

Remarque : pour les attributs textuels, la similarité est mesurée entre 0 et 1, 1 étant la valeur de similarité maximale, alors que pour les attributs numériques, la similarité maximale est identifiée par la mesure la plus petite. Par exemple les objets les géographiquement les plus proches ont la distance euclidienne le plus petite. Il est donc d'homogénéiser cette forme de convention d'abord en passant en pourcentages, ensuite en prenant le complémentaire de la mesure, donc (100-mesure%) pour les attributs textuels. Ainsi, plus la mesure de similarité textuelle est grande, et plus notre coordonnée sera petite. Par conséquent, les données proches en similarité seront aussi figurativement proches sur le graphique.

On note alors $I_i = \frac{1}{(nbobj_i^2 - nbobj_i)} \sum_{j=1}^{nbobj_i} \sum_{j'=1}^{nbobj_i} S_{i,j,j'}$ l'inertie de similarité du centre de gravité du cluster i. Il s'agit de la moyenne des mesures de similarité des objets qui le composent mesurés deux à deux.

Les lignes de navigation sismiques sont constituées d'attributs qualitatifs et d'attributs quantitatifs. Les deux attributs qualitatifs sont les attributs textuels : nom de ligne et nom de campagne.

Pour les attributs quantitatifs, on calcule l'ensemble des similarités moyennes attributaires servant à définir les attributs des barycentres.
L'opérateur somme doit se comporter différemment pour la définition de ces nouveaux objets que sont les centres de gravité selon qu'on traite des attributs qualitatifs ou des attributs quantitatifs (noté en rouge plus bas).

Pour les attributs qualitatifs, au sein du même cluster, le nom « moyen » sera défini comme étant le premier nom d'objet ayant la similarité la plus grande avec le plus grand nombre d'objets du même cluster. Pour les attributs quantitatifs, l'opérateur de l'addition est classique, euclidien, et est appliqué directement sur les valeurs attributaires.

De même que pour les centres de gravité des clusters, on calcule G le centre de gravité de l'ensemble des données. La notion de centre de gravité correspond au calcul de moyennes attributaires pondérées, tandis que la notion de centre des clusters désigne le meilleur des représentants des groupes. La différence entre les deux se formule en termes de représentativité. En effet, pour les attributs numériques, comme les coordonnées, la moyenne ou la médiane sont de bons indicateurs statistiques pour caractériser des groupes. Plus particulièrement dans notre cas, la moyenne reflète plus le nombre d'individus du groupe, qui est une information importante lorsqu'on travaille avec de gros volumes de données. Par contre, on ne peut pas calculer de moyenne ou de médiane sur les attributs textuels. Il est donc nécessaire d'utiliser la métrique de similarité pour élire les attributs textuels les plus proches de tous les autres du groupe comme représentants. Pour ces raisons, on considère que le centre de gravité de l'ensemble des données et le centre de gravité des centres des clusters sont approximativement confondus dans les résultats de nos calculs qui aboutissent à un représentant abstrait : ce n'est pas forcément une donnée initialement enregistrée dans la base (c'est le cas pour les clusters singletons par exemple). En termes d'intelligence artificielle, ces représentants seraient considérés comme des données internes générées par le système.

Alors $G = (\frac{1}{(nbclust)} \sum_{j=1}^{nbclust} g_{i,k})_{k\mathbb{C}[1..nbatt]}$

Nous aurons alors son inertie de similarité $I = \frac{1}{nbclust} \sum_{j=1}^{nbclust} I_i$

On construit le tableau SIF, reliant le sommet initial et le sommet final de chaque arête du graphe, en rattachant tous les points au centre de gravité du cluster, puis tous les centres de gravité de cluster au centre de gravité global G.

On calcule les coordonnées cartésiennes (x, y) de tous les objets à partir des coordonnées polaires (ρ, θ).

Ainsi, pour tout cluster i, les coordonnées d'un objet du cluster sont :

$\rho_i = Similarité(g_i, obj)\%$

$\theta_i = \frac{Similarité(obj_i, obj_{i+1})\% * 2\pi}{100}$

Alors les coordonnées cartésiennes, dans le repère cartésien d'origine g_i de l'objet sont :

$x_i = \rho_i \cos\theta_i$

$y_i = \rho_i \sin\theta_i$

Et les coordonnées polaires des centres de gravité, en prenant G comme origine sont :

$\rho_i = Similarité(g_i, G)\%$

$\theta_i = \frac{Similarité(g_i, g_{i+1})\% * 2\pi}{100}$

D'où les coordonnées cartésiennes de centres suivant les mêmes formules :

$x_i = \rho_i \cos\theta_i$

$y_i = \rho_i \sin\theta_i$

On affiche le semis de points tel quel.

Puis on affiche le graphe décrit par SIF.

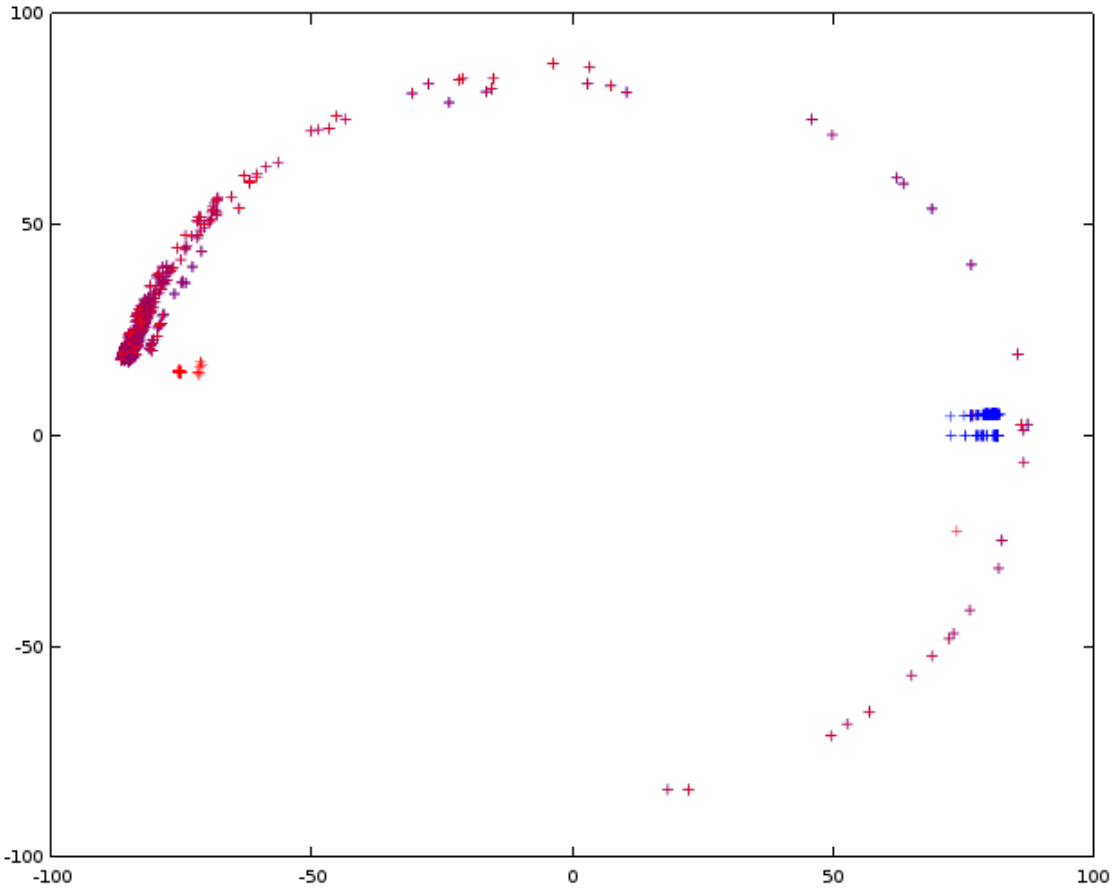


Figure 47 : Représentation graphique du graphe éclaté du jeu de données du Brésil. En bleu les données et en rouge les représentants de clusters.

Sur la figure 47, le système de coordonnées est polaire, calculé selon la mesure de ressemblance. La représentation graphique temporaire a été faite sur Octave et uniquement sur 2000 points.

On constate, d'une part, de nombreux points où les symboles rouges et les symboles bleus se superposent. Ce sont les données uniques. D'autre part, on aperçoit des points bleus assez clairement à droite du cercle qui ne se confondent pas avec leur représentant : ce sont des points appartenant à un même cluster.

En termes algorithmiques, on effectue ces calculs à partir de la construction des tables ci-dessous. Ces tables peuvent être utilisées comme des tables de calcul à la volée ou bien elles peuvent être stockées si le volume de données nécessite de réaliser des pré-calculs.

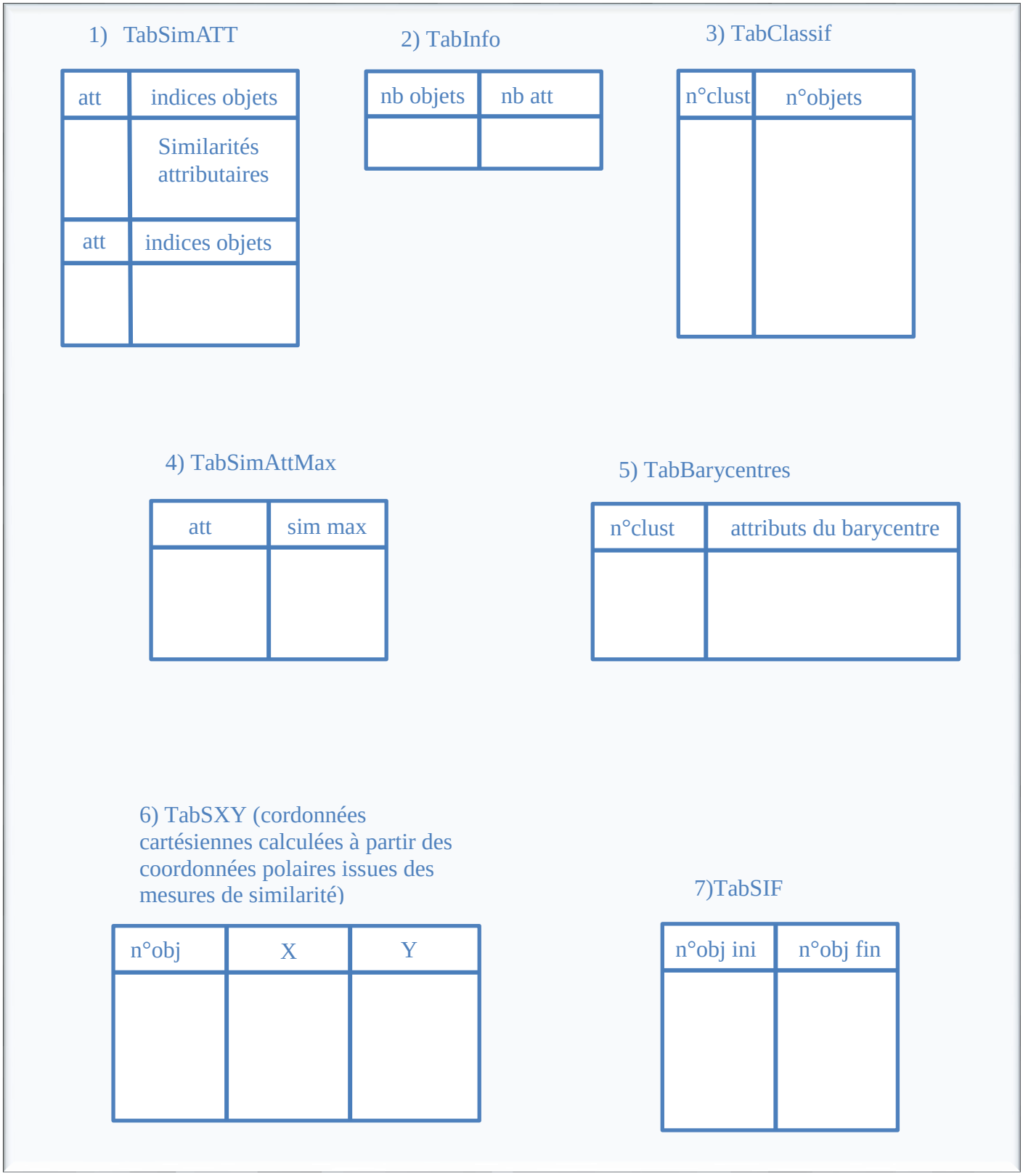


Figure 48 : Tables utilisées dans les algorithmes de calcul du graphe éclaté.

6.3.2) Algorithme de visualisation en image colorée – Etape 2

On effectue une triangulation de Delaunay, et récupère tous les triangles décrits par les numéros des points qui les composent. Ensuite, on calcule la charge de similarité de chaque triangle : il s'agit de la similarité entre les trois points qui le composent : moyenne des similarités entre les points comparés deux à deux. Réciproquement, chaque point porte comme charge de similarité la moyenne des charges de similarité des triangles auxquels il appartient.

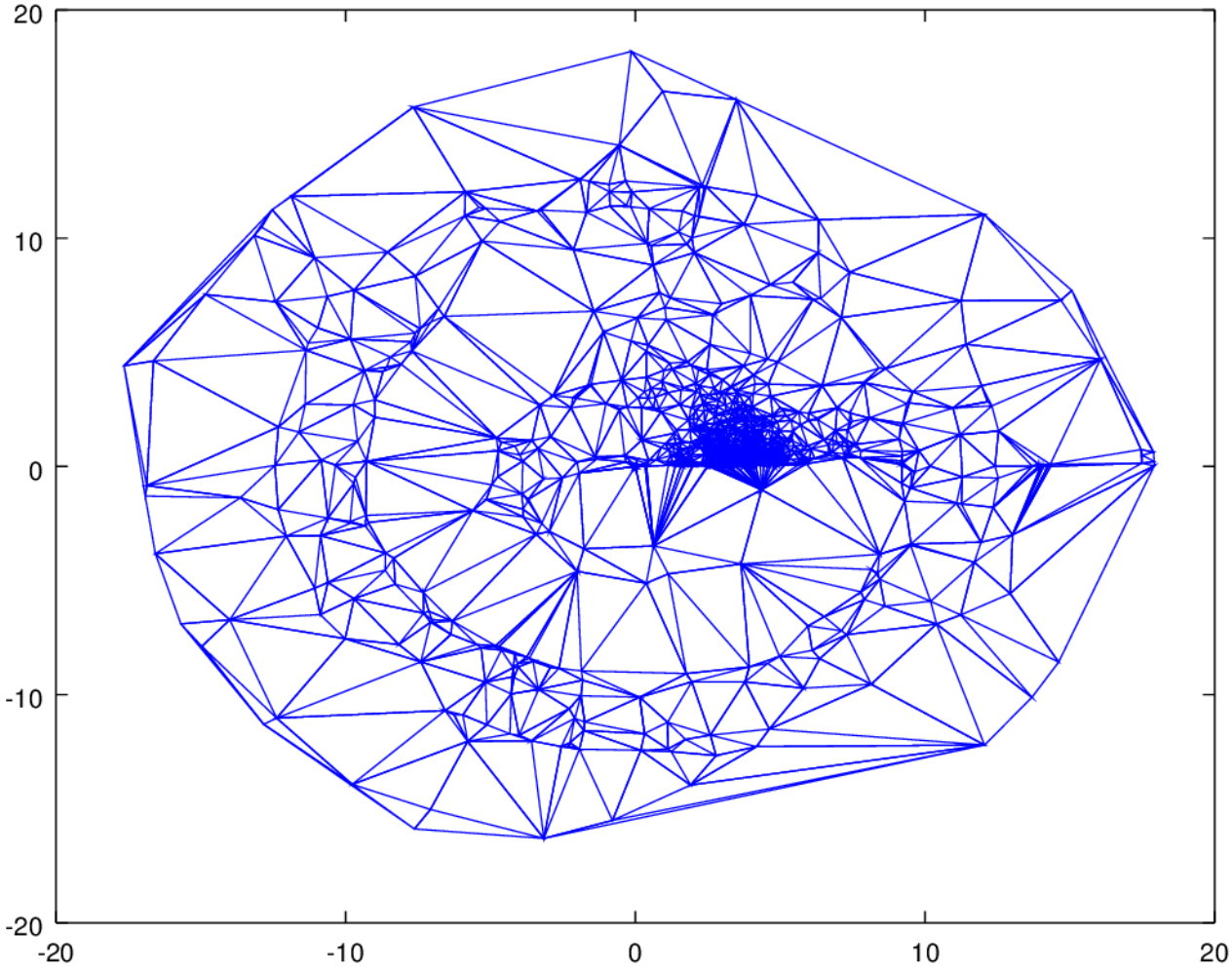


Figure 49 : Triangulation de Delaunay de l'ensemble des données du Brésil en incluant les représentants et à échelle réduite sur 20 unités représentées via leurs coordonnées polaires basées sur la mesure de similarité.

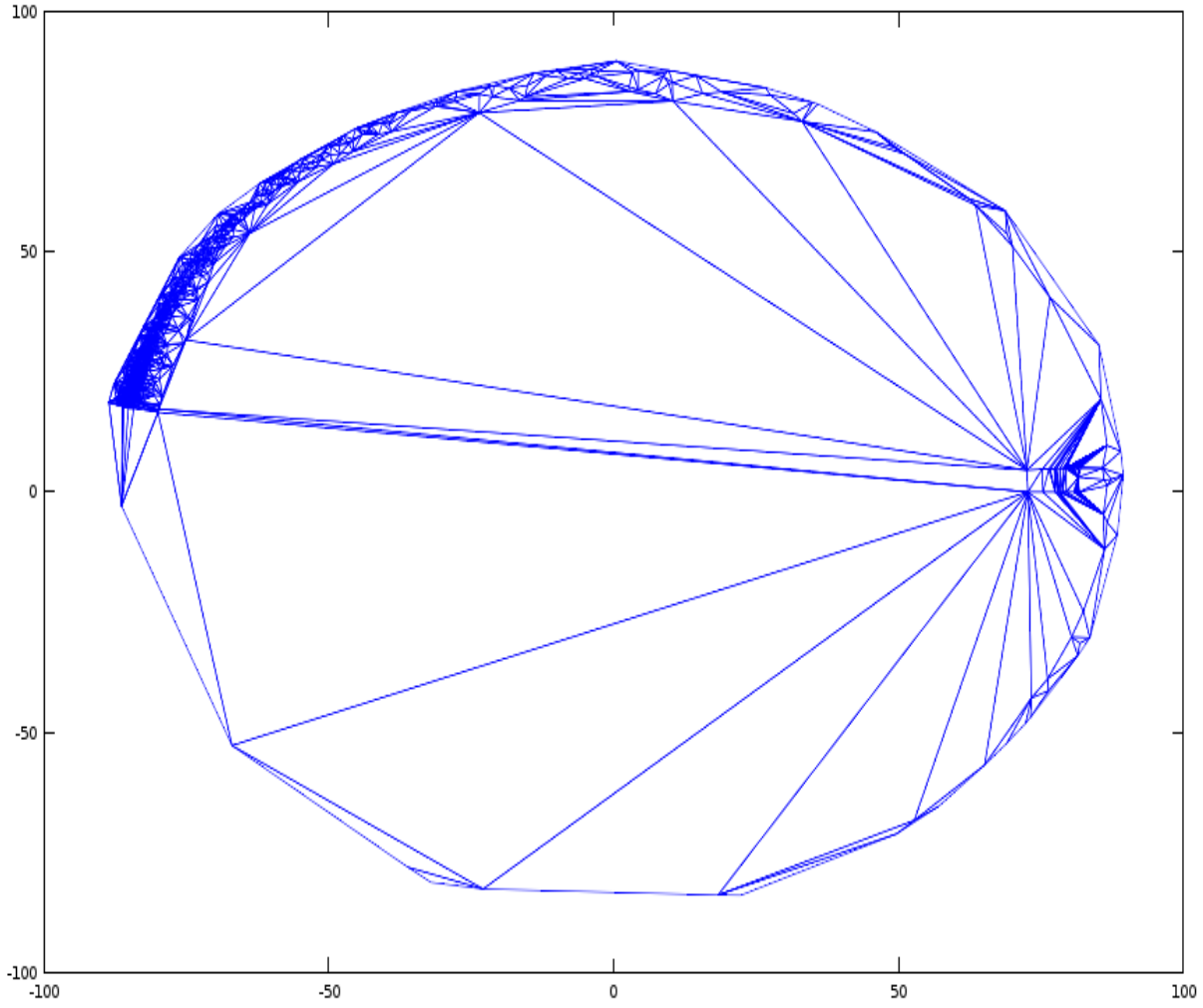


Figure 50 : Triangulation de Delaunay de l'ensemble des données du Brésil sans les représentants, sur une échelle de 100%, représentées via leurs coordonnées polaires basées sur la mesure de similarité.

Pour une meilleure compréhension, on peut imaginer la charge de similarité comme une couleur allant du blanc au rouge (voire rouge très sombre virant au noir) en passant par le vert, et plus elle va vers le rouge-noir, plus la similarité dans le voisinage du point est forte, donc plus la probabilité de trouver des clusters à forte inertie (potentiellement des doublons de données) interne est grande.

Dans l'étape suivante, on construit une matrice pixélisée, en définissant une grille sur l'espace de représentation des points et des triangles. Cette grille a pour origine (x0,y0) le point en bas à gauche du semi de points. La résolution de l'image, définie par la taille de la grille est arbitraire. On donne un nombre de pixels qui servira à définir le pas en abscisses dx et en ordonnées dy de cette grille.

Pour finir, on rattache chaque point à une maille de la grille, et la maille portera comme couleur la moyenne des charges de similarité des points. Les mailles deviennent alors des pixels colorés d'une couleur calculée entre 0 et 1 ou entre 0 et 255 pour le domaine RVB.

On peut constater qu'il était possible de colorer directement les triangles, mais cela privait de la notion de résolution, et liait les points entre eux tandis que l'approche pixellisée n'est plus conditionnée par les liens faits dans la triangulation. De plus, la lisibilité est plus grande avec une image plus habituelle, avec un repère orthogonal.

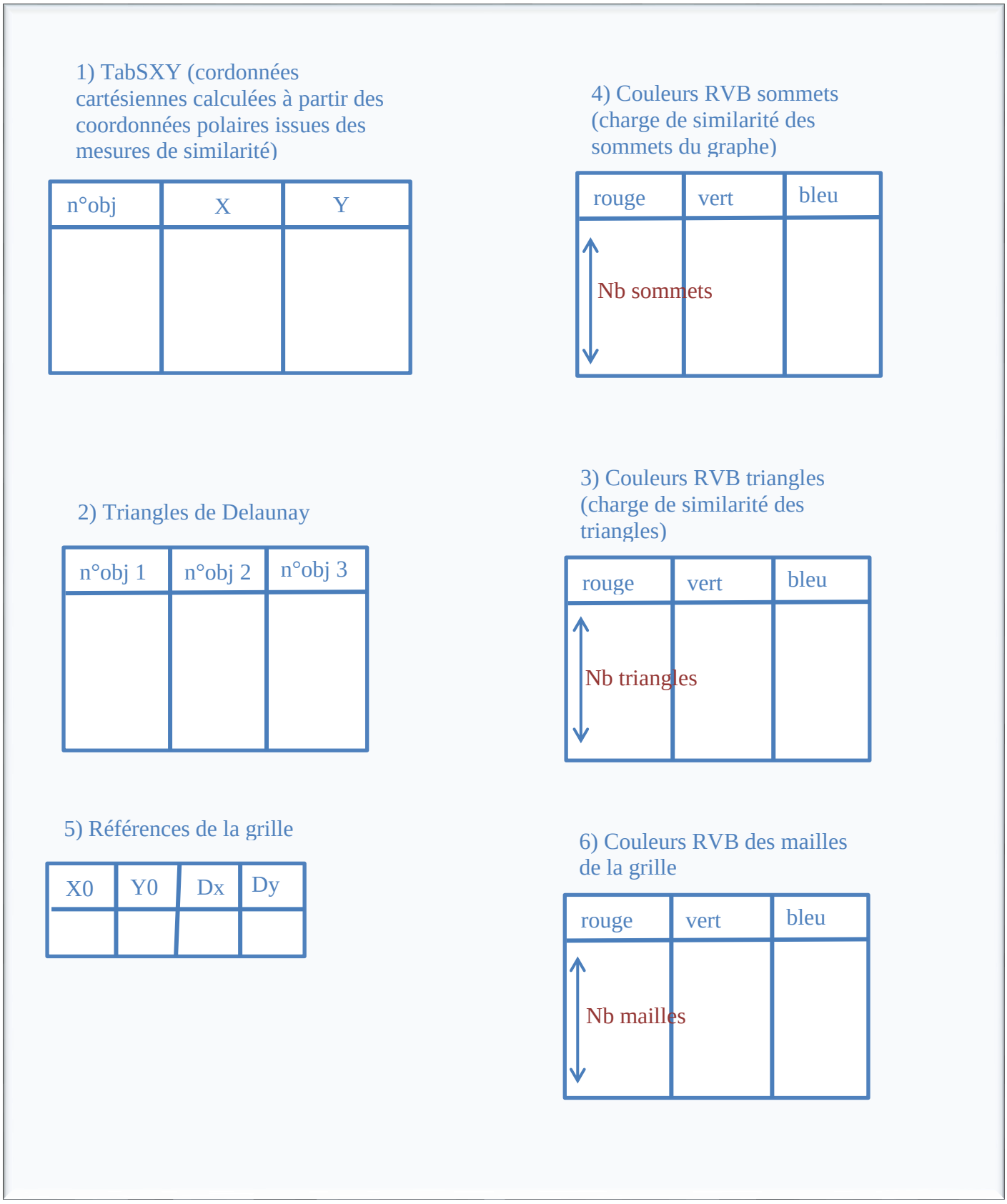


Figure 51 : Tables utilisées dans les algorithmes de calcul de l'image colorée.

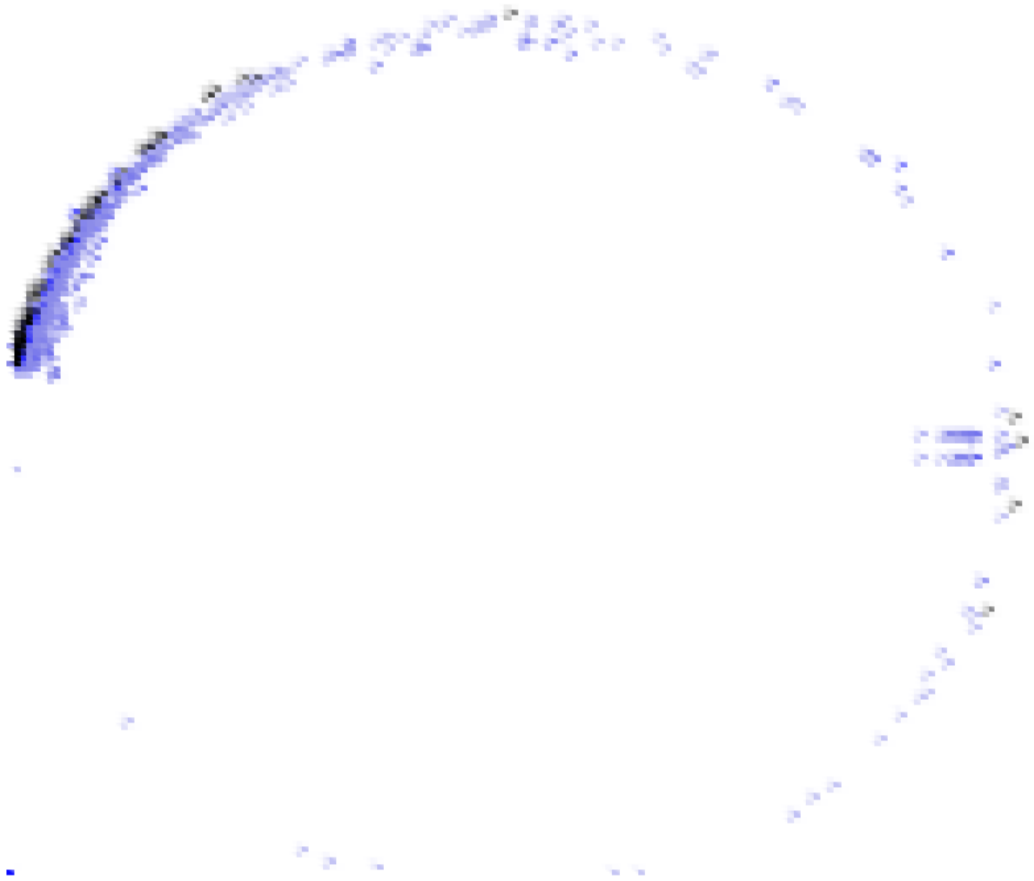


Figure 52 : Image 1000*1000 pixels de du jeu de données du Brésil sans inclure les représentants, et à échelle sur 100 unités.

Dans la figure 52 ci-dessus, la couleur du pixel indique la ressemblance des données qu'il contient. Les taches noires à gauche témoignent de fortes zones de duplications. On peut, de plus, remarquer une forte dispersion de la base des données du Brésil, en effet, le centre du cercle, représentant le barycentre de toutes les données, se trouve relativement éloigné des données. On repère deux pôles dans ces données : l'un étendu sur la partie haute à gauche du cercle, et l'autre sur la partie droite du cercle. Il est intéressant de voir que l'échelle de similarité choisie (20 unités ou 100 unités) ici va plus ou moins bien témoigner de la dispersion

de la base. Sur 20 l'image est comme contractée, et sur 100 elle est plus fidèle à la géométrie interne de la base de données.

On peut réaliser des images à résolutions différentes, donc à nombre et taille de pixels différents. Les couleurs varient alors en fonction de la charge de similarité des points contenus dans chaque maille. Les couleurs des mailles ne contenant pas de points sont interpolées. Plus la couleur tend vers le blanc et moins on a d'information.

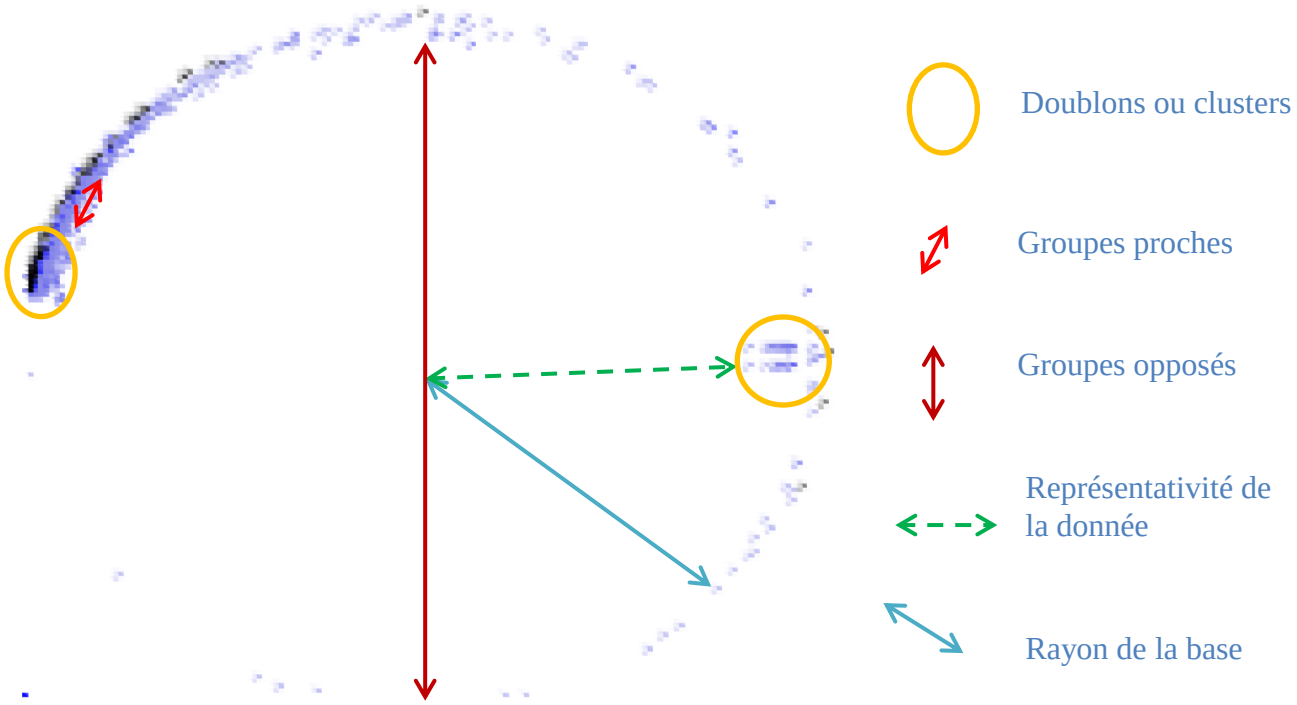


Figure 53 : Quelques informations à tirer de l'image d'une base de données.

Cette image présente tout d'abord un rayon indiquant l'homogénéité des informations dans la base. En effet, le rayon est défini comme l'éloignement en termes de similairité du le centre de l'image qui est l'emplacement du représentant abstrait global des données, à chacun aux points les plus en marge. Ainsi, lorsqu'on compare différentes bases ou jeux de données, on peut avoir une idée de la représentativité de leur centre et de leur homogénéité. De plus, lorsuq'on envisage de factoriser de l'information entre différentes bases de données, on peut

mesurer la similarité entre leurs centres. Plus ils sont proches et plus il y aura de chances que les bases contiennent des informations redondantes, donc factorisables.

On peut noter l'intérêt d'automatiser une telle gestion des connaissances dans un système expert afin de réaliser des catégories logiques, par exemple, sans parler du faible coût calculatoire dans la comparaison des représentants globaux des bases. De plus, l'ingestion de données nouvelles et leur classement dans les catégories déjà existantes peut s'effectuer de manière optimale en comparant les nouvelles données non pas à toutes celles de la base mais uniquement aux représentants des différents clusters. Cela peut même s'opérer de manière échelonnée via l'utilisation d'hyper-clusters. Notons également que lorsque de nouvelles données arrivent et qu'elles ne ressmeblent suffisamment à aucune autre, elles vont former une nouvelle catégorie, voire plusieurs. Cela permettrait donc l'identification d'un nouveau comportement, mécanisme précieux à l'intelligence artificielle.

Dans cette image ressort également la représentativité des données par rapport à l'ensemble de la base. En effet, plus une donnée est proche du centre de la base, et plus elle est représentative.

Bien entendu, la couleur sombre de certains pixels indique la présence de doublons : l'information est nécessaire pour l'harmonisation des donnés. Il serait alors intéressant pour d'éventuelles implantations logicielles de rendre l'image interactive afin de pouvoir cliquer sur un pixel pour en voir le détail des données, voire directement y valider la suppression des données redondantes.

Les deux dernières informations mises en évidence dans l'image sont les aspects d'attraction et de repulsion des données ou des groupes de données. Les données diamétralement opposées sont également les plus éloignées en termes de ressemblance. Inversement, les groupes de données contingentes ont une similarité inter-clusters plus grande.

Pour finir, l'image finale est construite en s'appuyant sur la segmentation faite par LAC. On reprend cependant les mesures de similarité brutes pour colorer les pixels. Cela permet de compenser le biais entre la segmentation issue des seuils de tolérance paramétrés et la distance de similarité « complète » renseignée par la couleur du pixel.

On pourra même afficher sur cette image le contour des groupes formés par la segmentation et le contour des classes issues d'une détection automatique de contours pour voir le biais entre les deux.

Cluster number	survey_name	line_name
70	BRO	BRO_28rl151
70	BRO	00BRO_28RL151
70	BRO	BRO_2-8RL151
70	BRO	BRO_28RL151
70	BRO	BRO28RL151
70	BRO	BRO_28RL15100

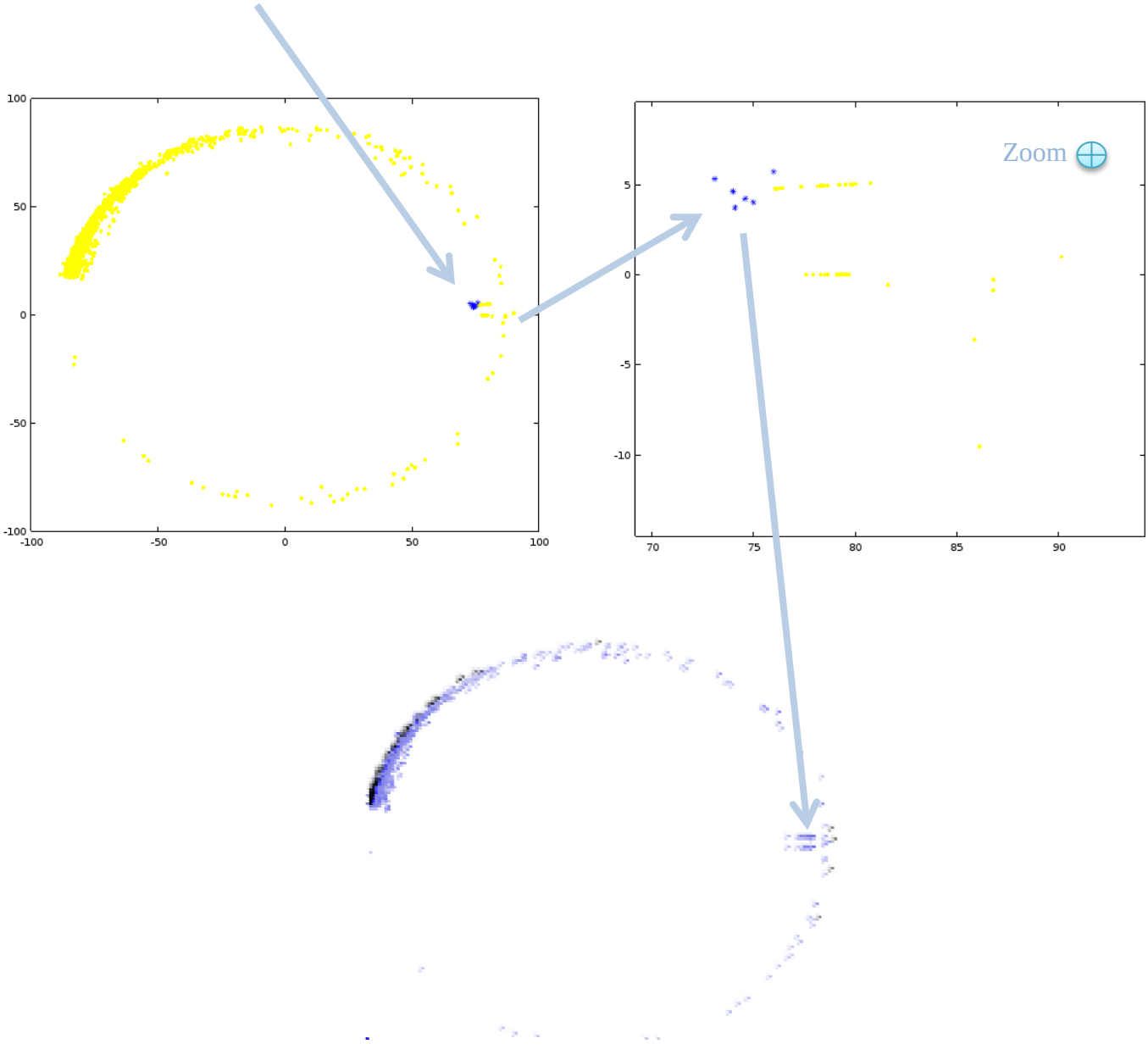


Figure 54 : Exemple de suivi de données depuis le fichier d'entrée jusque dans l'image. Sur le graphe éclaté, en bleu on marque les données très similaires, appartement ici au même cluster. En jaune les autres données.

Comme on peut le voir sur la figure 54, il est possible d'avoir un suivi direct entre les données en entrée et leur représentation dans l'espace de similarité. L'exemple pris dans cette figure est simple : nous avons introduit des doublons artificiels ayant juste des changements de noms de lignes, et quelques variations dans les intervalles moyens entre traces. Ensuite, nous suivons d'une part l'évolution de ces données dans la classification. Elles sont classées dans le même groupe. Ensuite, sur le graphe éclaté elles sont très proches, et on pourra finir par remarquer que dans l'image pixellisée, les pixels correspondant à ces données sont foncés, indiquant une forte similarité. Pour l'expérience, on a fait un traçage de données qu'on savait similaires. Dans le cadre de l'utilisation de l'imagerie, le processus sera inversé. On partira des pixels les plus foncés pour aboutir aux données redondantes.

Remarquons que dans le cas de la réconciliation entre les sources, ou pour des problématiques d'achats, les données que l'on souhaiterait analyser sont non pas les clusters, mais les données uniques. Ce sont ces données que l'on n'a pas en base de référence, ou que l'on ne peut réconcilier entre deux bases. Pour les analyser, on prendra le « négatif » de l'image, en renversant le code couleur : les données les plus uniques seront les plus foncées.

Conclusion

On se situe dans un environnement technologique où Big Data, Data Mining, Master Data Management prennent de l'essor et où le hardware ainsi que les réseaux internet locaux ou globaux donnent une ampleur économiquement fondamentale à l'information. L'imminence de l'Internet Of Things en est une illustration, comme on le sait. Cependant, les technologies logicielles aujourd'hui sur le marché sont lourdes à mettre en place, et demandent de réorganiser les infrastructures de gestion et d'accès à l'information, elles nécessitent aussi maints investissements en formations et consultations. Si les algorithmes, les concepts et les mécanismes relatifs aux systèmes experts sont apparus il y a une quarantaine d'années, l'intérêt non pas futuriste mais pragmatique des industries ne se manifeste que de nos jours. En effet, depuis les robots spéculateurs, jusqu'au web profiling commercial en passant par l'industrie d'exploitation des ressources naturelles, des bases de données s'enrichissent, parfois de manière chaotique, en même temps que l'intérêt de les exploiter oriente les prises de décisions.

Le problème de l'harmonisation des données industrielles soulève des questions concernant l'identification de redondances dans de volumineuses bases de données complexes. Il a d'abord été nécessaire de chercher s'il existait des mesures de similarité adaptées à ce type de données portant des attributs hétérogènes. Dans les travaux de recherche, l'aspect de complexité n'était présent que dans les approches de similarité contextuelle. Cependant aucune de ces métriques ne traitait le cas des données géophysiques. Il a donc été nécessaire de développer une métrique de similarité contextuelle spécifique pour le cas des objets géo-scientifiques industriels. Il s'agit d'un système de filtrage à tamis conçu grâce à une hiérarchie attributaire imbriquée dans un arbre de décision. Ce système de filtrage fait appel à des métriques élémentaires adaptées à chaque type d'attribut, pour optimiser le rapport entre l'exhaustivité et la précision des mesures, en particulier pour les attributs textuels qui sont plus difficiles à traiter que les attributs numériques. L'avantage de cette nouvelle approche est de fournir une mesure de similarité à plusieurs échelles : pour les attributs, pour les objets, et pour les ensembles d'objets.

L'identification des redondances demandait aussi de brasser de gros volumes de données tandis que le procédé existant dans les équipes de Data Management chez TOTAL était semi-manuel. Cela devenait trop lent face à l'évolution des besoins et prenait trop de

temps aux géophysiciens dont les compétences pouvaient être mieux valorisées sur d'autres activités. Ainsi a-t-on introduit des méthodes de classification mettant automatiquement en route les mécanismes de mesure de similarité par le système de filtrage. Le temps de travail sur un jeu de données comme celui du Brésil avec environ 4000 lignes de navigation sismique et 1380000 points de tir est passé de trois semaines à deux jours pour une personne. Sur cette même étude, les tests de performance comparatifs avec le logiciel de qualité de données InnerLogix de Schlumberger ont montré, pour des critères fixes de comparaison une précision et une rapidité légèrement supérieures dans le Logiciel Automatique de Comparaisons (LAC) spécialement développé pour le problème de l'harmonisation des données industrielles. La rapidité était en partie due aux algorithmes de synthèse des informations des lignes de navigation sismiques par le traitement segmenté de fichiers. Une collaboration avec les équipes de Schlumberger travaillant sur InnerLogix a abouti à la création d'une passerelle entre les deux outils afin de compléter certains aspects de chiffrement de la similarité dans InnerLogix. Depuis trois ans, LAC est industrialisé et déployé au service de Data Management chez TOTAL, mais n'a plus de maintenance technique depuis un an. Les nouvelles fonctionnalités d'imagerie de base de données qui ont été développées dans cette thèse n'y sont pas encore intégrées, mais devraient permettre une meilleure visualisation des phénomènes.

Bien entendu, c'est l'expert géophysicien, utilisateur de LAC, qui doit paramétrer les seuils de tolérance du système et en analyser les résultats. Les décisions finales concernant les données seront les siennes. Cela a motivé l'élaboration d'une méthodologie d'harmonisation par Automatisation des Mesures de Ressemblance (AMR) indiquant comment constituer une hiérarchie attributaire pouvant être validée à l'aide d'une analyse en composantes principales. Cette méthodologie est généralisable à d'autres types de données que les lignes de navigation sismiques. Par exemple, elle a été adaptée et utilisée pour la réconciliation de bases de documents ainsi que pour leur géo-référencement, dans le service de Documentation Technique chez TOTAL.

Après avoir identifié et regroupé des redondances, il est nécessaire d'aider l'expert géophysicien à trouver le meilleur représentant d'un groupe de données, notamment afin de savoir lesquelles garder et lesquelles supprimer. Pour cela, une méthode de calcul du meilleur représentant fondée sur la mesure de similarité par filtrage à tamis a été proposée. Elle permet de caractériser un groupe soit par un individu réel du groupe soit par un représentant abstrait du groupe. Cette forme de traitement de données à la fois internes et externes au système avec des

mécanismes de reconnaissance (mesures de ressemblance) oriente ce travail vers les principes d'intelligence artificielle et de systèmes experts. En effet, ce type de représentants permet de réaliser une gestion par catégories d'une base de connaissances. On s'approche notamment de la notion de moteur de recherche pour des données complexes industrielles.

La problématique de l'harmonisation des données est fortement liée à la qualité des bases de données. Il s'agit d'abord d'évaluer la situation d'une base statique, par exemple grâce aux nouvelles techniques d'imagerie proposées dans ce mémoire et fondées sur un espace de représentation plan de données pluridimensionnelles défini par la mesure de similarité décrite plus haut. Une image de base de données permet de détecter de manière visuelle certaines caractéristiques de la base, ainsi que des zones de doublons à harmoniser. Par ailleurs, on peut déjà comparer deux bases distinctes grâce à ces images. Ensuite, il serait envisageable de gérer les flux entrants et sortants des bases de données tout en surveillant l'évolution de leur qualité grâce à une analyse dynamique des images des bases. Pour l'avenir, il serait certainement rentable de développer davantage cette imagerie de bases de données et d'en tirer, par exemple, de nouveaux modes d'apprentissage automatique.

Développer des solutions synthétiques d'accès non pas à la donnée mais à la compréhension de celle-ci, voire à la compréhension des interactions entre ensembles est apparu comme un défi incontournable. Car sans ces solutions-là, on peut rapidement se perdre dans la multitude et la dispersion. C'est pourquoi ce doctorat a naturellement convergé vers la recherche d'un mode de représentation candidat à une vue de synthèse, sorte de carte de sens plus que de positionnement, proche de la famille des analyses factorielles, pouvant être branchée à des bases et peut-être à proposer comme un écran de pilotage. C'est dans cette direction, via le développement de l'imagerie pour les bases de données que j'aimerais poursuivre mes recherches.

Bibliographie

BANDYOPADHYAY Sanghamitra, SAHA Sriparna, Unsupervised Classification: Similarity measures, Classical and Metaheuristic Approches, and Applications, 2013, Springer

BEARD Matthew, *Expert Systems: An Introduction*, Kindle Edition, 2014

BERSON Alex, Lawrence DUBOV, Master Data Management and Customer Data Integration for a Global Enterprise, 2007, McGraw Hill

BILENKO, M. MOONEY, R. Adaptative Name Matching in Information Integration. *IEEE Intelligent Systems*, 2003, volume 18

BOUILLE Francois, « Un modèle universel de banque de données, simultanément partageable, portable et répartie », Thèse de doctorat ès Sciences Mathématiques-Informatique, Université P.M. Curie, 28 Avril 1977, 550p. (Jury : Arsac, Dahl, Delobel, Kitagawa, Robinet, Rocher, Vignes).

BOUILLE Francois, ASU Workshop on « SIMULA and Data Bases », Paris : « A new generation of data bases using the Hypergraph-Based Data Structure », 2-3 May 1977, in Proceed. P.123-151.

BOUILLE Francois, ASU Workshop on « SIMULA and Data Bases », Paris : « The paleontological collection data bank at the Université Pierre et Marie Curie », 2-3 May 1977, *in collab.*, in Proceed. P.153-173.

BOUILLE Francois, International Seminar on Intelligent Question-Answering and Data Bases Systems, Bonas (Gers) : « VLDB design with HBDS, EXEL and SIMULA 67 », 21-30 Juin, in Proceed. IRIA, p.88-100 .

BOUILLE Francois, « Cartographie thèmatique, Informatique, Applications » , Conférence invitée à la Sociedade de Geografia de Lisboa, 14 Juillet 1977, in Boletim da Sociedade de Geografia de Lisboa, série 96a, n°1-6, p. 5-54 .

BOUILLE Francois, 7ten Jahrestagung des Gesellschaft für Informatik, Nürnberg : « The Hypergraph-Based Data Structure : a new approach to data bases modeling and applications », 26-28 September, in Springer-Verlag, p.37-55 .

BOUILLE Francois, International Conference on Cybernetics and Society, Tokyo : « Fuzzy data processing with the Hypergraph-Based Data Structure », 3-7 November 1978, in IEEE Proceed., vol. 2, p.1222-1227.

BOUILLE Francois, US Air Force European Office of Aerospace and Development : « The Hypergraph-Based Data Structure and its applications to data structuring and complex systems modeling », September 1979, Report, 167p.

BOUILLE Francois, 7th Biennal CODATA Conference, Kyoto : « Integrating data from different sources at different scales », 8-11 October 1980, in Pergamon Press, 4p.

BOUILLE Francois, 2nd Int. HBDS Sem., Richmond (Virginia): « A multilayered HBDS with new extensions », 1981, March 9-13, 11pp. in Proceed.

BOUILLE Francois, 9th Int. CODATA Conf., Jerusalem : « A structured expert system for geodata banking », 1984, June 24-27, and in the Role of Data in Scientific Progress, Elsevier, p.417-420.

BOUILLE Francois, AGIT'94 Symposium, Salzburg : « Fuzziness structuring and processing in an object-oriented GIS », 1994, Proceed. Salzburger Geographische Materialien, Heft 21, p.113-122.

BOUILLE Francois, GIS Int.Conf. « Europe in Tansition », Brno : « Object-oriented methodology of structuring multiscale embedded networks », 1994, Aug.28-31, Proceed., chap.I, p.2-18.

BOUILLE Francois, 14th Int. CODATA Conf., Chambéry : « The use of GIS in petroleum industry for oil exploration and production », 1994, Sept.18-22, 6p.

BOUILLE Francois, CONGRESS on « Remote sensing of environment. Data processing and interpretation », Acad.of Sc., Moscow, « HBDS in GIS : overview on the model, methodology and applications » , invited paper, 1996, 17-18 May, 66p.

BOUILLE Francois, INTERCARTO-II International Conference, Irkutsk : « O-O methodology for 2-D images recovering the 3-rd dimension », 1996, 26-29 June, 16p.
Francois BOUILLE, Int. Sem. on Spatial Data Handling, Zürich : « Architecture of a geographic expert system », 1984, August 20-24, Proceed., 24pp.

BOUILLE Francois, Int. Conf. EUROCARTO VII, Enschede: « Developping strategies in GIS, by problem-solving methods based on a structured expert system », 1988, Sept.19-22, Proceed. 12pp. and in Environmental Applications of Digital Mapping, ITC.Pub., N°8, p.42-50.

BOUILLE Francois, EGIS/MARI'94 Int.Conf., Paris : « Methodology of building a class-based integrated platform for GIS design and development », 1994 , March 29-April, Proceed., Vol.1, p.909-918.

BOUILLE Francois, International Conference « Intercarto-6, 2000 », Apatity (Mourmansk Region, Russia) : « Reflex handling in O-O GIS », August 22-24, 2000, 12p. in Proceedings.

BOUILLE Francois, ISPRS-III-SSIDPCVN, München: « Principles of automated learning in a GIS, using an illimited set of object-oriented persistent neurons », 1994, Sept.5-9, Proceed. Ebner-Heipke-Eder Editors, Vol.30, Part 3/1, p.69-76.

COHEN, W.W., RAVIKUMAR, P., FIENBERG, S.E., *A comparison of String Distance Metrics for Name-Matching Tasks*, 2003

Cours de l'Ecole Nationale des Sciences Géographiques (ENSG)
http://fad.ensg.eu/moodle/course/category.php?id=12

DE JOINVILLE Olivier, *Visualisation, techniques d'amélioration de la visualisation des images numériques*, ENSG

ESCOFIER Brigitte, PAGES Jérôme, Analyses factorielles simples et multiples : Objectifs, méthodes et interprétations, 2008, DUNOD

HERMANSEN, J.C., Advanced Global Name Recognition. *White paper IBM*

HOLMES, David. MCCABE, Catherine. Improving Precision and Recall for Soundex Retrieval, 2002, pp 22-26

HUSSON Francois, Lê Sébastien, PAGES Jérôme, Analyse de données avec R, 2009, Presses Universitaires

JONES, C.B., SMART, Ph., TWAROCH, F., Deliverable 6.5: *Final Toponym Ontology Prototype*. Livrable de projet, Université de Cardiff, 2009

KESSLER, Carsten. Similarity Measurement in Context. *Lecture Notes in Computer Science*, 2007, pp 277-290

L.KIDD Alison, *Knowledge Acquisition for Expert*, Springer, 1987

OLTEANU, Ana-Maria. *Fusion de connaissances imparfaites pour l'appariement de données géographiques*, thèse, IGN/Paris-Est, 2008, 268 p.33

RODRIGEZ-BACHILLER Augustin, John Glasson, *Expert Systems and GIS for Impact Assessment*, Taylor&Francis Inc, 2004

SOARES Sunil, Data Governance Tools: Evaluation Criteria, Big Data Governance, and Alignement with Enterprise Data Management, 2014, MC Press

WILLIAMS, R.W. *Similarity measures for geographical place names*. Mémoire de projet de fin d'études, Université de Cardiff, 21/04/2011, 76 p.

Glossaire

ACP : analyse en composante principales, méthode d'analyse statistique appartenant à la famille des analyses factorielles et dont le but est d'ordonner les caractéristiques des données (attributs) de celle qui caractérise le plus le données à celle qui les caractérise le moins.

ADL : Algorithmic Descriptive Language, langage algorithmique mis au point par le Professeur Bouillé afin d'écrire des algorithmes de manière synthétique et en se libérant des spécificités des différents langages de programmation.

Arbre de décision : arbre binaire, ou automate, permettant dans le cadre de cette thèse de savoir quelle métrique de similarité et quels facteurs de tolérance on applique si à l'étape précédente les objets comparés ont été jugés similaires, et quels métrique et facteurs on applique sinon.

Automatisation : Mise en place de mécanismes pour que des chaînes de traitement et/ou de gestion de la donnée fonctionnent par elles-mêmes, et sans intervention de l'utilisateur selon le degré d'automatisation choisie ou possible.

CDP : Common Depth Point, point de réflexion de l'onde sismique sur la couche géologique.

Centroide : barycentre des SP d'une ligne de navigation sismique

Cluster : groupe de similarité, ensemble d'objets évalués comme étant semblables les uns aux autres et formant un groupe, après classification.

CMP : Common Middle Point, point milieu, correspondant au point à la surface de la Terre, à l'aplomb du point de réflexion de l'onde sismique, en approximant les couches géologiques à des droites parallèles à la surface du sol

Harmonisation de données : ensemble de procédés d'analyses, traitements et sélection des données d'une base pour en retirer les doublons et en harmoniser le contenu en termes de qualité des données.

ILX : InnerLogix, logiciel Schlumberger de synchronisation de bases de données et d'harmonisation de données.

LAC : Logiciel Automatique de Comparaisons, logiciel réalisé pendant cette thèse et commencé lors d'un stage chez TOTAL précédant le début du travail de recherche. Il est industrialisé depuis janvier 2013 chez TOTAL, en particulier utilisé dans les services de Data Management et de Documentation Technique. Il a été mis sous contrat de maintenance pendant un an, et ne l'est plus aujourd'hui. Son objectif est d'automatiser et rendre plus précis l'harmonisation des lignes de navigation sismiques.

Ligne/profil de navigation sismique : courbe à la surface de la Terre définie par un ensemble de points dont les coordonnées correspondent au dispositif d'acquisition sismique

Méthodologie AMR : Automatisation de la mesure de ressemblance, méthodologie fondée dans ce travail de thèse et mise en application pour l'harmonisation des bases de données de lignes de navigation sismique chez TOTAL.

Métrique de similarité : formule mathématique, ou algorithme, permettant de comparer deux objets selon un ou plusieurs critères afin d'évaluer s'ils sont similaires ou non et d'en quantifier la ressemblance.

Optimisation : amélioration d'un algorithme, programme, processus pour qu'il soit plus efficace, plus précis, plus exhaustif ou plus rapide, etc.

Rapport SP/CDP : nombre de points de tir dans un bin

Receiver Point : point récepteur du signal sismique, généralement un géophone

Ressemblance : la ressemblance peut être définie comme le jugement ou l'évaluation de la proximité conceptuelle entre objets, selon une résolution donnée. C'est une notion proche de celle de distance dans un espace attributaire. 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 , l'orthodromie ou la loxodromie… L'adverbe

« suffisamment » est relatif à la résolution à laquelle on évalue la ressemblance, ou encore au facteur de tolérance utilisé. On peut considérer que deux points sur une carte géographique sont semblables si la distance qui les éloigne est inférieure à 1 mètre par exemple. Deux objets sont proches s'ils tendent l'un vers l'autre (notion bien proche de celle de limite en analyse fonctionnelle mathématique). Par rapport à la distance géographique, la différence dans ces travaux de recherche est qu'on ne se place pas sur l'éllipsoide mais dans un « espace de similarité ».

SP : Shot Point, point de tir, source du signal sismique

Segmentation : méthode et mécanismes algorithmiques de traitement d'un fichier par blocs ou morceaux traités successivement.

Trace sismique : signal sismique capté par un géophone et correspondant à une source donnée pour un tir donné, et à une longueur d'onde donnée.

Article paru dans GES Journal en janvier 2014 (Geography, Environment, Sustainability)

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

Alba FUGA alba.fuga@neuf.fr

Laboratoire Sisyphe, UMR 7619, Université Pierre et Marie Curie Paris VI,

Boîte 105, 4 place Jussieu, 4 Tour 56 75252 Paris Cedex 05, France

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

1. 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.

2. 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 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.

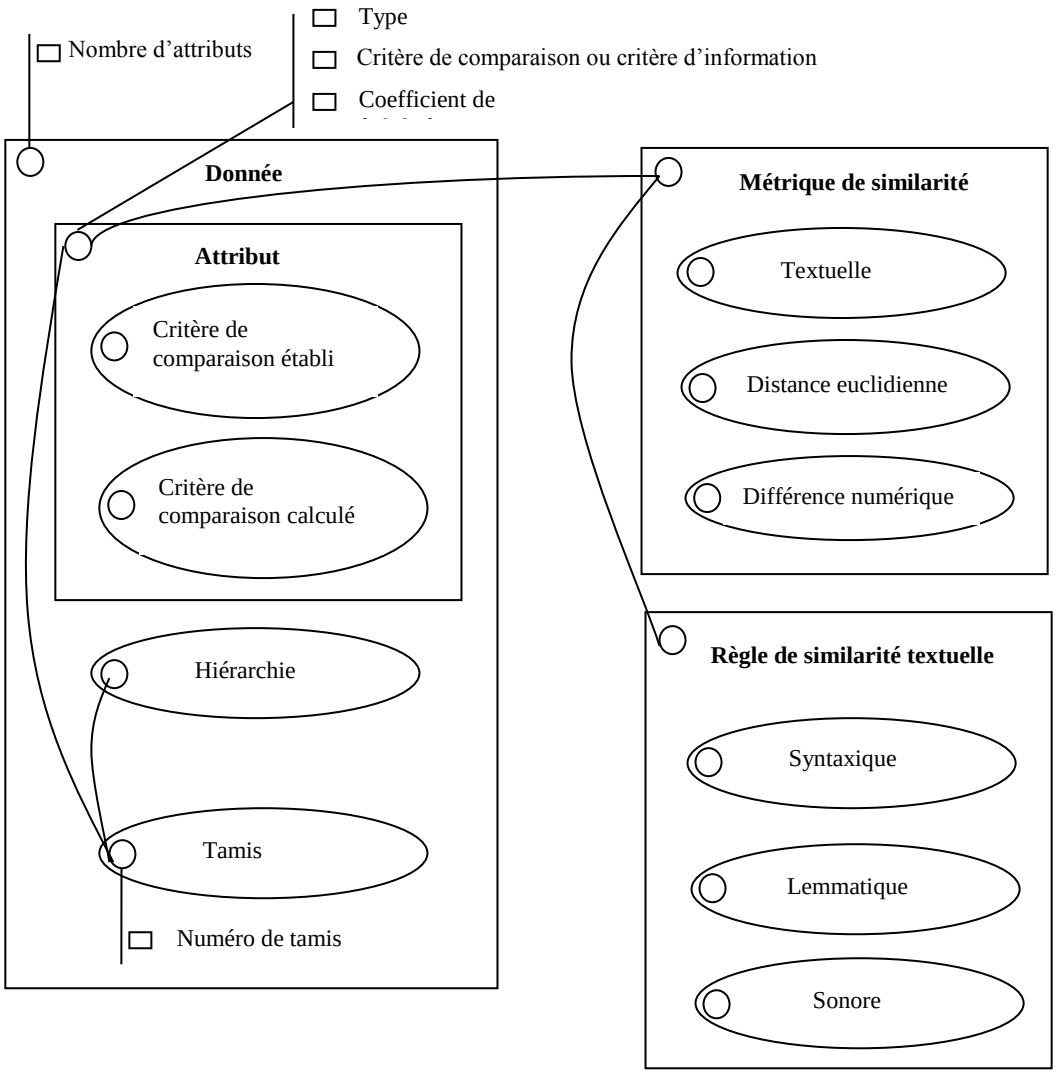


Figure 55 : Schéma HBDS de la méthodologie LAC

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.

3. 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.

4. 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

4. 1. 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.

4. 2. 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.

4. 3. 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 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.

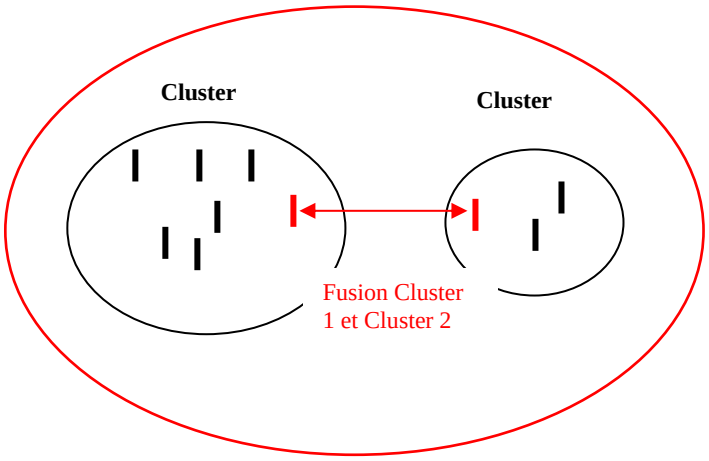


Figure 56 : Exemple de fusion entre deux clusters. Phénomène de « contagion »

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.

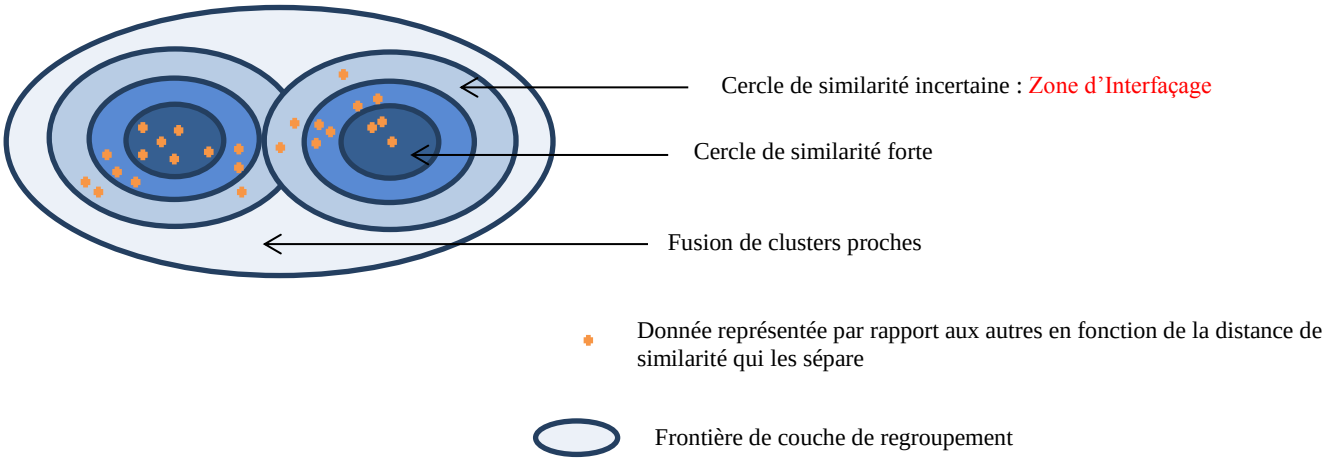


Figure 3 : Représentation de données dans un graphique de similarité, avec les contours de clusters formés selon différents vecteurs de similarité.

4. 4. 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.

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.

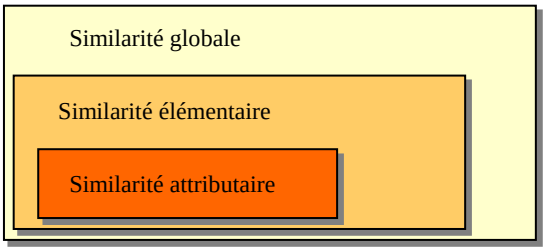


Figure 4 : Les trois échelles de mesure de Ressemblance

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 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.

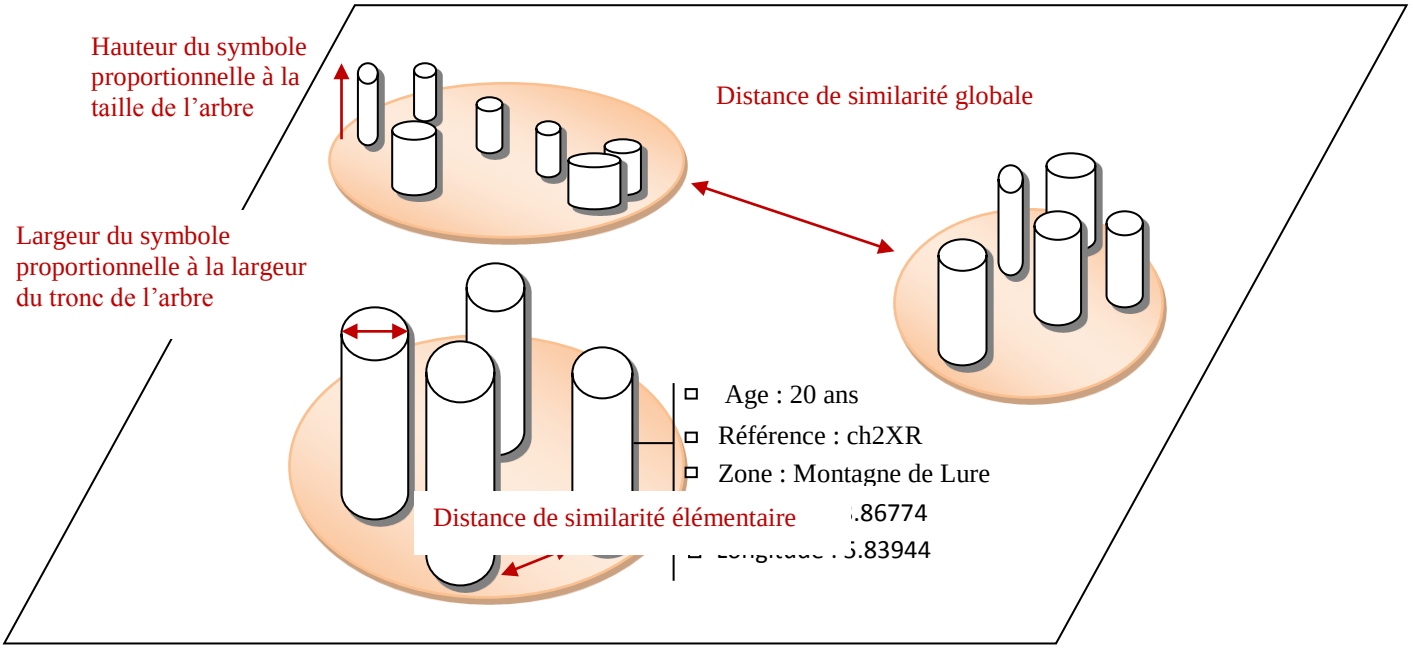


Figure 5 : Une possibilité de visualisation de clusters. Exemple sur des arbres que l'on a classifié selon leur taille, la largeur de leur tronc et leur positionnement géographique.

5. 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.

Algorithmes

Réorganisation

Voici l'algorithme de réorganisation de fichier d'entrée contenant des lignes de navigation sismiques, sous format SISMAGE.

Reorganisation(Fichier_in, Fichier_out, N):

∉

reprise ⊗ 0 ; ligne_reprise ⊗ 10 ; xc ⊗ 0 ; yc ⊗ 0 ; latc ⊗ 0 ; lonc ⊗ 0 ; sp_min ⊗ 0 ;

sp_max ⊗ 0 ; sp_min_x ⊗ 0 ; sp_min_y ⊗ 0 ; sp_max_x ⊗ 0 ; sp_max_y ⊗ 0 ; sp_min_lat

⊗ 0 ; sp_min_lont⊗ 0 ; sp_max_lat ⊗ 0 ; sp_max_lon ⊗ 0 ;

cdp_min ⊗ 0 ; cdp_max ⊗ 0 ; sp_nb ⊗ 0 ; div ⊗ 0 ; sp_cdp_preced ⊗ 0 ; ln_preced ⊗ 0 ;

length ⊗ 0 ; dernier_nav_x ⊗ 0 ; dernier_nav_y ⊗ 0 ; segmente ⊗ FALSE; imprime ⊗

FALSE;

⊆ \segmente Tant qu'on est toujours dans le segment courant

Reprise=0 ? Initialisation de la réorganisation : c'est le tout premier segment.

num_ligne ⊗ 0 ; numfin ⊗ 0 ; FIN est le vecteur dans lequel on stocke les lignes du fichier du segment avec les attributs, et donc tous les SP.

ligne_reprise ⊗ ligne_reprise + 1 ; à chaque ligne de fichier, on incrémente l'indice de la ligne de reprise. Elle évolue au fur et à mesure.

⊆ \LECTURE_SUIVANT(fichier_in, ligne)

numligne < 9 ? 9 lignes d'en-tête

ECRIT (fichier_out, ligne) ; |′

(numligne >= ligne_reprise-1) ≥ (numligne-ligne_reprise+1 < N) < 9 ?

Tant que la ligne de fichier est comprise entre la ligne de début du segment et la ligne de fin du segment, alors on effectue un traitement.

$\subseteq_{i:1}^{|ligne|}$ |ligne(i)|=0 ? ♥ ↓′

FIN_{num} ’AJOUT(ligne(i)); $\in_i$; numfin ⊗ numfin+1 ;

| (numligne – ligne_reprise +1) > N ? ! sortie de la boucle conditionnelle |′

numligne ⊗ numligne+1 ;

′

Maintenant, puisqu’on possède les informations du segment de fichier dans FIN, on commence la factorisation.

Condition de segmentation

| FIN | <N ? segmente ⊗ FALSE ;

Tant qu’on reste dans le domaine du segment courant, on ne passe pas à un autre segment.

| reprise ⊗ reprise +1 ; ′

Si on dépasse le segment courant, on incrémente le compteur de reprise.

Début de la factorisation de la ligne de navigation

i ⊗ 0 ;

$\subseteq^{\backslash i<|FIN|}$ length ⊗ 0 ; imprime ⊗ FALSE ;

A chaque donnée on réinitialise le déclencheur d’écriture dans le fichier de sortie à FALSE (donc pas d’écriture)

nom_ini ⊗ $FIN_{i,1}$; Premier nom de ligne de navigation, correspondant au nom de la première ligne de SP du segment. On repère les SP appartenant à la même ligne de navigation grâce au nom de la ligne qu’ils ont tous en commun. Remarque : ceci ne fonctionne pas si deux lignes de navigation portant le même nom se suivent. Elles seraient considérées concaténées.

$\subseteq_{j\,:\,i}^{|FIN|}$ nom_fin $\otimes$ $FIN_{j,1}$;

nom_ini != nom_fin ? i $\otimes$ j ; imprime $\otimes$ TRUE ; ! ; A l'instant où on ne trouve plus le même nom de ligne de navigation, alors on sait qu'on vient d'en factoriser une et on l'écrit dans le fichier final.

| Tant que les lignes du segment portent le même nom de ligne de navigation, on est sur une même ligne à factoriser.

$\subseteq_{k\,:\,0}^{|FIN_j|}$ ligne_nav'AJOUT($FIN_{j,k}$);$\in_k$;

nln $\otimes$ nln+1 ; i $\otimes$ i+1 ;

$'\in_j$;

lns $\otimes$ |ligne_nav| ; div $\otimes$ lns;

lns nous sevira de dividende pour faire la moyenne des coordonnées (calcul du barycentre), mais aussi pour accéder au dernier élément de ligne_navigation.

ln_preced $\otimes$ ligne_nav;

ligne_reprise $\otimes$ ligne_reprise+lns ;

Mise à jour de ligne_reprise grâce à la taille de ligne_navigation. La ligne de reprise vaut la dernière valeur de ligne_navigation, à laquelle on ajoute au fur et à mesure le nombre de lignes de fichier traitées par factorisation des lignes de navigation.

Initialisation des attributs à calculer.

sp_min $\otimes$ $ligne\_nav_{0,2}$; sp_max $\otimes$ $ligne\_nav_{0,2}$;

sp_min_x $\otimes$ $ligne\_nav_{0,4}$; sp_max_x $\otimes$ $ligne\_nav_{0,4}$;

sp_min_y $\otimes$ $ligne\_nav_{0,5}$; sp_max_y $\otimes$ $ligne\_nav_{0,5}$;

$ligne\_nav_{0,6}$ = « NON_RENSEIGNE » ?

sp_min_lat $\otimes$ « NON_RENSEIGNE » ; sp_max_lat $\otimes$ « NON_RENSEIGNE » ;

| sp_min_lat ⊗ $ligne\_nav_{0,6}$; sp_max_lat ⊗ $ligne\_nav_{0,6}$; ′

$ligne\_nav_{0,7}$ = « NON_RENSEIGNE » ?

sp_min_lon ⊗ « NON_RENSEIGNE » ; sp_max_lon ⊗ « NON_RENSEIGNE » ;

| sp_min_lon ⊗ $ligne\_nav_{0,7}$; sp_max_lon ⊗ $ligne\_nav_{0,7}$; ′

$ligne\_nav_{0,3}$ = « NON_RENSEIGNE » ?

cdp_min ⊗ « NON_RENSEIGNE » ; cdp_max ⊗ « NON_RENSEIGNE » ;

| cdp_min ⊗ $ligne\_nav_{0,3}$; cdp_max ⊗ $ligne\_nav_{0,3}$; ′

sp_nb ⊗ 0; xc ⊗ 0; yc ⊗ 0; latc ⊗ 0; lonc ⊗ 0; length ⊗ 0;

dernier_nav_x ⊗ 0; dernier_nav_y ⊗ 0;

Fin de l'initialisation des attributs à calculer.

Calcul des minima et maxima

$\subseteq_{j:0}^{lns}$ sp_min>= $ligne\_nav_{j,2}$?

sp_min ⊗ $ligne\_nav_{j,2}$; sp_min_x ⊗ $ligne\_nav_{j,4}$; sp_min_y ⊗ $ligne\_nav_{j,5}$;

$ligne\_nav_{j,6}$ = « NON_RENSEIGNE » ? sp_min_lat ⊗ "NON_RENSEIGNE";

| sp_min_lat ⊗ $ligne\_nav_{j,6}$; ′

$ligne\_nav_{j,7}$ = « NON_RENSEIGNE » ? sp_min_lon ⊗ "NON_RENSEIGNE";

| sp_min_lon ⊗ $ligne\_nav_{j,7}$; ′

$ligne\_nav_{j,3}$ = « NON_RENSEIGNE » ? cdp_min ⊗ "NON_RENSEIGNE";

| cdp_min ⊗ $ligne\_nav_{j,3}$; ′

|

$ligne\_nav_{j,2}$ > sp_max ?

sp_max $\otimes$ $ligne\_nav_{j,2}$; sp_max_x $\otimes$ $ligne\_nav_{j,4}$; sp_max_y $\otimes$ $ligne\_nav_{j,4}$; ′

$ligne\_nav_{j,6}$ = « NON_RENSEIGNE » ? sp_max_lat $\otimes$ "NON_RENSEIGNE";

| sp_max_lat $\otimes$ $ligne\_nav_{j,6}$; ′

$ligne\_nav_{j,7}$ = « NON_RENSEIGNE » ? sp_max_lon $\otimes$ "NON_RENSEIGNE";

| sp_max_lon $\otimes$ $ligne\_nav_{j,7}$; ′

$ligne\_nav_{j,3}$ = « NON_RENSEIGNE » ? cdp_max $\otimes$ "NON_RENSEIGNE";

| cdp_max $\otimes$ $ligne\_nav_{j,3}$; ′

′ $\in_j$;

Fin des calculs des minima et maxima

nav_sort : Vecteur portant les lignes de navigation factorisées, prêtes à être imprimées.

Remplissage avec les attributs jusqu'à présent disponibles.

nav_sort'AJOUT($ligne\_nav_{0,0}$); nom de campagne

nav_sort'AJOUT ($ligne\_nav_{0,1}$);nom de ligne

nav_sort'AJOUT ($ligne\_nav_{0,8}$); intervalle inter-traces

nav_sort'AJOUT (sp_min); numéro du sp min

nav_sort'AJOUT (sp_max); numéro du sp max

nav_sort'AJOUT (sp_min_x); coordonnée x du du sp min

nav_sort'AJOUT (sp_min_y); coordonnée y du du sp min

nav_sort'AJOUT (sp_max_x); coordonnée x du du sp max

nav_sort'AJOUT (sp_max_y); coordonnée y du du sp max

nav_sort'AJOUT (sp_min_lat); latitude du du sp min

nav_sort’AJOUT (sp_min_lon); longitude du du sp min

nav_sort’AJOUT (sp_max_lat); latitude du du sp max

nav_sort’AJOUT (sp_max_lon); longitude du du sp max

nav_sort’AJOUT (cdp_min); numéro du cdp min

nav_sort’AJOUT (cdp_max); numéro du cdp max

Calcul du rapport sp/cdp

sp_cdp $\otimes$ 0;

lns>1 ?

ligne_nav$_{0,3}$ $\otimes$ « NON_RENSEIGNE » ?

sp_cdp $\otimes$ "NON_RENSEIGNE";

| sp_cdp$\otimes$(ligne_nav$_{0,2}$-ligne_nav$_{1,2}$)/(ligne_nav$_{0,3}$- ligne_nav$_{1,3}$) ;′

sp_cdp_preced $\otimes$sp_cdp;

| sp_cdp $\otimes$ 1;

′

Ajout de sp/cdp dans nav_sort

sp_cdp != « NON_RENSEIGNE » ? nav_sort’AJOUT(|sp_cdp|);

| nav_sort’AJOUT (« NON_RENSEIGNE »);′

Calcul du centroïde

$\subseteq_{j\,:\,0}^{lns}$

xc $\otimes$ xc+ ligne_nav$_{j,4}$ sommation des x

yc $\otimes$ yc+ ligne_nav$_{j,5}$ sommation des y

ligne_nav$_{j,6}$ = «NON_RENSEIGNE » ? gestion des latitudes non renseignées

latc⊗ « NON_RENSEIGNE »;

| latc ⊗ latc+ ligne_nav$_{j,6}$; sommation des latitudes ′

ligne_nav$_{j,7}$ = « NON_RENSEIGNE » ? gestion des longitudes non renseignées

lonc ⊗ "NON_RENSEIGNE";

| lonc ⊗ lonc+ ligne_nav$_{j,7}$ sommation des longitudes′

$\in_j$;

xc ⊗ xc/lns;

yc ⊗ yc/lns;

latc != « NON_RENSEIGNE » ? latc ⊗ latc/lns;|′

lonc != « NON_RENSEIGNE » ? lonc ⊗ lonc/lns; |′

intégration du centroïde dans nav_sort

nav_sort’AJOUT(xc);

nav_sort’AJOUT (yc);

nav_sort’AJOUT (latc);

nav_sort’AJOUT (lonc);

Rédaction de la ligne de navigation sismique telle qu’on l’écrira dans le fichier de sortie (il s’agit juste de concaténer les éléments de nav_sort)

ligne ⊗ "";

$\subseteq_{j\,:\,0}^{|nav\_sort|}$ ligne ⊗ligne+nav_sort$_j$+" ";$\in_j$

sp_nb ⊗ sp_nb+lns;

ligne ⊗ ligne+sp_nb+" ";

Ajout du nombre de sp à la ligne à écrire

ligne ⊗ ligne +"0 ";

Calcul de la longueur linéaire, à partir des coordonnées planes

$\subseteq_{j\,:\,1}^{|ligne\_nav|}$

X1 $\otimes$ $ligne\_nav_{j-1,4}$;

Y1 $\otimes$ $ligne\_nav_{j-1,5}$;

X2 $\otimes$ $ligne\_nav_{j,4}$;

Y2 $\otimes$ $ligne\_nav_{j,5}$;

length $\otimes$ length+Distance(X1, Y1, X2, Y2); Distance euclidienne

$\in_j$

Ajout de la longueur linéaire à la ligne à écrire

ligne $\otimes$ ligne + length+" ";

Calcul de la densité de sp par unité de longueur linéaire de ligne

densite $\otimes$ length/div;

Ajout de la densité à la ligne à écrire

ligne $\otimes$ ligne + densite;

Mise à jour des coordonnées du dernier sp de la ligne traitée

dernier_nav_x $\otimes$ $ligne\_nav_{|ligne\_nav|,4}$

dernier_nav_y $\otimes$ $ligne\_nav_{|ligne\_nav|,5}$

imprime ?

S'il est temps d'imprimer, on imprime et réinitialise les paramètres qui serviront à la nouvelle ligne de navigation sismique à factoriser

ECRIT(ligne); impression de la ligne

Réinitialisation des paramètres

div ⊗ 0; sp_min ⊗ 0; sp_max ⊗ 0; sp_min_x ⊗ 0; sp_max_x ⊗ 0;

sp_min_y ⊗ 0; sp_max_y ⊗ 0;

sp_min_lat != « NON_RENSEIGNE » ? sp_min_lat ⊗ 0; |′

sp_max_lat != « NON_RENSEIGNE » ? sp_max_lat ⊗ 0; |′

sp_min_lon != « NON_RENSEIGNE » ? sp_min_lon ⊗ 0; |′

sp_max_lon != « NON_RENSEIGNE » ? sp_max_lon ⊗ 0; |′

cdp_min != « NON_RENSEIGNE » ? cdp_min ⊗ 0; |′

cdp_max != « NON_RENSEIGNE » ? cdp_max ⊗ 0; |′

sp_nb ⊗ 0; latc ⊗ 0; lonc ⊗ 0; length ⊗ 0; dernier_nav_x ⊗ 0;

dernier_nav_y ⊗ 0; ligne_nav'CLEAR ; nln ⊗ 0; |′

∈ ; Fin boucle conditionnelle

Ajustement de la ligne de reprise pour le raccordement (on commence le nouveau segment à la dernière ligne du dernier segment pour traiter les cas de ligne sur plusieurs segments)

ligne_reprise ⊗ ligne_reprise - 1;

| Ajustement Cas où on travaille sur des segments différents du tout premier, et où l'on est susceptible de résoudre des cas de raccordement entre segments
Recopiage préliminaire du fichier issu de la factorisation des segments précédents dans un fichier temporaire de relais pour ne pas écraser les lignes déjà factorisées.
ligne_nav.clear();Réinitialisation du vecteur de la dernière ligne de navigation factorisées

Début du traitement du segment courant

numligne ⊗ 0;

numfin ⊗ 0;

⊆ \LECTURE_SUIVANT(fichier_in, ligne)

numligne>=ligne_reprise-1 ≥ (numligne-ligne_reprise+1 < N) ?

ECRIT (fichier_out, ligne) ; |′

(numligne >= ligne_reprise-1) ≥ (numligne-ligne_reprise+1 < N) < 9 ?

Tant que la ligne de fichier est comprise entre la ligne de début du segment et la ligne de fin du segment, alrors on effectue un traitement.

$\subseteq_{i:1}^{|ligne|}$ |ligne(i)|=0 ? ⌊′

FIN$_{num}$ ’AJOUT(ligne(i)); ∈$_i$; numfin ⊗ numfin+1 ;

| (numligne – ligne_reprise +1) > N ? ! sortie de la boucle conditionnelle |′

numligne ⊗ numligne+1 ;

′

∈ Fin boucle conditionnelle de lecture et remplissage de FIN

Conditions de segmentation

|FIN| < N ? segmente ⊗ FALSE; | reprise ⊗ reprise+1;′

Factorisation

i ⊗ 0;

$\subseteq^{i < |FIN|}$ imprime ⊗ false;

A chaque donnée, on réinitialise le déclencheur d’écriture à faux (donc pas d’écriture)

nln⊗0;

Premier nom de ligne de navigation ; correspondant au nom de ligne de navigationde la première ligne (SP) du segment

nom_ini⊗$FIN_{i,1}$;

$\subseteq_{j:1}^{|FIN|}$ nom_fin ⊗ FIN.get(j).get(1);

nom_ini != nom_fin A l'instant où l'on ne trouve plus le même nom de ligne de navigation, on sait qu'on vient d'en factoriser une et on l'imprime dans le fichier final.

i⊗j; imprime = true; !

| Tant que les lignes du segment portent le même nom de ligne de navigation, on est sur une même ligne de navigation à factoriser.

$\subseteq_{k:1}^{|FIN_j|}$ $ligne\_nav_{nln}$'AJOUT($FIN_{j,k}$);$\in_k$;

nln⊗nln+1; i⊗i+1; $'\in_j$;

lns nous servira de dividende pour la moyenne, mais aussi pour accéder au dernier élément de ligne_navigation

lns ⊗ |ligne_navigation|

Mise à jour progressive de l'indice de ligne de reprise

ligne_reprise ⊗ ligne_reprise+lns;

Si la dernière ligne du segment précédent porte un nom différent de la ligne courante, cela veut dire que la dernière ligne de navigation du segment précédent n'est pas prolongée dans le segment courant. Alors on peut l'imprimer.

!imprime ≥ ln_preced. != $ligne\_nav_{0,1}$ ≥ ligne != "" ? ECRIT(FOUT,ligne); div ⊗ 0;|$'$

Dans le cas où la dernière ligne du segment précédent porte le même nom que la première ligne du segment courant, alors il faut procéder au raccordement. On raccorde la longueur linéaire en calculant la distance entre la dernière ligne du segment précédent et la première ligne du segment courant.

ln_preced = $ligne\_nav_{0,1}$?

length ⊗length+Distance(dernier_nav_x, dernier_nav_y, $ligne\_nav_{0,4}$, $ligne\_nav_{0,5}$) ;

| length ⊗ 0; ′

Le raccord pour les autres attributs se fait automatiquement car ici on ne les réinitialise pas. On ne les initialise que si la ligne de navigation est complète. Si la dernière ligne du segment précédent porte un nom différent de celui de la ligne courante, il n'y a pas de raccordement à faire, et on traite le segment.

ln_preced != $ligne\_nav_{0,1}$?

Initialisation des paramètres

sp_min ⊗ $ligne\_nav_{0,2}$; sp_max ⊗ $ligne\_nav_{0,2}$;

sp_min_x ⊗ $ligne\_nav_{0,2}$; sp_max_x ⊗ $ligne\_nav_{0,4}$;

sp_min_y ⊗ $ligne\_nav_{0,5}$; sp_max_y ⊗ $ligne\_nav_{0,5}$;

$ligne\_nav_{0,6}$ ⊗ "NON_RENSEIGNE" ?

sp_min_lat ⊗ "NON_RENSEIGNE"; sp_max_lat ⊗ "NON_RENSEIGNE";

| sp_min_lat ⊗ $ligne\_nav_{0,6}$; sp_max_lat ⊗ $ligne\_nav_{0,6}$;

′

$ligne\_nav_{0,7}$ ="NON_RENSEIGNE" ?

sp_min_lon ⊗ "NON_RENSEIGNE"; sp_max_lon ⊗ "NON_RENSEIGNE";

| sp_min_lon ⊗ $ligne\_nav_{0,7}$; sp_max_lon ⊗ $ligne\_nav_{0,7}$;

′

$ligne\_nav_{0,3}$ = "NON_RENSEIGNE" ?

cdp_min ⊗ "NON_RENSEIGNE"; cdp_max ⊗ "NON_RENSEIGNE";

| cdp_min ⊗ $ligne\_nav_{0,3}$; cdp_max ⊗ $ligne\_nav_{0,3}$; ′

sp_nb ⊗ 0; xc ⊗ 0.; yc ⊗ 0.; latc ⊗ 0.; lonc ⊗ 0.;

′

Calcul des min, max de x, y lon, lat

$\subseteq_{j:1}^{lns}$ sp_min >= $ligne\_nav_{j,2}$?

sp_min ⊗ $ligne\_nav_{j,2}$; sp_min_x ⊗ $ligne\_nav_{j,4}$; sp_min_y ⊗ $ligne\_nav_{j,5}$;

$ligne\_nav_{j,6}$ = "NON_RENSEIGNE" ?

sp_min_lat ⊗ "NON_RENSEIGNE"; | sp_min_lat ⊗ $ligne\_nav_{j,6}$; ′

$ligne\_nav_{j,7}$ = "NON_RENSEIGNE" ? sp_min_lon ⊗ "NON_RENSEIGNE";

| sp_min_lon ⊗ $ligne\_nav_{j,7}$; ′

$ligne\_nav_{j,3}$ = "NON_RENSEIGNE" ?

cdp_min ⊗ "NON_RENSEIGNE"; | cdp_min ⊗ $ligne\_nav_{j,3}$; ′

| sp_max < $ligne\_nav_{j,2}$? sp_max ⊗ $ligne\_nav_{j,2}$;

sp_max_x ⊗ $ligne\_nav_{j,4}$;

sp_max_y ⊗ $ligne\_nav_{j,5}$;

$ligne\_nav_{j,6}$ = "NON_RENSEIGNE" ?

sp_max_lat ⊗ "NON_RENSEIGNE"; | sp_max_lat ⊗ $ligne\_nav_{j,6}$; ′

$ligne\_nav_{j,7}$ = "NON_RENSEIGNE" ?

sp_max_lon ⊗ "NON_RENSEIGNE"; | sp_max_lon ⊗ $ligne\_nav_{j,7}$; ′

ligne_nav$_{j,3}$ = "NON_RENSEIGNE" ?

cdp_max ⊗ "NON_RENSEIGNE"; | cdp_max ⊗ ligne_nav$_{j,3}$;′

′ ∈$_j$;

Remplissage de nav_sort, la table des lignes de navigation sismique factorisées

nav_sort’AJOUT (ligne_nav$_{0,0}$); survey name

nav_sort’AJOUT (ligne_nav$_{0,1}$); line name

nav_sort’AJOUT (ligne_nav$_{0,8}$); average trace interval

nav_sort’AJOUT (sp_min);

nav_sort’AJOUT (sp_max);

nav_sort’AJOUT (sp_min_x);

nav_sort’AJOUT (sp_min_y);

nav_sort’AJOUT (sp_max_x);

nav_sort’AJOUT (sp_max_y);

nav_sort’AJOUT (sp_min_lat);

nav_sort’AJOUT (sp_min_lon);

nav_sort’AJOUT (sp_max_lat);

nav_sort’AJOUT (sp_max_lon);

nav_sort’AJOUT (cdp_min);

nav_sort’AJOUT (cdp_max);

Rem Calcul de SP/CDP

sp_cdp ⊗ 0;

lns>0 ? j⊗0;

lns = ?

sp_cdp ⊗sp_cdp_preced;

| ligne_nav$_{j,3}$ = "NON_RENSEIGNE" ?

sp_cdp ⊗ "NON_RENSEIGNE";

| sp_cdp ⊗ (ligne_nav$_{j,2}$- ligne_nav$_{j+1,2}$)/(ligne_nav$_{j,3}$- ligne_nav$_{j+1,3}$); ′

sp_cdp_preced ⊗ sp_cdp;

′

| sp_cdp ⊗ 1; Rem En cas d'élément non renseigné, on attribue un rapport sp/cdp = 1

′

sp_cdp != "NON_RENSEIGNE" ?

nav_sort'AJOUT(|sp_cdp|); | nav_sort'AJOUT("NON_RENSEIGNE"); ′

div ⊗ div+lns;

Calcul du centroide

$\subseteq_{j:1}^{lns}$ xc ⊗ (xc+ ligne_nav$_{j,4}$);

yc ⊗ (yc+ ligne_nav$_{j,5}$);

ligne_nav$_{j,6}$ = "NON_RENSEIGNE" ? latc ⊗ "NON_RENSEIGNE";

| latc ⊗ latc + ligne_nav$_{j,6}$;′

ligne_nav$_{j,7}$ = "NON_RENSEIGNE" ? lonc ⊗ "NON_RENSEIGNE";

| lonc ⊗ lonc + ligne_nav$_{j,7}$; ′

$\in_j$;

Imprime ? xc ⊗ xc/div; yc ⊗ yc/div;

latc != "NON_RENSEIGNE" ? latc ⊗ latc/div;|′

lonc != "NON_RENSEIGNE" ? lonc ⊗ lonc/div; |′

|′fin du remplissage de nav_sort avec le centroide

nav_sort'AJOUT(xc);
nav_sort'AJOUT(yc) ;
nav_sort'AJOUT(latc) ;
nav_sort'AJOUT(lonc) ;

Constitution de la chaine de caractères qui sera écrite dans le fichier de sortie. Il s'agit de la concaténation des attributs de la ligne factorisée.

ligne $\otimes$ "";

$\subseteq_{j:1}^{|nav\_sort|}$ ligne $\otimes$ ligne+nav_sort$_j$+" ";$\in_j$;

sp_nb $\otimes$ sp_nb+lns;

ligne $\otimes$ ligne + sp_nb)+" ";

ligne $\otimes$ ligne +"0 "; chaine Ajout de l'azimut à 0 car non calculé pour le moment.

Calcul de la distance linéaire.

$\subseteq_{j:1}^{lns}$ X1 $\otimes$ ligne_nav$_{j-1,4}$; Y1 $\otimes$ ligne_nav$_{j-1,5}$; X2 $\otimes$ ligne_nav$_{j,4}$; Y2 $\otimes$ ligne_nav$_{j,5}$;

length $\otimes$ length+Distance(X1, Y1, X2, Y2); $\in_j$;

Ajout de la distance linéaire à la ligne de fichier.

ligne $\otimes$ ligne + length+" ";

densite $\otimes$ length/div;

Mise à jour de la ligne précédente qui devient la ligne courante avant que la courante devienne celle qui lui succède

lns>0? ln_preced $\otimes$ ligne_nav$_{lns-1,1}$; | ln_preced $\otimes$ ligne_nav$_{lns,1}$; ′

Mise à jour des coordonnées de la dernière ligne de navigation du segment avec celle du segment courant avant de faire passer ce dernier au segment suivant.

dernier_nav_x $\otimes$ ligne_nav$_{|ligne_nav|-1,4}$;

dernier_nav_y $\otimes$ ligne_nav$_{|ligne_nav|-1,5}$;

Ajout de la densité à la ligne de fichier de sortie.

ligne $\otimes$ ligne + densite;

Lorsqu'il est temps d'écrire dans le fichier de sortie, et si la ligne n'est pas vide, on écrit et on réinitialise les paramètres.

imprime ≥ ligne != "" ? ECRIT(ligne, FOUT); dernier_nav_x ⊗ 0; dernier_nav_y ⊗ 0;

length ⊗ 0;′

Si on a imprimé une ligne, on réinitialise les paramètres pas encore réinitialisés.

Imprime ? div ⊗ 0; ligne_nav'CLEAR; ligne ⊗ ""; nln ⊗ 0; ′

∈ Fin boucle conditionnelle sur i

ligne_reprise ⊗ ligne_reprise - 1 ; on reprendra le traitement pour le segment suivant à la dernière ligne du segment précédent.

′ ∈ Fin boucle conditionnelle sur la segmentation

∠

Classification - Couples

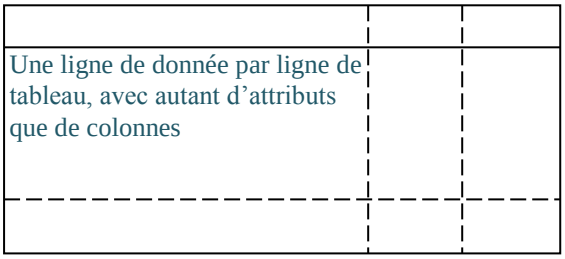
RegroupementCouples (TR, Fichier_in):

∉ TR est la colonne contenant les seuils de tolérance, ordonnée selon la hiérarchie attributaire.

Remplissage du tableau de chargement des données grâce à la lecture du fichier en entrée.

FIN ⊗ Lecture_Lignes(Fichier_in) ;

Structure de FIN :



numclust ⊗ 0 ; compteur du nombre de clusters

|FIN| |FIN|
⊆i : 1 ⊆j : 1

Remplissage de la table d'attributs TATT des deux lignes comparées (i, j). La première colonne contient les attributs de i. La seconde colonne contient les attributs de j. Les attributs sont ordonnés selon la hiérarchie attributaire.

TATT ⊗ Extrait_Att(FIN, i, j) ;

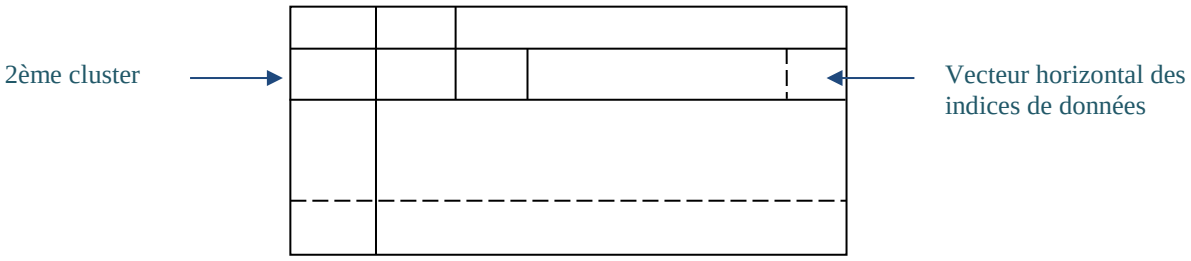
Calcul du coefficient de similarité global du couple (i, j) et du booléen disant si (i, j) passe le système de seuillage permettant de considérer les deux lignes comme effectivement semblables.

LEX ⊗ Regles_Filtrage(TR,TATT) ;

Structure de LEX : couple (booléen, mesure)

Clust ⊗ 0 ; Tableau des clusters, décrits par les indices dans FIN des lignes qui composent ces clusters

Structure de Clust :

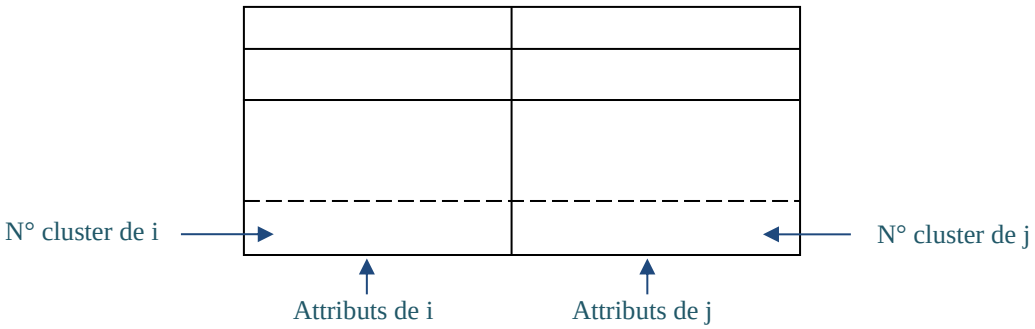


LEX_1 ? Le couple (i, j) est suffisamment similaire par rapport aux seuils de tolérance des critères de comparaison sélectionnés.

$TATT_{i,0}$!= $TATT_{j,0}$?

On place dans un même cluster uniquement les lignes dont les noms de campagnes sont exactement différents

Structure de TATT :



Clust $_{|numclust|}$'AJOUT(i);

Clust $_{|numclust|}$'AJOUT(j);

FIN(i, |FIN$_j$|) ⊗ numclust;

FIN(j, |FIN$_j$|) ⊗ numclust;

TATT(i, |FIN$_i$|) ⊗ numclust;

TATT(j, |FIN$_j$|) ⊗ numclust;

numclust ⊗ numclust+1 ;

′ Fin test sur les noms de campagne

′ Fin test sur la ressemblance entre i et j

$\in_j$; $\in_i$; ∠

Complexité de l'algorithme O(n^2).

Classification – Groupes Assymétriques

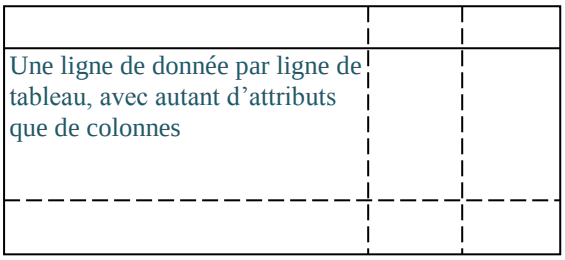
ClassificationAsymetrique (TR, Fichier_in):

∉ TR est la colonne contenant les seuils de tolérance, ordonnée selon la hiérarchie attributaire.

Remplissage du tableau de chargement des données grâce à la lecture du fichier en entrée, avec retrait des doublons exacts.

FIN ⊗ Lecture_Lignes(Fichier_in)'DEDOUBLONNE ;

Structure de FIN :



$\subseteq_{i:1}^{|FIN|}$ Clust(i, |Clust$_i$|)'AJOUT(i) $\subseteq_{j:1}^{|FIN|}$

Remplissage de la table d'attributs TATT des deux lignes comparées (i, j). La première colonne contient les attributs de i. La seconde colonne contient les attributs de j. Les attributs sont ordonnés selon la hiérarchie attributaire.

TATT ⊗ Extrait_Att(FIN, i, j) ;

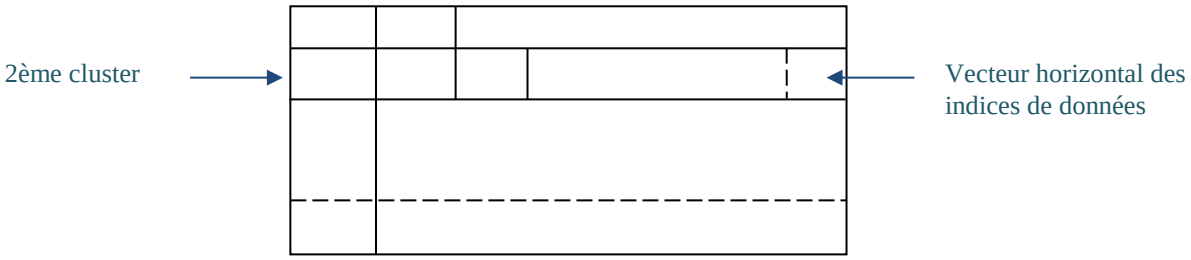
Calcul du coefficient de similarité global du couple (i, j) et du booléen disant si (i, j) passe le système de seuillage permettant de considérer les deux lignes comme effectivement semblables.

LEX ⊗ Regles_Filtrage(TR,TATT) ;

Structure de LEX : couple (booléen, mesure)

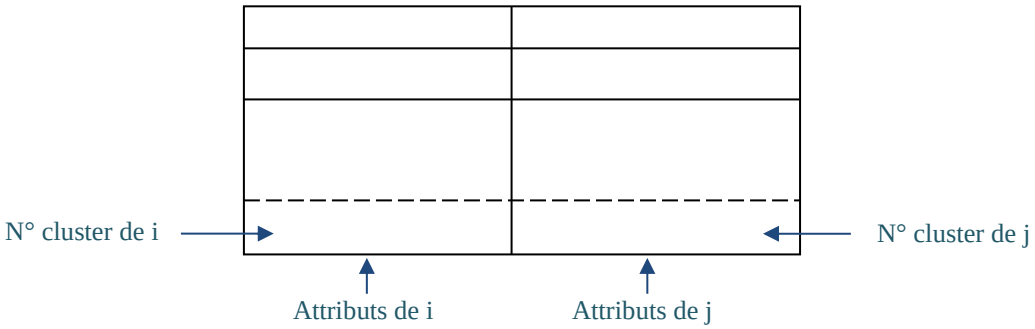
Clust ⊗ 0 ; Tableau des clusters, décrits par les indices dans FIN des lignes qui composent ces clusters. Ici il y a autant de clusters que de lignes dans FIN. On souhaite garder les clusters vides, élément important d'analyse car ils permettent de détecter les données uniques, par exemple contenues dans une base mais pas dans l'autre.

Structure de Clust :



LEX_1 ? Le couple (i, j) est suffisamment similaire par rapport aux seuils de tolérance des critères de comparaison sélectionnés.

Structure de TATT :



Clust(i, |Clust $_i$|)’AJOUT(j);

|′ Fin test sur LEX

$\in_j$; $\in_i$; ∠

Complexité de l’algorithme en O(n^2).

Clustering

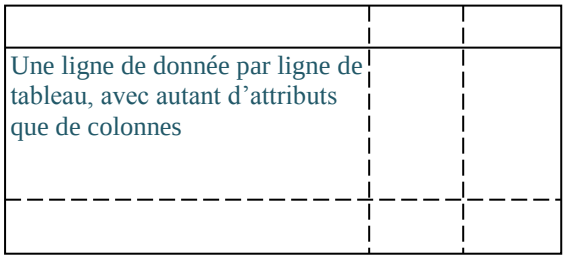
ClassificationCluster (TR, Fichier_in):

∉ TR est la colonne contenant les seuils de tolérance, ordonnée selon la hiérarchie attributaire.

Remplissage du tableau de chargement des données grâce à la lecture du fichier en entrée.

FIN ⊗ Lecture_Lignes(Fichier_in) ;

Structure de FIN :



numclust ⊗ 0 ; compteur du nombre de clusters

|FIN| |FIN|
⊆i : 1 ⊆j : 1

Remplissage de la table d'attributs TATT des deux lignes comparées (i, j). La première colonne contient les attributs de i. La seconde colonne contient les attributs de j. Les attributs sont ordonnés selon la hiérarchie attributaire.

TATT ⊗ Extrait_Att(FIN, i, j) ;

Calcul du coefficient de similarité global du couple (i, j) et du booléen disant si (i, j) passe le système de seuillage permettant de considérer les deux lignes comme effectivement semblables.

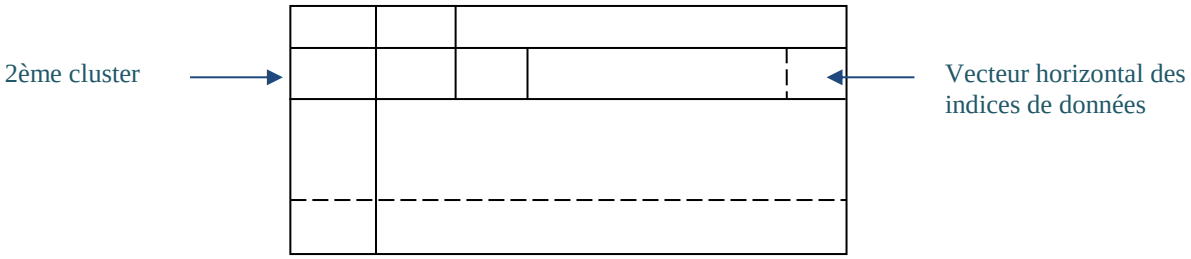
LEX ⊗ Regles_Filtrage(TR,TATT) ;

Structure de LEX : couple (booléen, mesure)

Booléen du passage du système de filtrage	Mesure de similarité globale

Clust ⊗ 0 ; Tableau des clusters, décrits par les indices dans FIN des lignes qui composent ces clusters

Structure de Clust :



Classification par densité de similarité

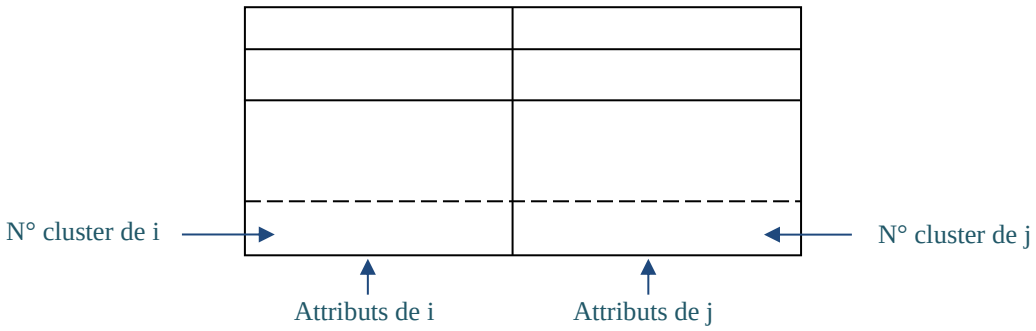
LEX_1 ? Le couple (i, j) est suffisamment similaire par rapport aux seuils de tolérance des critères de comparaison sélectionnés.

$FIN_{i,|FIN|}$= -1 ? i appartient déjà à un cluster.

$FIN_{j,|FIN|}$= -1 ≥ $FIN_{i,|FIN|}$? j appartient déjà à un cluster, mais différent de celui de i. On fusionne alors les deux clusters.

Borne ⊗ |Clust $_{TATT(1,\ |TATT|)}$|;

Structure de TATT :



borne
⊆d : 1 Clust $_{TATT(1,\ |TATT|)}$ ’AJOUT(Clust $_{TATT(1,\ |TATT|)}$, d);

FIN(Clust $_{TATT(1,\ |TATT|),\ d,}$ |TATT|) ⊗ FIN(j, |FIN_j|) ;

$TATT_{|TATT|,\ 1}$⊗ FIN(j, |FIN_j|) ;

Vidage du cluster de i

Clust $_{FIN(i,\ |FINi|)}$’ CLEAR;

∈d ;

' Fin test sur l’appartenance de j à un cluster

| i n’appartient à aucun cluster

FIN(j, |FIN_j|) != -1 ?

j appartient déjà à un cluster, alors on place i dans le cluster de j

Ajout du numéro de la donnée i dans le vecteur des numéros des données du cluster de j

Clust $_{FIN(i,\ |FINi|)}$’AJOUT(i);

FIN(i, |FIN_i|) ⊗ FIN(j, |FIN_j|) ;

$TATT_{|TATT|,\ 1}$⊗ FIN(j, |FIN_j|) ;

| ni i ni j n’appartiennent à un cluster. Il faut donc créer un nouveau cluster pour ce couple

Clust $_{|Clust|,0}$ ’AJOUT(numClust);

Clust $_{|Clust|,1}$ ’AJOUT(i);

Clust $_{|Clust|,2}$ ’AJOUT(j);

FIN(i, |FIN$_i$|) $\otimes$ numClust ;

FIN(j, |FIN$_j$|) $\otimes$ numClust ;

TATT$_{|TATT|,\ 1}$ $\otimes$ numClust ;

TATT$_{|TATT|,\ 2}$ $\otimes$ numClust ;

′ Fin test sur l’appartenance de j à un cluster

′ Fin test sur l’appartenance de i à un cluster

′ Fin test sur la ressemblance entre i et j

$\in_j$; $\in_i$; $\angle$

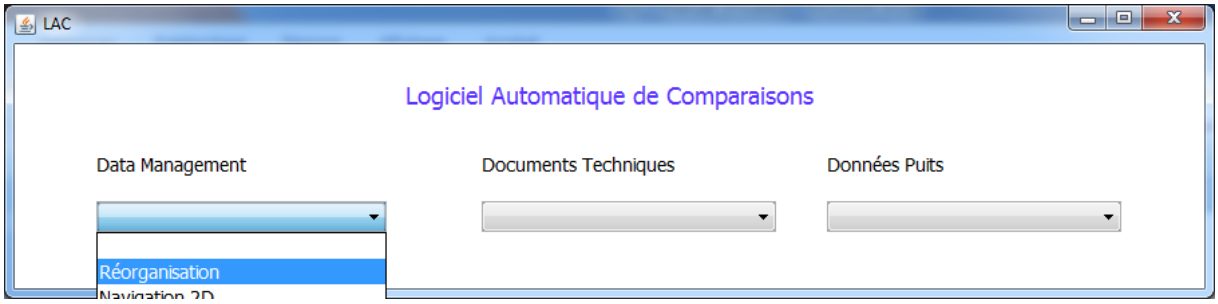
Complexité de l’algorithme en O(n^2).

Mode d'emploi LAC_DM (pour le Data Management)

Généralités

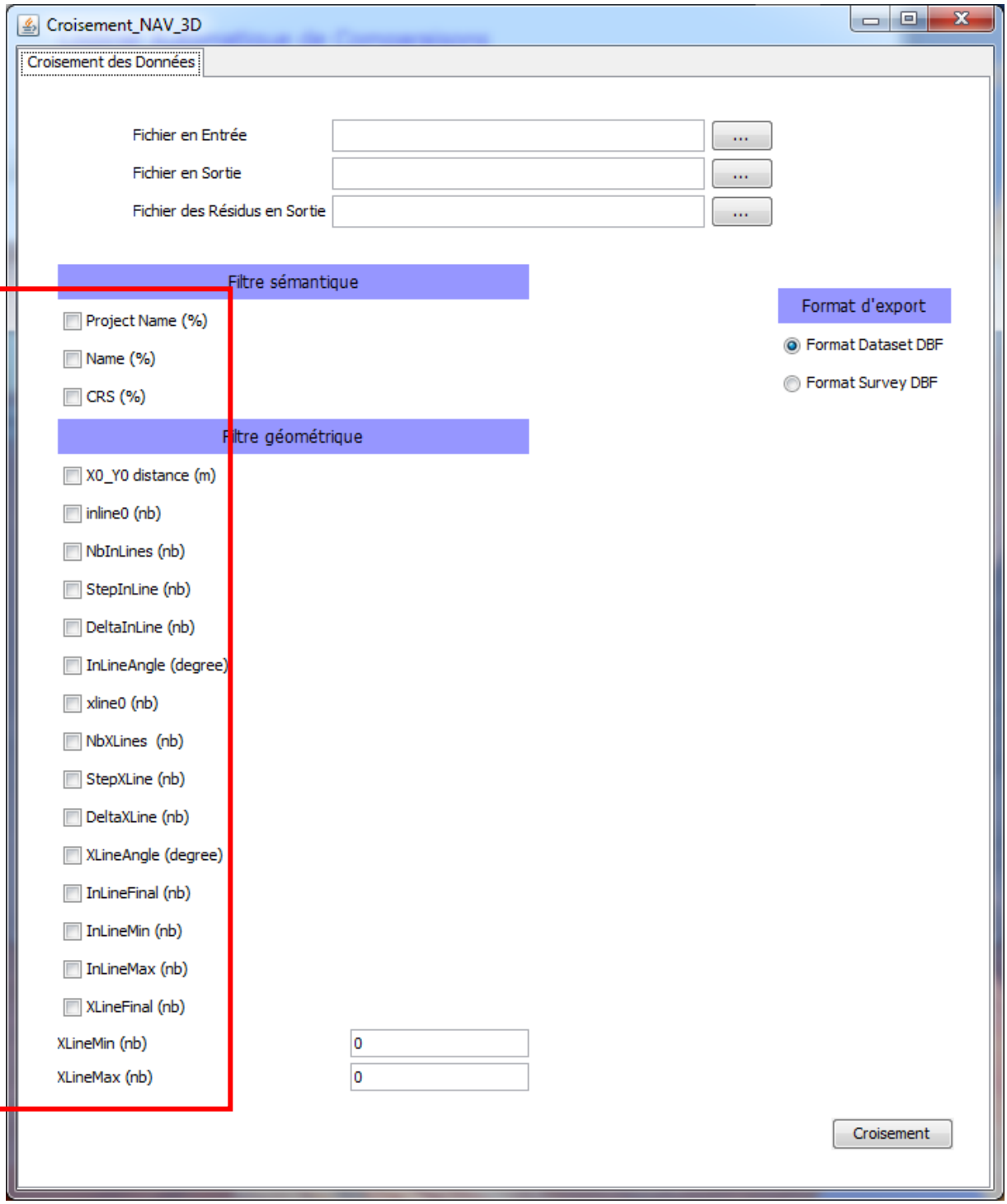
Installation : Il suffit de copier-coller l'exécutable de LAC. Dans la version actuelle, il est nécessaire d'avoir dans un même dossier l'exécutable et les fichiers à traiter.

Lancement : Double click sur l'exécutable LAC. Une première fenêtre apparaît où l'on choisit le module de l'application (DM_ Navigation 2D, DT_Rattachement, DP_LDB/Disk, …).

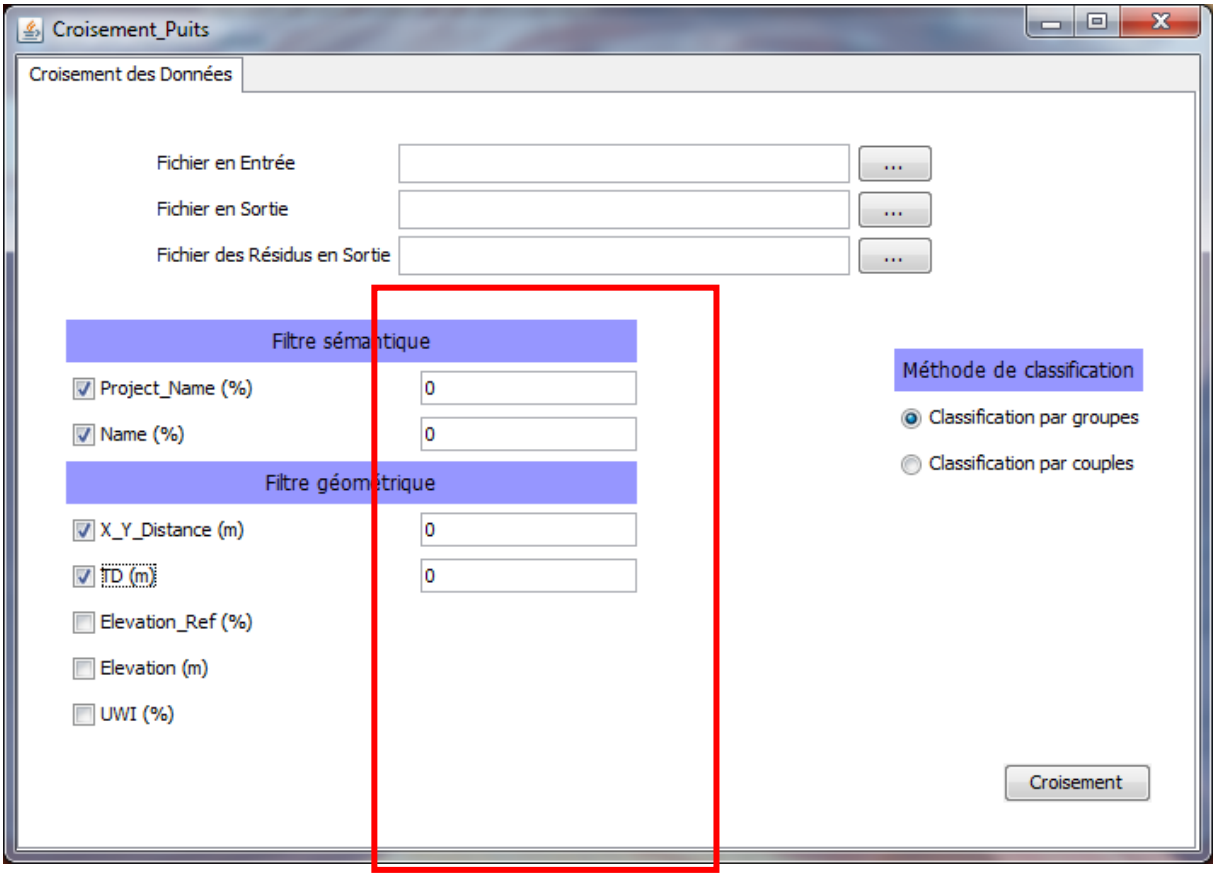


Paramétrage :

Critères de comparaison : Pour chaque thème, un jeu de critères de comparaison spécifique est disponible. Il est nécessaire de cocher les critères sur lesquels on souhaite effectuer le croisement, dans la liste hiérarchisée à gauche de la fenêtre qui apparaît lorsqu'on choisit le thème.


Croisement_NAV_3D
Croisement des Données
Fichier en Entrée
Fichier en Sortie
Fichier des Résidus en Sortie
...
Filtre sémantique
Project Name (%)
Name (%)
CRS (%)
Filtre géométrique
X0_Y0 distance (m)
inline0 (nb)
NbInLines (nb)
StepInLine (nb)
DeltaInLine (nb)
InLineAngle (degree)
xline0 (nb)
NbXLines (nb)
StepXLine (nb)
DeltaXLine (nb)
XLineAngle (degree)
InLineFinal (nb)
InLineMin (nb)
InLineMax (nb)
XLineFinal (nb)
XLineMin (nb)
0
XLineMax (nb)
0
Format d'export
Format Dataset DBF
Format Survey DBF
Croisement

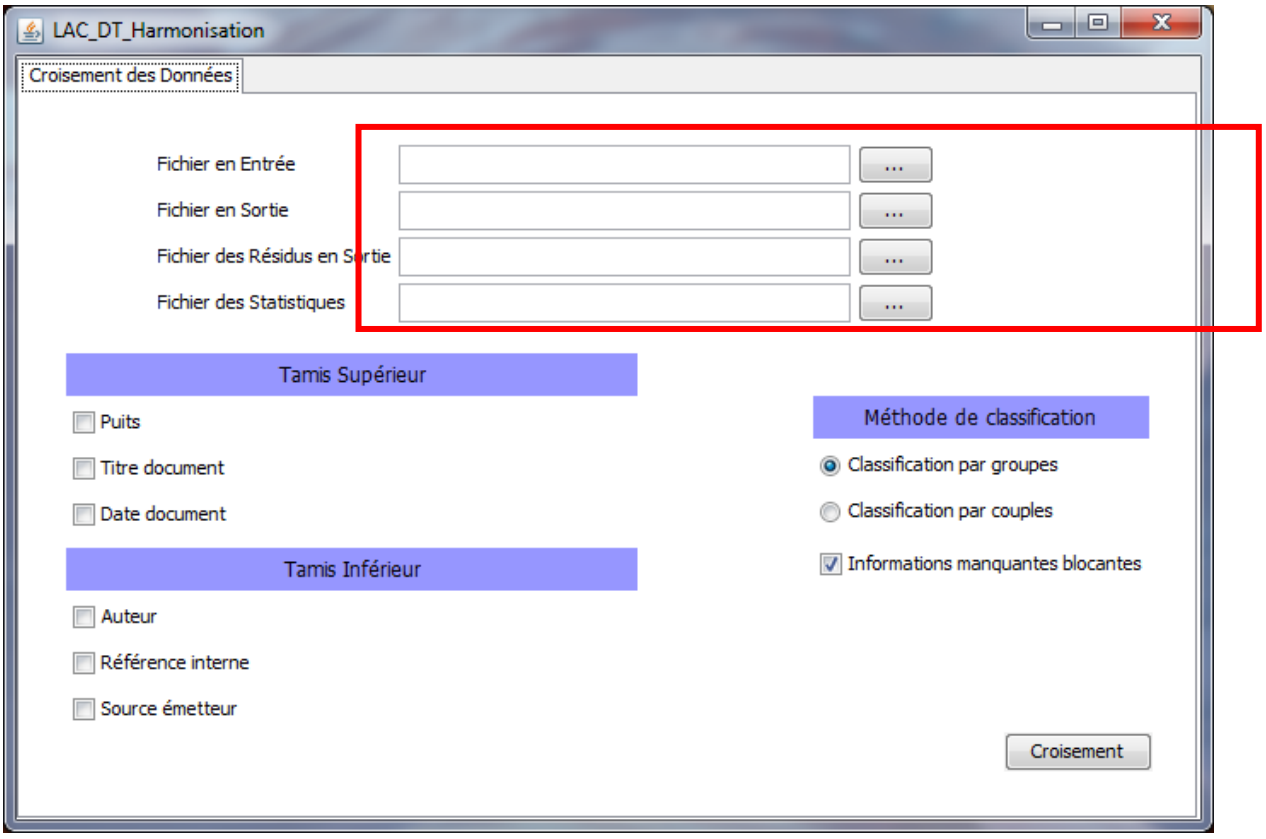
Seuils de tolérance : Pour chaque critère de comparaison sélectionné, il faut spécifier un seuil de tolérance, suivant l'unité demandée. Les entités considérées comme similaires seront celles qui auront une ressemblance supérieure au pourcentage de tolérance donné pour les critères textuels, et une différence inférieure aux seuils de tolérance des autres critères (mètres, degrés, unités…). L'espace de spécification du seuil de tolérance apparaît lorsqu'on sélectionne le critère de comparaison.



Entrées et Sorties : Le nom du fichier à traiter doit être renseigné à l'emplacement indiqué en haut à gauche de la fenêtre de paramétrage. Ensuite, il est nécessaire de spécifier le nom du rapport d'analyse contenant l'ensemble des groupes de similarité, et d'indiquer le nom du fichier de résidus contenant tout élément semblable à aucun autre, selon le traitement effectué.

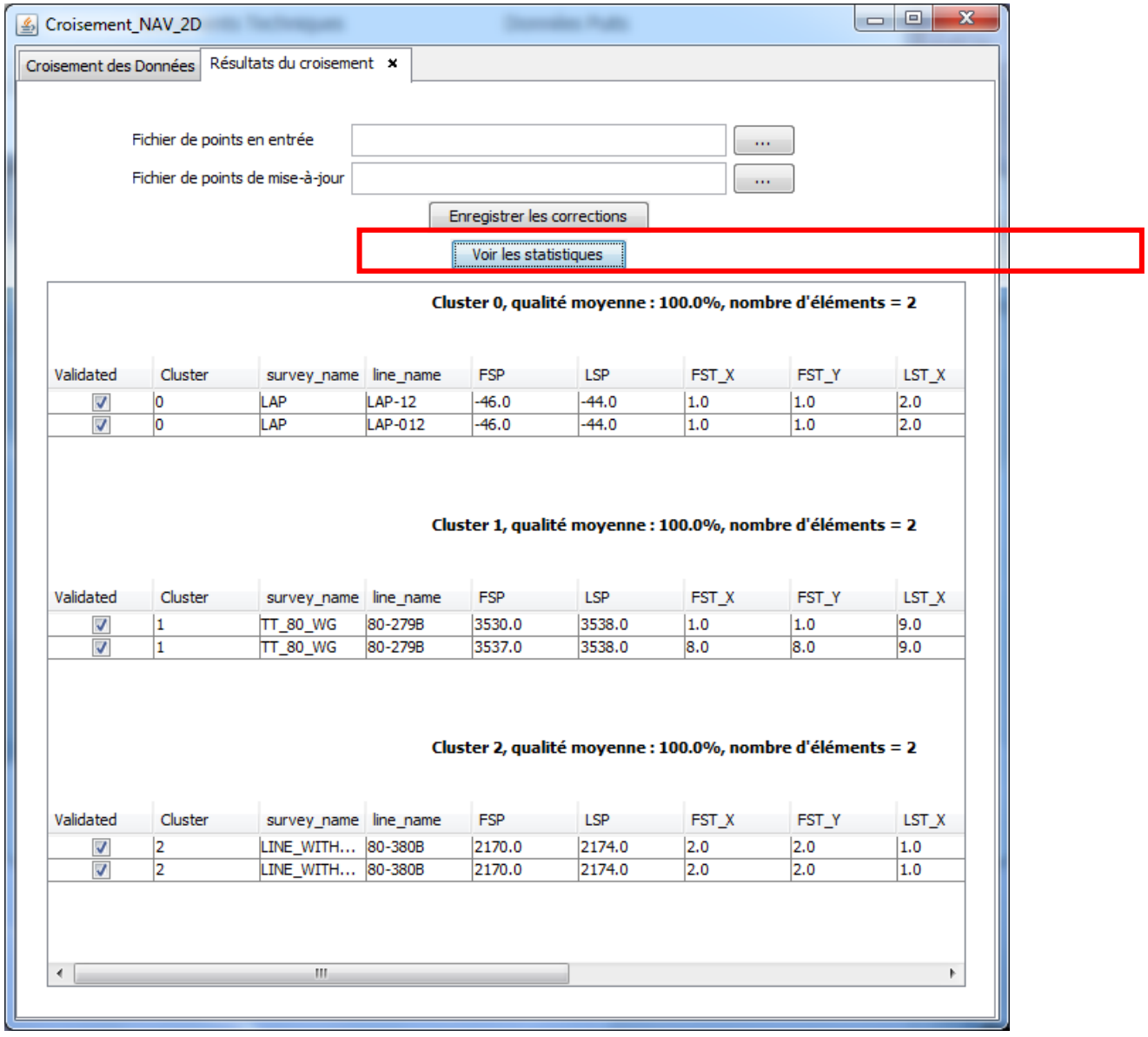
Les rapports d'analyse sont créés dans le dossier contenant l'exécutable de LAC. Pour lancer le traitement, il faut cliquer sur le bouton « Croisement ». Le traitement finit toujours lorsque le bouton « Croisement » redevient orange. Une interface temporaire d'exposition de résultats de croisement peut apparaître, cependant le document de référence pour analyser ou lire les résultats reste le rapport d'analyse.

Attention : le rapport de sortie ne doit pas être ouvert pendant que le traitement a lieu, sinon les résultats n'y seront pas écrits.

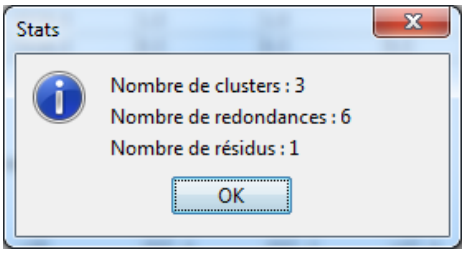


Affichage des résultats via l'interface

Les résultats de classification s'affichent dans un onglet sur la même fenêtre. De cette manière on peut lancer plusieurs traitements de données par l'interface de croisement, et pour chaque traitement, le résultat sera affiché sur un nouvel onglet. On pourra donc comparer via l'interface les différents traitements et paramétrages des traitements.



Le bouton « Voir Statistiques » permet d'afficher des statistiques de traitement soit généraux, soit spécifiques à un type de données.

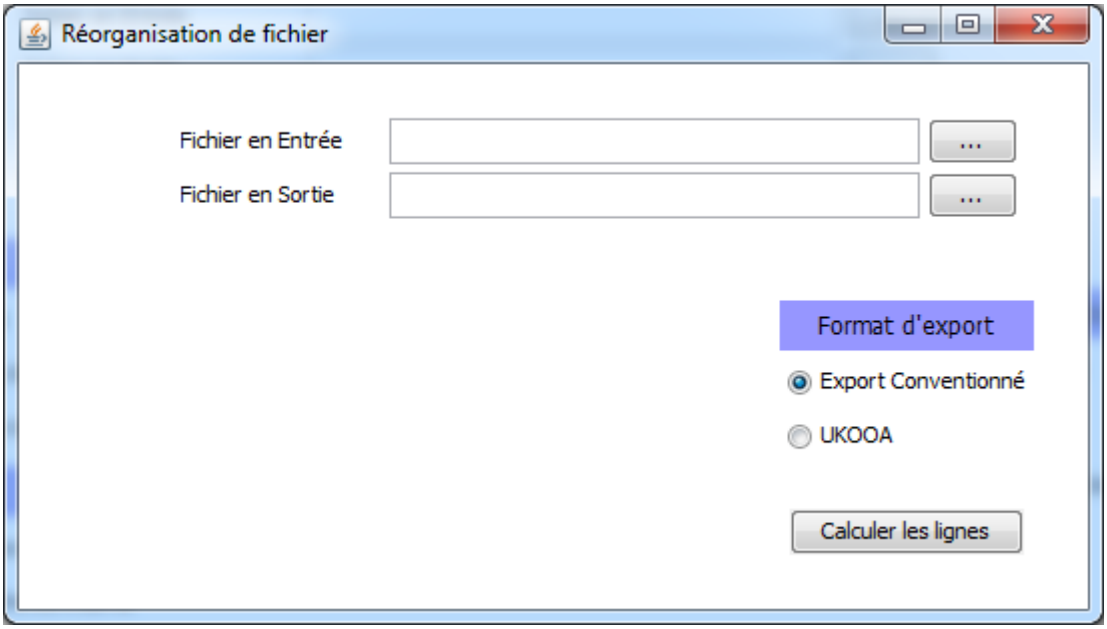


LAC_DM pour les lignes de navigation 2D

Prétraitement : Afin de comparer des profils de navigation sismique 2D, le point de départ est un export des lignes de navigation sismique avec tous leurs SP et CDP. Cependant, la comparaison ne se fait pas point par point, mais ligne par ligne.
Il est donc nécessaire, avec d'effectuer le croisement, de :

- Calculer des critères de comparaison supplémentaires caractérisant chaque ligne de manière globale, comme le centroïde, la longueur linéaire de la ligne, le rapport SP/CDP lorsqu'il est calculable, le premier et le dernier SP et CDP de chaque ligne.
- Factoriser une ligne de navigation sismique en une ligne de fichier (car dans l'export de départ, une ligne de navigation sismique se déploie sur autant de lignes de fichier qu'elle contient de SP)
- Dans le processus de factorisation, on comprend aussi une phase d'organisation des colonnes du fichier de sortie du prétraitement.

Le prétraitement s'effectue en cliquant sur le bouton « Réorganisation_Nav » de la fenêtre de démarrage. Il est nécessaire d'y renseigner le nom du fichier d'export contenant tous les SP, et le fichier de sortie du prétraitement contenant les lignes factorisées avec les critères de comparaison complémentaires.
On doit aussi cocher la case « Export Conventionné » pour les fichiers d'entrée formatés comme un export ASCII Sismage, ou bien la case « Export UKOOA » pour les fichiers d'entrée de format UKOOA. Donc seuls les fichiers SISMAGE/UKOOA sont reconnus.
Le bouton « Calculer Lignes » permet de lancer le prétraitement. Il se grise au cours du traitment et redevient orange une fois le prétraitement terminé.



Formats entrée :

Format d'entrée suivant la convention d'extraction de Sismage pour le prétraitement :

```
------Begin Header----------
Export of geometry info for 2D line= LAP-12
Datum: WGS 84
Reference Meridian: Greenwich
Projection System: UTM zone 20N   63W
Unit: m
Average trace interval: 51.04752947062986
------End Header----------

survey_name          line_name SP          CDP     X       Y       latitude  longitude  average_trace_interval
LAP         LAP-12   -46.0     1.0         1       1       10.013  -60.897   51.047
LAP         LAP-12   -44.0     2.0         2       2       10.013  -60.000   51.047
TT_80_WG             80-279B   3530.0      356.0   1       1       11.098    -60.139   12.7361
TT_80_WG             80-279B   3531.0      356.0   2       2       11.098    -60.139   12.736
```

Format de sortie du prétraitement, et format d'entrée pour le traitement :

```
------Begin Header----------
Export of geometry info for 2D line= LAP-12
Datum: WGS 84
Reference Meridian: Greenwich
Projection System: UTM zone 20N  63W
Unit: m
Average trace interval: 51.04752947062986
------End Header----------

survey_name  line_name  trace_interval  fst_sp  last_sp  fst_x  fst_y  lst_x  lst_y  fst_lat  fst_lon  lst_lat  lst_lon  fst_cdp  lst_cdp  sp/cdp  centroid_x  centroid_y  centroid_lat  centroid_lon  sp_nb  AZ  length  density
LAP          LAP-12     51.047          -46.0   -44.0    1.0    1.0    2.0    2.0    10.013   -60.897  10.013   -60.     1.0      2.0      2.0     1.5         1.5         10.013        -60.897       2.0    0   1.414   0.707
TT_80_WG     80-279B    12.736          3530.0  3538.0   1.0    1.0    9.0    9.0    11.098   -60.139  11.097   -60.138  356.0    365.0    2       5.0         5.0         11.097        -60.138       9.0    0   11.313  1.2572
```

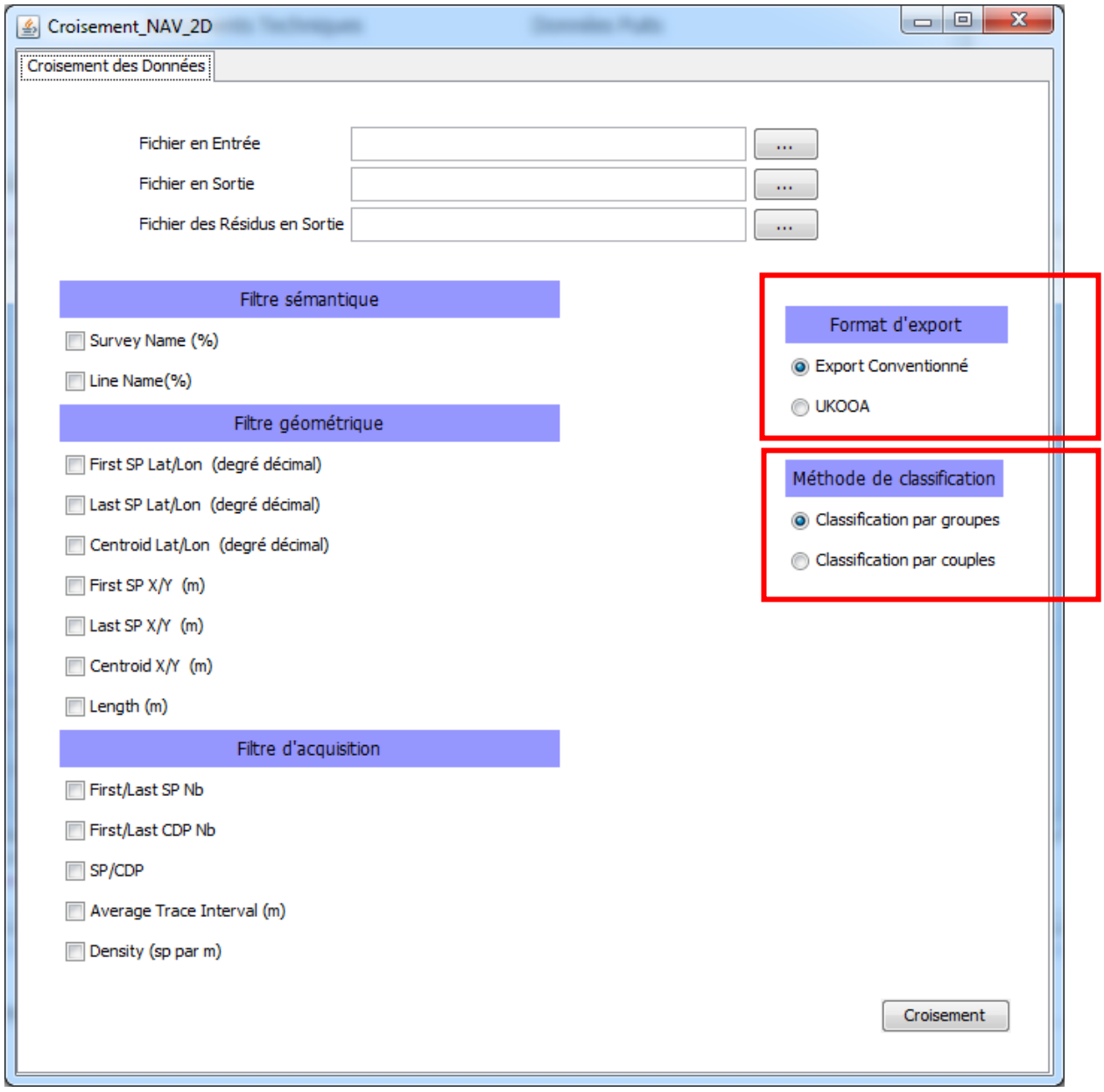
Remarque : Il faut bien 10 lignes d'en-tête à chaque fois. Si on travaille avec des fichiers n'ayant pas ces 10 lignes, il suffit de sauter 10 lignes (en allant à la ligne 10 fois avant les données).

Format du rapport d'analyse :

cluster	survey_name	line_name	min_name	max_name	moy_name	FSP	LSP	FST_X	FST_Y	LST_X	LST_Y	FCDP	LCDP	C_x	C_y	length
0	LAP	LAP-12	70	70	70	-46.0	-44.0	1.0	1.0	2.0	2.0	1.0	2.0	1.5	1.5	1.414
0	LAP	LAP-012	70	70	70	-46.0	-44.0	1.0	1.0	2.0	2.0	1.0	2.0	1.5	1.5	1.414
1	TT_80_WG	80-279B	100	100	100	3530.0	3538.0	1.0	1.0	9.0	9.0	356.0	365.0	5.0	5.0	11.313
1	TT_80_WG	80-279B	100	100	100	3537.0	3538.0	8.0	8.0	9.0	9.0	365.0	365.0	8.5	8.5	1.414
2	LINE_WITH_NO_XYs	80-380B	100	100	100	2170.0	2174.0	2.0	2.0	1.0	1.0	1926.0	1934.0	1.5	1.5	1.414
2	LINE_WITH_NO_XYs	80-380B	100	100	100	2170.0	2174.0	2.0	2.0	1.0	1.0	1926.0	1934.0	1.5	1.5	1.414

Dans le rapport d'analyse, on identifie les différents groupes de similarité par leur identifiant de cluster dans la première colonne du tableau.
Ensuite, les éléments contenus dans un groupe sont placés successivement ligne par ligne dans le rapport. Les attributs qui sont renseignés sur ces éléments sont : le nom de la campagne, le nom de la ligne, le premier SP, le dernier SP, les coordonnées X et Y des premier et dernier SP, le premier CDP, le dernier CDP, les coordonnées X et Y du centroïde et la longueur linéaire.
De plus, pour chaque critère de comparaison coché dans l'interface, apparaissent en rapport d'analyse le minimum, le maximum et la moyenne de la mesure de similarité faite sur un cluster donné, pour ce critère (en orange dans le tableau).

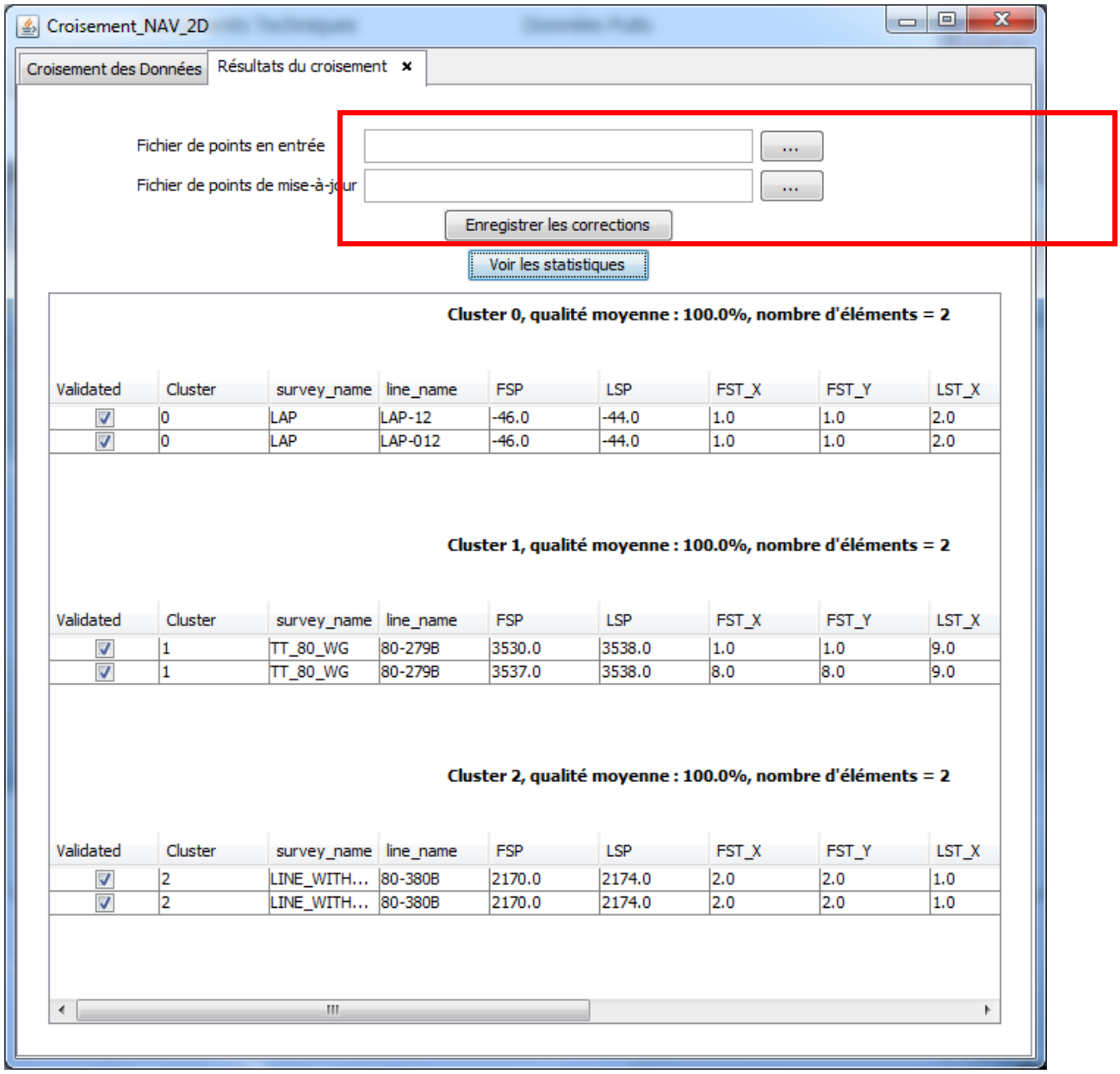
Spécificités de traitement : On a le choix entre deux méthodes de classifications, une classification par couples de lignes ne provenant pas de la même campagne, et une classification par groupes de similarité de deux lignes ou plus, sans contraintes sur la campagne.
Comme pour le prétraitement, il est nécessaire de préciser si le format de départ provenait d'un fichier UKOOA ou bien d'un export conventionnel Sismage (sans oublier d'exporter la survey name).



Par ailleurs, pour la Navigation 2D il est possible de générer un fichier des données corrigées (c'est-à-dire ne comportant pas les lignes que l'on souhaite retirer du projet) avec l'intégralité des SP par lignes, format identique au fichier initial exporté de SISMAGE.

Pour cela, il faut :

- Spécifier le nom du fichier de départ par SP, exporté de SISMAGE
- Spécifier le nom du fichier de données corrigées que l'on souhaite créer
- Décocher les lignes que l'on souhaite retirer des données de départ.


Croisement_NAV_2D
Croisement des Données
Résultats du croisement
Fichier de points en entrée
Fichier de points de mise-à-jour
...
...
Enregistrer les corrections
Voir les statistiques
Cluster 0, qualité moyenne : 100.0%, nombre d'éléments = 2
Validated Cluster survey_name line_name FSP LSP FST_X FST_Y LST_X
0 LAP LAP-12 -46.0 -44.0 1.0 1.0 2.0
0 LAP LAP-012 -46.0 -44.0 1.0 1.0 2.0
Cluster 1, qualité moyenne : 100.0%, nombre d'éléments = 2
Validated Cluster survey_name line_name FSP LSP FST_X FST_Y LST_X
1 TT_80_WG 80-279B 3530.0 3538.0 1.0 1.0 9.0
1 TT_80_WG 80-279B 3537.0 3538.0 8.0 8.0 9.0
Cluster 2, qualité moyenne : 100.0%, nombre d'éléments = 2
Validated Cluster survey_name line_name FSP LSP FST_X FST_Y LST_X
2 LINE_WITH... 80-380B 2170.0 2174.0 2.0 2.0 1.0
2 LINE_WITH... 80-380B 2170.0 2174.0 2.0 2.0 1.0

LAC_DM pour les lignes de navigation 3D

Format d'entrée : pour l'instant on travaille sur des exports en format texte, séparateur tabulation, de fichiers DBF.

Format du rapport d'analyse : il fonctionne comme pour la navigation 2D. Pour la 3D, tous les attributs sont renseignés. Lorsqu'un critère de comparaison est sélectionné, minimum, maximum et moyenne de mesures de similarité sur un cluster sont affichés sur un attribut donné.

Spécificités de traitement : uniquement une classification par groupes de similarité a été nécessaire jusqu'à présent.

LAC_DM pour les puits

Format d'entrée :

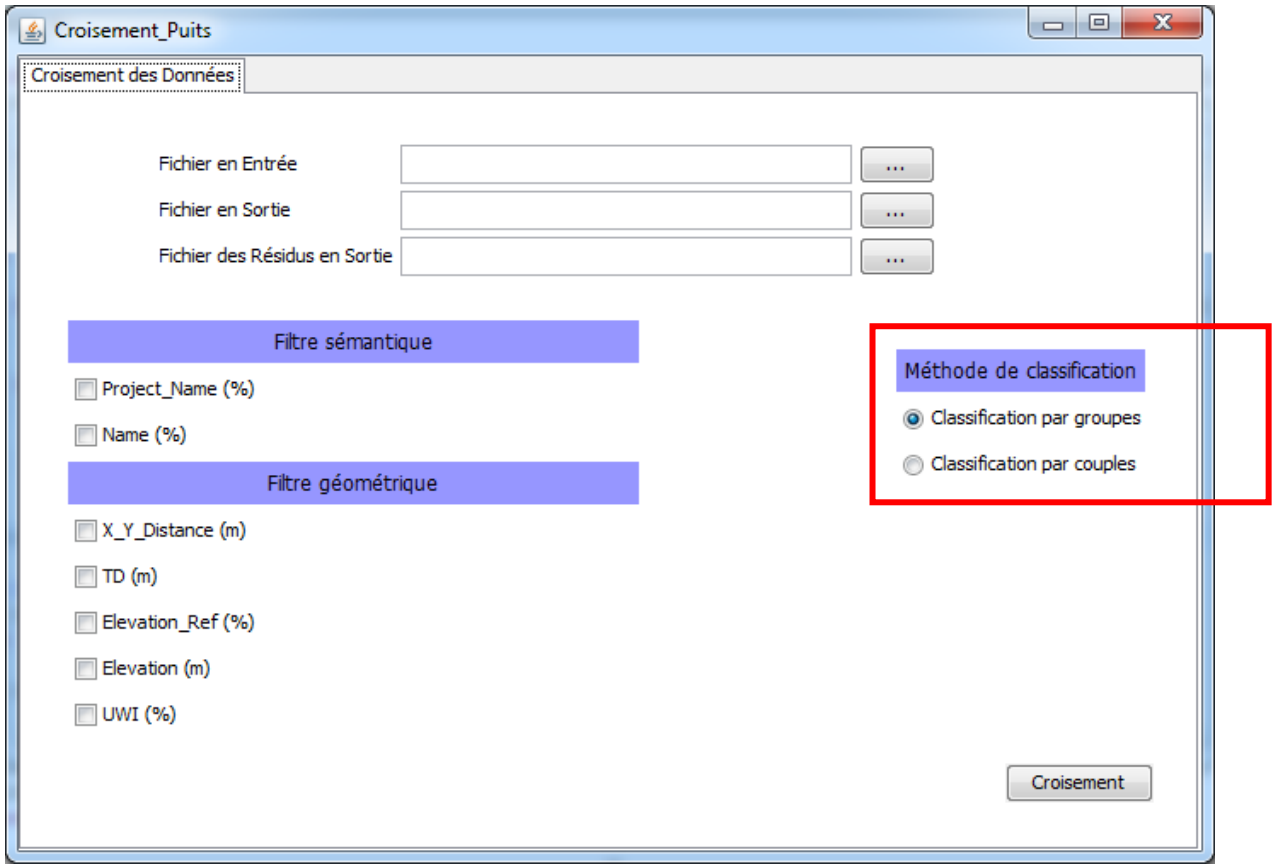
Ce format nécessite une ligne d'en-tête (une seule).

Proj_Name	Name	X	Y	TD	Elevation_Ref	Elevation	UWI	Source
Proj_Name1	Name1	1	2	5	Ref	5	5	Source
Proj_Name2	Name2	1	1	5	Ref	5	5	Autre

Format du rapport d'analyse :

Le rapport d'analyse fonctionne comme la navigation 3D, et contient tous les attributs.

Spécificités de traitement : Deux méthodes de classification possibles, comme pour la navigation 2D : classification par groupes ou classification par couples.



Mode d’emploi de LAC_DT (pour la Documentation Technique)

LAC_DT pour l’harmonisation : comparaison de fonds Siège-Filiale, harmonisation de fonds

Formats d’entrée :

Aussi bien le fichier Filiale que le Fichier Siège doivent être formatés selon l’exemple suivant. Ce format nécessite une ligne d’en-tête (une seule).

Noms de puits	Titre	Auteurs	Date	Référence_interne	Source_émettrice	eDoc	BIA	Fichier_source
Noms de puits1	Titre1	Auteurs1	Date1	Référence_interne1	Source_émettrice1	eDoc1	BIA1	FILIALE
Noms de puits2	Titre2	Auteurs2	Date2	Référence_interne2	Source_émettrice2	eDoc2	BIA2	FILIALE
Noms de puits3	Titre3	Auteurs3	Date3	Référence_interne3	Source_émettrice3	eDoc3	BIA3	SIEGE

Format du rapport d’analyse :

Le rapport d’analyse contient, en plus des caractéristiques du document, le taux de similarité globale pour chaque cluster et le vecteur de qualité (similarité attributaire moyenne) pour chaque cluster, comme dans l’exemple suivant, pour le rapport d’analyse.

Si on réalise une classification par groupes, alors des groupes de similarité seront créés quelle que soit la source de la donnée et quel que soit le nombre d’éléments du groupe, tant qu’ils sont considérés similaires par rapport aux facteurs de tolérance.

Si on réalise une classification par couples, seuls des couples de données similaires SIEGE-FILIALE seront créé

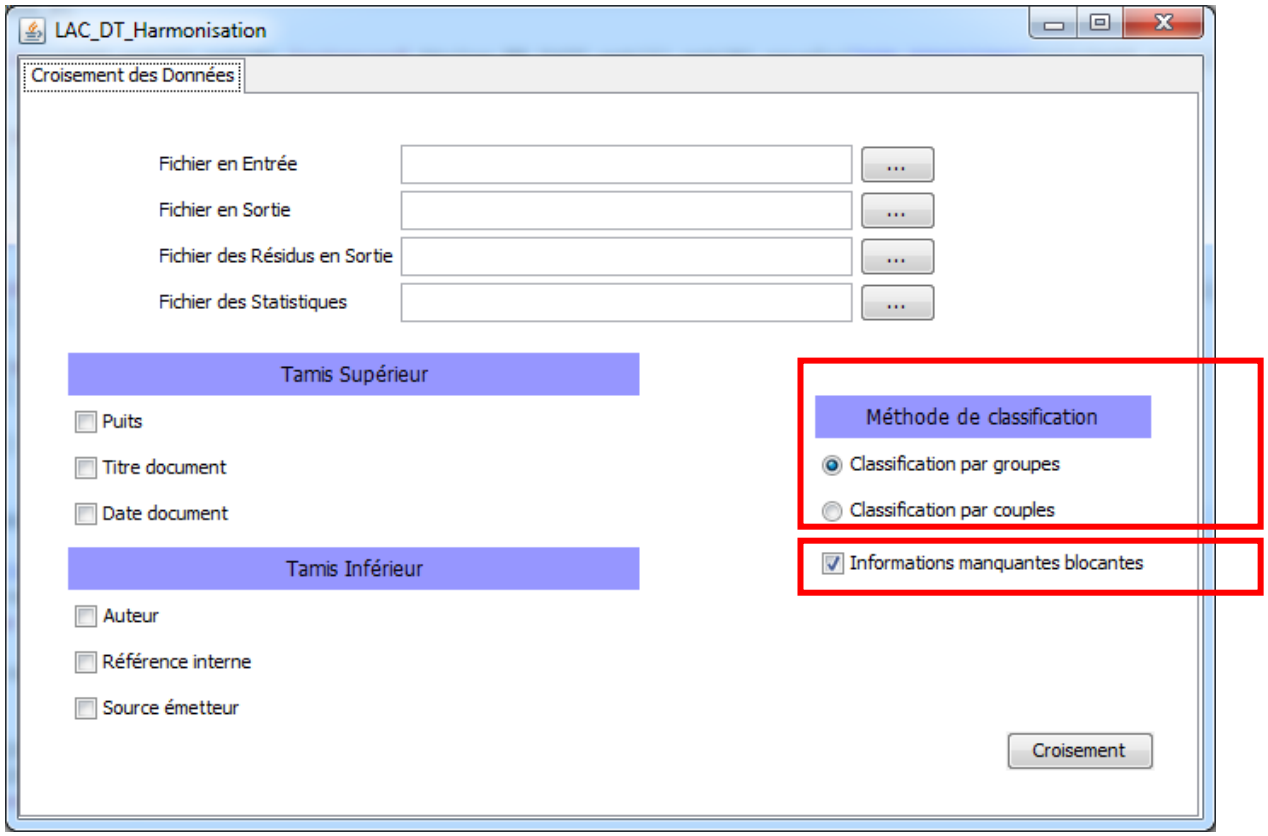
Taux de Similarité	Vecteur de qualité	Cluster ID	Puits	Titre Document	Date	Auteur	Référence Interne	Source Emetteur	eDoc	BIA	Fichier Source
100.0	[100.0, 100.0, 100.0, 100.0, 100.0, 1.0]	1	"[SARAJEH-3,SH003,SH-3]@[ALBORZ-6,ARZ006,AR-6]	ETUDE BIOSTRATIGRAPHIQUE EN LAMES MINCES D' ECHANTILLONS DE CAROTTES DE LA FORMATION DE QUM DES SONDAGES YORTEH SHAH , TALKEH , SARAJEH 3 , SORKEH , ALBORZ 6 (IRAN)	4900	100	WB061393	FORAMINIFERE	TITLE	NON_REN SEIGNE	SIEGE
100.0	[100.0, 100.0, 100.0, 100.0, 100.0, 1.0]	1	"[SARAJEH-3,SH003,SH-3]@[ALBORZ-6,ARZ006,AR-6]	ETUDE BIOSTRATIGRAPHIQUE EN LAMES MINCES D' ECHANTILLONS DE CAROTTES DE LA FORMATION DE QUM DES SONDAGES YORTEH SHAH , TALKEH , SARAJEH 3 , SORKEH , ALBORZ 6 (IRAN)	4900	100	WB061393	FORAMINIFERE	TITLE	NON_REN SEIGNE	FILIALE
100.0	[100.0, 100.0, 100.0, 100.0, 100.0, 1.0]	2	"[SARAJEH-3,SH003,SH-3]@[ALBORZ-6,ARZ006,AR-6]	NOTE SUR LA REINTERPRETATION DE LA STRUCTURE DE POINTE-WEZE A LA SUITE DU FORAGE DE WZ.1	4900	100	WB061393	FORAMINIFERE	TITLE	NON_REN SEIGNE	SIEGE
100.0	[100.0, 100.0, 100.0, 100.0, 100.0, 1.0]	2	"[SARAJEH-3,SH003,SH-3]@[ALBORZ-6,ARZ006,AR-6]	NOTE SUR LA REINTERPRETATION DE LA STRUCTURE DE POINTE-WEZE A LA SUITE DU FORAGE DE WZ.1	4900	100	WB061393	FORAMINIFERE	TITLE	NON_REN SEIGNE	SIEGE

Deux autres fichiers de sortie sont créés :

- un fichier des résidus, c'est-à-dire des documents qui n'ont trouvé aucune référence homologue.
- Un fichier de statistiques tel que l'exemple ci-après.

Nombre total de redondances	Documents uniques Siège	Documents uniques Filiale
14	6	0

Spécificités de traitement : Deux méthodes de classification possibles : classification par groupes ou classification par couples. On peut également spécifier le statut des informations manquantes. En effet, lorsqu'un champ n'est pas rempli, il s'agit de savoir si un texte non renseigné est exactement similaire ou exactement différent d'un texte renseigné. Si on souhaite que les informations non renseignées soient considérées comme distinctes des informations à renseigner, alors il faut cocher la case « Informations manquantes bloquantes ».



LAC_DT pour le géo-référencement : comparaison d'une liste de mots IHS avec une liste de mots e-Search

Remarque : IHS et eSearch sont deux systèmes de stockage des informations relatives à la documentation technique.

Formats d'entrée :

Aussi bien le fichier IHS que le Fichier eSearch doit être formatés selon l'exemple suivant. Ce format nécessite une ligne d'en-tête (une seule). La première colonne comporte le mot IHS ou eSearch. Dans l'en-tête on peut spécifier s'il s'agit de champs, de bassins, de permis… La seconde ligne spécifie la source du mot. Ici, un fichier à traiter doit avoir une même source. Ainsi, dans l'interface de paramétrage, on devra entrer deux noms de fichiers en entrée : celui du fichier IHS et celui du fichier eSearch.

Exemple pour un fichier IHS

Champ IHS	Source
Ablette Marine 1	IHS
Ablette Ouest Marine 1	IHS
Aigle 1 Bis	IHS
Akoum Marin B-1	IHS
Alewana	IHS

Exemple pour un fichier eSearch

hamp_eSearch	Source
ABLETTE MARINE	eSearch
ABLETTE OUEST MARINE	eSearch
AIGLE MARINE	eSearch
AIGLE SUD MARINE	eSearch
ALEWANA	eSearch

Format du rapport d'analyse :

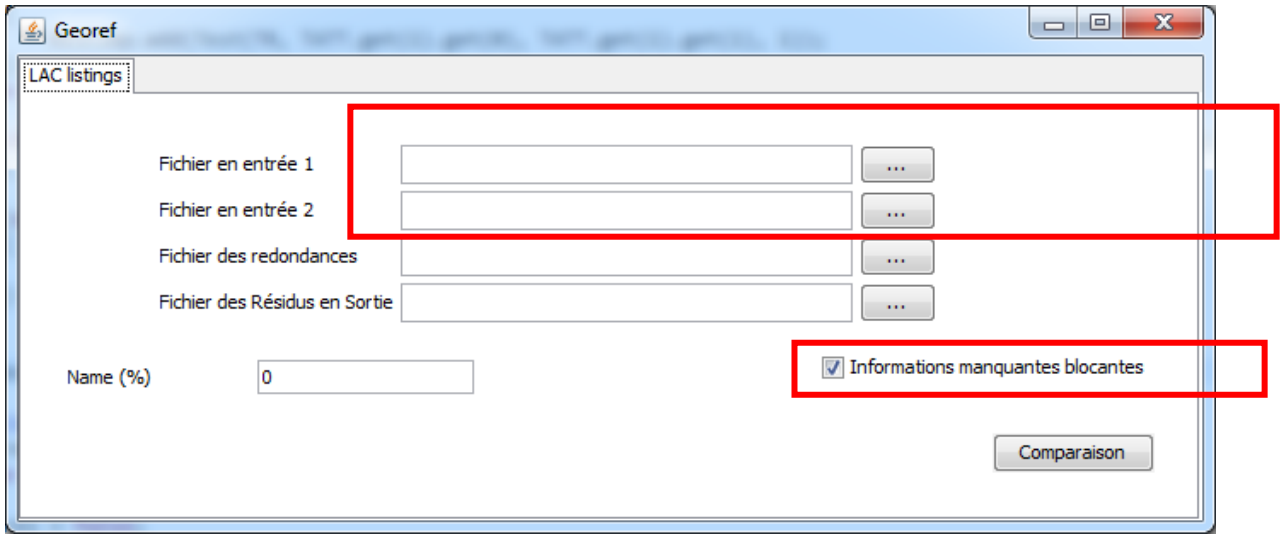
Le rapport d'analyse se présente comme dans l'exemple du tableau ci-dessous.

Cluster ID	Nom	Source
0	ALEWANA	eSearch
0	Alewana	IHS
1	ANGUILLE MARINE	eSearch
1	Anguille Marine	IHS

Un fichier des résidus est aussi créé.

Spécificités de traitement :

Il est nécessaire d'indiquer deux noms de fichiers en entrée, celui du fichier formaté pour LAC et provenant IHS et celui du fichier eSearch. On peut également spécifier le statut des informations manquantes. En effet, lorsqu'un champ n'est pas rempli, il s'agit de savoir si un texte non renseigné est exactement similaire ou exactement différent d'un texte renseigné. Si on souhaite que les informations non renseignées soient considérées comme distinctes des informations à renseigner, alors il faut cocher la case « Informations manquantes bloquantes ».



LAC_DT pour le rattachement : rattachement de noms de puits à des documents techniques

Formats d'entrée :

Il est nécessaire d'exporter les informations eSerach (base de données des documents techniques) ainsi que les noms de puits sur Excel.

Le fichier à traiter doit être formaté comme l'exemple ci-dessous, en mettant dans le même fichier les attributs des références de documents techniques et la liste de noms de puits à côté (l'alignement entre un nom de puits et un document technique n'a pas d'importance). Les informations non renseignées doivent porter la mention **NON_RENSEIGNE**. Il est ensuite nécessaire d'exporter ce tableau en fichier texte, séparateur tabulation, afin de réaliser le traitement dans LAC.

Titre	Mot libre	Doc_ID	Puits:Alias puits
IVORY COAST - WELL B-3X - LISTING OF TWO-WAY TRAVEL TIME AND DEPTH BELOW DATUM OF MEAN SEA LEVEL - VELOCITIES - REFLECTION COEFFICIENTS - TWO-WAY TRANSMISSION LOSS	DRILLING ; PRODUCTION	DO006035	B1-003 ; B1-3X ; ESP B1-3X ; ESPOIR B1-3X
SYNTHETIC SEISMOGRAM REPORT - PUITS A-2X	DRILLING ; LOG ; PRODUCTION	DO006045	A-2X ; A002 ; ESPOIR A 2
CORRELATION DE VITESSES DES PUITS A-8X - K1-X - K1-2X - IVCO-20 220 - 12177 FT	VITESSES	475448	A-8X; A008 ; ESPOIR A 8 IVCO-20 ; IVCO020 ; IVORY COAST 20 ; IVORY-COAST-20 ; IVORY-COAST-OFFSHORE-20 K1-002 ; K1-2X K1-001 ; K1-1X
GEOLOGICAL COMPLETION REPORT IVORY COAST OFFSHORE - IVCO 11	NON_RENSEIGNE	342047	IVCO-11 ; IVCO011 ; IVORY COAST 11 ; IVORY-COAST-11 ; IVORY-COAST-OFFSHORE-11

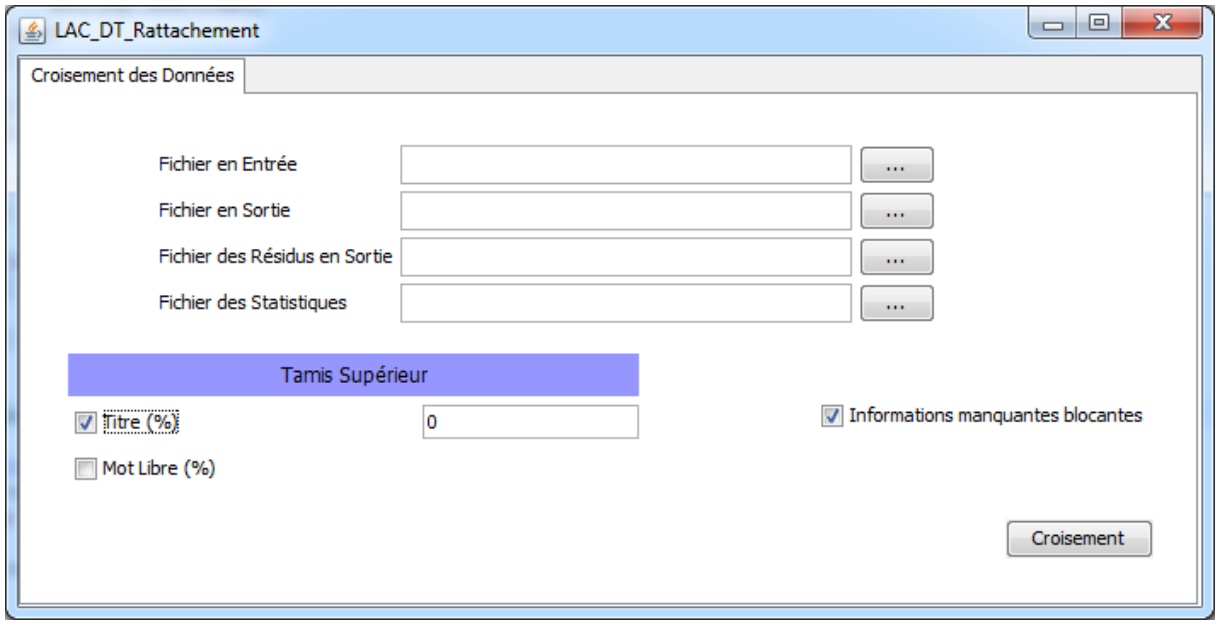
Le rapport d'analyse contient, sur chaque ligne, la référence concernée précédée d'une case contenant les noms de puits qui lui ont été rattachés. Ces noms de puits sont séparés par le caractère spécial « @ »

Puits	Titre	Mot_Libre	DocID	DocID
B-3@B-3X@B003@LB003@	IVORY COAST - WELL B-3X - LISTING OF TWO-WAY TRAVEL TIME AND DEPTH BELOW DATUM OF MEAN SEA LEVEL - VELOCITIES - REFLECTION COEFFICIENTS - TWO-WAY TRANSMISSION LOSS	DRILLING ; PRODUCTION	DO006035	DO006035
A-2X@A002@	SYNTHETIC SEISMOGRAM REPORT - PUITS A-2X	DRILLING ; LOG ; PRODUCTION	DO006045	DO006045
A-8X@A008@IVCO-20@IVCO020@K1-002@K1-2X@IVCO-2@IVCO002@	CORRELATION DE VITESSES DES PUITS A-8X - K1-X - K1-2X - IVCO-20 220 - 12177 FT	VITESSES	475448	475448
IVCO-11@IVCO011@IVCO-1@IVCO001@	GEOLOGICAL COMPLETION REPORT IVORY COAST OFFSHORE - IVCO 11	NON_RENSEIGNE	342047	342047

Un fichier des résidus est aussi créé.

Spécificités de traitement :

Le fichier en entrée doit être le fichier texte formaté pour LAC comme expliqué ci-dessus. On peut également spécifier le statut des informations manquantes. En effet, lorsqu'un champ n'est pas rempli, il s'agit de savoir si un texte non renseigné est exactement similaire ou exactement différent d'un texte renseigné. Si on souhaite que les informations non renseignées soient considérées comme distinctes des informations à renseigner, alors il faut cocher la case « Informations manquantes bloquantes ».



Mode d'emploi de LAC_DP (pour la Données Puits)

LAC_DT pour la comparaison entre la base LogDB et un disque de stockage

Formats d'entrée :

Les fichiers d'export de LDB et du disque doivent être formatés comme dans l'exemple ci-dessous. Il faut mettre la valeur **NON_RENSEIGNE** dans les champs sans information. Il est aussi nécessaire que ces deux fichiers soient en format TXT séparateur tabulation.

Exemple pour un fichier réorganisé issu d'un export de Disque

Name	Name	Path	SEQ Name	Fichier_Source
BB002101V-AA7281-10_15599-MCAL-DEPT-30Nov58.LAS	BB002101V-AA7281-10_15599-MCAL-DEPT-30Nov58.LAS	E:\1.0 Fields' Data\1.1 Bab\Bab Petrophysical\Raw Log Data - Digital\Bab Raw Log Data\	NON_RENSEIGNE	DISK
BB002101V-AA7281-11_15600-MCAL-DEPT-30Nov58.LIS	BB002101V-AA7281-11_15600-MCAL-DEPT-30Nov58.LIS	E:\1.0 Fields' Data\1.1 Bab\Bab Petrophysical\Raw Log Data - Digital\Bab Raw Log Data\	NON_RENSEIGNE	DISK
BB002101V-AA7281-12_15601-MCAL-DEPT-30Nov58.LAS	BB002101V-AA7281-12_15601-MCAL-DEPT-30Nov58.LAS	E:\1.0 Fields' Data\1.1 Bab\Bab Petrophysical\Raw Log Data - Digital\Bab Raw Log Data\	NON_RENSEIGNE	DISK

Exemple pour un fichier réorganisé issu d'un export de LogDB

FILE NAME	SERVICE NAME	FIELD NAME	SEQ NAME	Fichier_Source
JARIM-1-57.LAS_21987	LAS	UNIDENTIFIED	LAS .001	LDB
JARIM-158.LAS_22038	LAS	UNIDENTIFIED	LAS .001	LDB
JARIM-159.LAS_22039	LAS	UNIDENTIFIED	LAS .001	LDB

Format du rapport d'analyse :

Le rapport d'analyse contient, en plus des caractéristiques du document, le taux de similarité globale pour chaque cluster et le vecteur de qualité (similarité attributaire moyenne) pour chaque cluster, comme dans l'exemple suivant, pour le rapport d'analyse.

Si on réalise une classification par groupes, alors des groupes de similarité seront créés quelle que soit la source de la donnée et quel que soit le nombre d'éléments du groupe, tant qu'ils sont considérés similaires par rapport aux facteurs de tolérance.

Si on réalise une classification par couples, seuls des couples de données similaires SIEGE-FILIALE seront créés.

Cluster ID	Nom Fichier	Outil	Path ou Filed name	SEQ Name	Source
0	SY1631-8_77438	PLT	SHAH	N.A.	LDB
0	SY001102S_SY1631-8_77438_PLT-RESMON_PLT-GR-RAW_15-Jun-10_Stations-Logs-Flowing.DLIS	SY001102S_SY1631-8_77438_PLT-RESMON_PLT-GR-RAW_15-Jun-10_Stations-Logs-Flowing.DLIS	E:\1.0 Fields' Data\1.13 Shah\Shah Petrophysical\Raw Log Data - Digital\Shah Raw Log Data\	NON_RENSEIGNE	TOR
1	SY1631-8_77438	PLT	SHAH	LAS .001	LDB
1	SY001102S_SY1631-8_77438_PLT-RESMON_PLT-GR-RAW_15-Jun-10_Stations-Logs-Flowing.DLIS	SY001102S_SY1631-8_77438_PLT-RESMON_PLT-GR-RAW_15-Jun-10_Stations-Logs-Flowing.DLIS	E:\1.0 Fields' Data\1.13 Shah\Shah Petrophysical\Raw Log Data - Digital\Shah Raw Log Data\	NON_RENSEIGNE	TOR
2	SY1631-8_77438	PLT	SHAH	LIS .001	LDB
2	SY001102S_SY1631-8_77438_PLT-RESMON_PLT-GR-RAW_15-Jun-10_Stations-Logs-Flowing.DLIS	SY001102S_SY1631-8_77438_PLT-RESMON_PLT-GR-RAW_15-Jun-10_Stations-Logs-Flowing.DLIS	E:\1.0 Fields' Data\1.13 Shah\Shah Petrophysical\Raw Log Data - Digital\Shah Raw Log Data\	NON_RENSEIGNE	TOR
3	SY1631-9_77439	PLT	SHAH	LAS .001	LDB
3	SY001102S_SY1631-9_77439_PLT-RESMON_PLT-GR-RAW_15-Jun-10_Stations-Logs-Shutin.DLIS	SY001102S_SY1631-9_77439_PLT-RESMON_PLT-GR-RAW_15-Jun-10_Stations-Logs-Shutin.DLIS	E:\1.0 Fields' Data\1.13 Shah\Shah Petrophysical\Raw Log Data - Digital\Shah Raw Log Data\	NON_RENSEIGNE	TOR

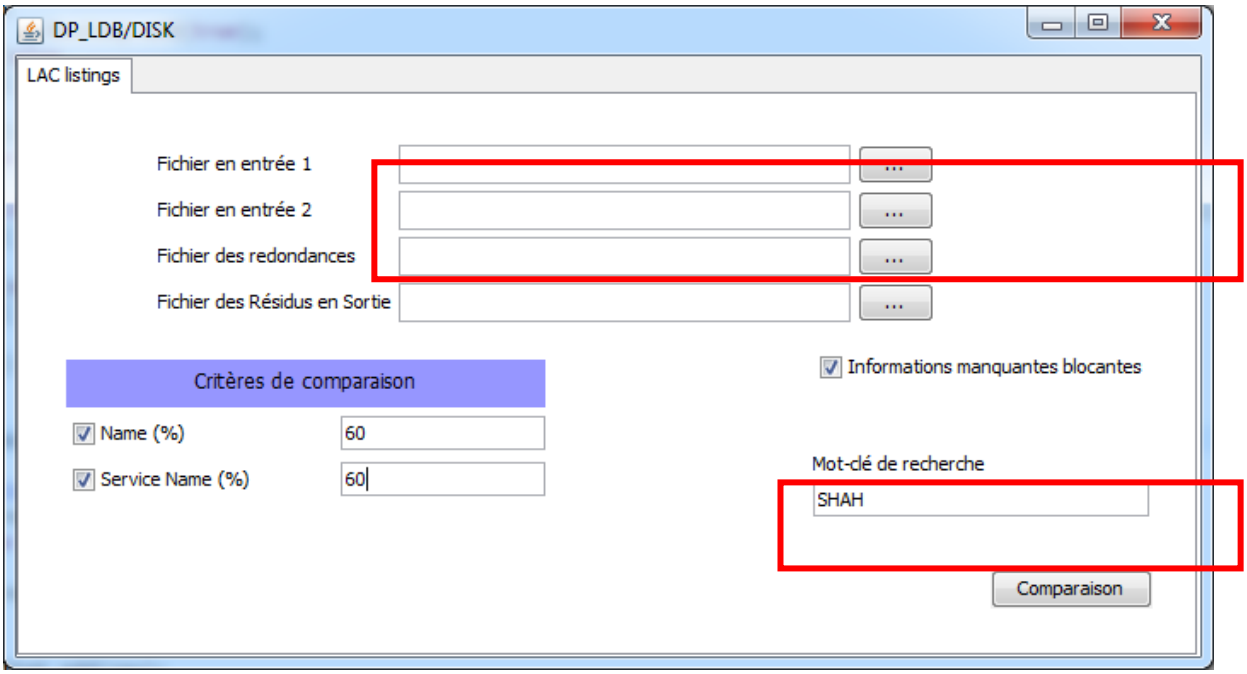
Un fichier des résidus est aussi créé.

Spécificités de traitement :

Les fichiers en entrée doivent être formatés pour LAC comme expliqué ci-dessus. On peut également spécifier le statut des informations manquantes. En effet, lorsqu'un champ n'est pas rempli, il s'agit de savoir si un texte non renseigné est exactement similaire ou exactement différent d'un texte renseigné. Si on souhaite que les informations non renseignées soient considérées comme distinctes des informations à renseigner, alors il faut cocher la case « Informations manquantes bloquantes ».

Il est nécessaire de spécifier le fichier provenant de LDB comme Fichier en Entrée 1, et le fichier provenant du disque en second.

En particulier, on peut réaliser le traitement LAC sur une zone géographique donnée, grâce au filtrage par mot-clé de recherche. Si on souhaite réaliser le traitement sur toutes les zones, il est nécessaire de remplir le champ de mot-clé par **TOUTES_LES_ZONES**.


DP_LDB/DISK
LAC listings
Fichier en entrée 1
Fichier en entrée 2
Fichier des redondances
Fichier des Résidus en Sortie
...
Critères de comparaison
Name (%)
60
Service Name (%)
60
Informations manquantes blocantes
Mot-clé de recherche
SHAH
Comparaison

LAC_DP pour la recherche de doublons dans LogDB

Formats d'entrée :

L'export de LogDB doit contenir les éléments suivants exactement dans cet ordre.

Field Name
Well Name
UWI
Contractor
Activity Date
Activity Type
Activity remarks
Service Name
Service Remarks
Sequence Name
Sequence Remarks
INDEX Type
INDEX Name
TOP DEPTH
BOTTOM DEPTH
INDEX Unit
File Name
Format
Image Size
Tape Remarks
Tape Date Loaded

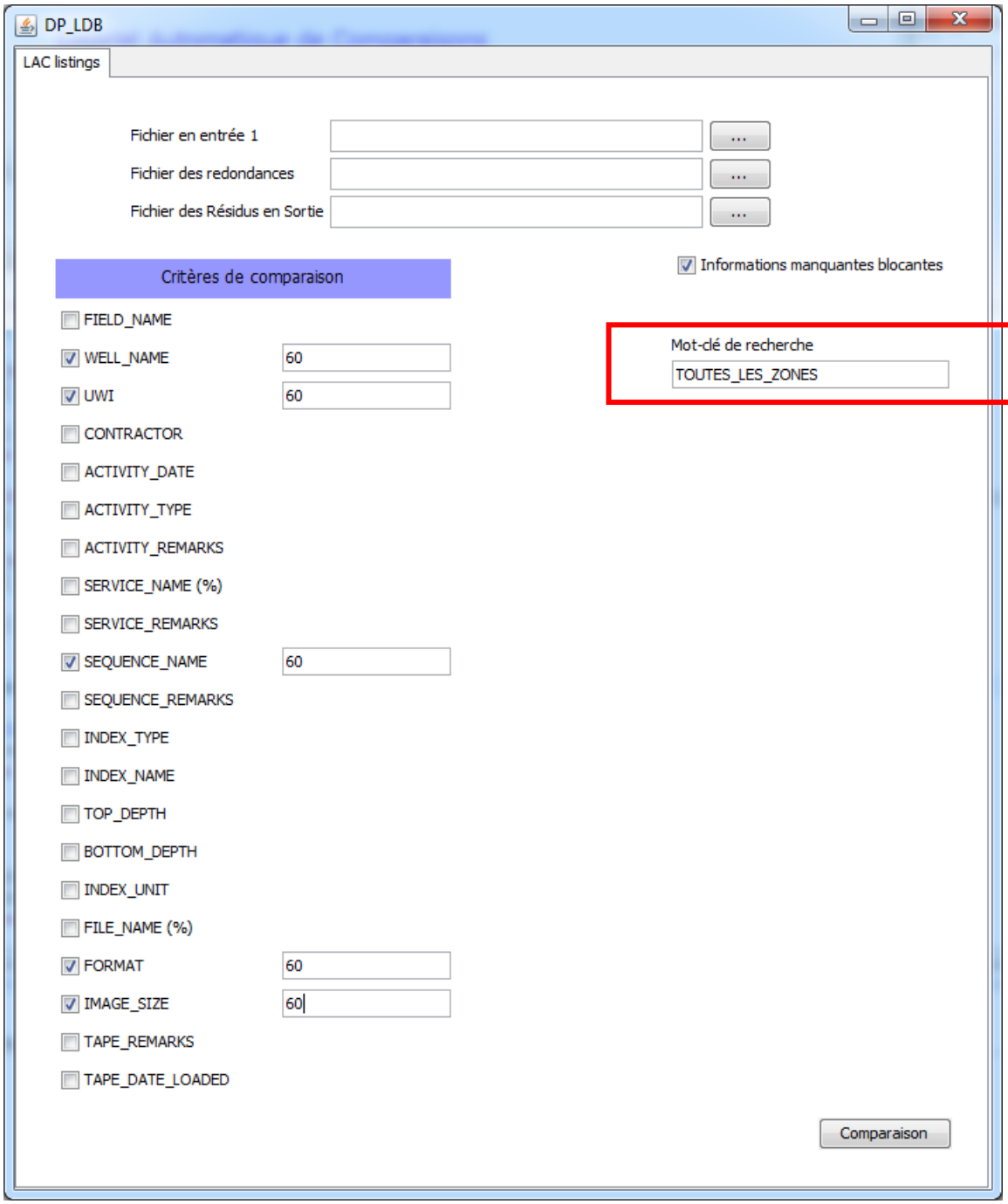
Ces attributs doivent être stockés dans cet ordre sur chaque ligne de fichier. Le fichier doit contenir une ligne d'en-tête spécifiant ces noms d'attributs. Le fichier comportant toutes les informations doit être un fichier texte dont le séparateur est la tabulation.Les informations non remplies devront porter la mention **NON_RENSEIGNE**.

Format du rapport d'analyse :

Le fichier des résultats de la classification a le même format que le fichier d'entrée, en y ajoutant une colonne au début de celui-ci pour indiquer le numéro du cluster auquel appartiennent les lignes.

Spécificités de traitement :

On peut spécifier le statut des informations manquantes. En effet, lorsqu'un champ n'est pas rempli, il s'agit de savoir si un texte non renseigné est exactement similaire ou exactement différent d'un texte renseigné. Si on souhaite que les informations non renseignées soient considérées comme distinctes des informations à renseigner, alors il faut cocher la case « Informations manquantes bloquantes ». On peut aussi réaliser le traitement LAC sur une zone géographique donnée, grâce au filtrage par mot-clé de recherche. Si on souhaite réaliser le traitement sur toutes les zones, il est nécessaire de remplir le champ de mot-clé par **TOUTES_LES_ZONES**.



Remarque : Si on souhaite trouver les doublons de séquence (donc à l'échelle sub-fichier), on coche la case de comparaison des noms de fichier (critère « FILE_NAME »).

Si on souhaite trouver uniquement les doublons de fichiers, alors il est nécessaire de NE PAS cocher le critère de comparaison « File_NAME ».

Table des illustrations

Figure 1 : Illustration des bases de données chez TOTAL. Source : Intranet TOTAL, en 2013 10
Figure 2 : Schéma de rebondissement de l'onde sismique selon la géométrie du terrain. On y représente les shot point SP, les common depth point CDP, géophones et masses émettrices. . 14
Figure 3 : HBDS des composantes géographiques et géophysiques d'une acquisition sismique et de son stockage 15
Figure 4 : Modèle HBDS de la typologie attributaire de l'AMR 18
Figure 5 : Composition attributaire d'une ligne de navigation sismique 19
Figure 6 : Classification des critères de comparaison des lignes de navigation sismique, les catégories sont rangées par fiabilité décroissante pour la comparaison 22
Figure 7 : Schéma représentant le travail de réorganisation des données pour passer de l'ensemble des points sismiques d'un fichier issu de Sismage à la représentation de la ligne de navigation sismique 26
Figure 8 : Schéma représentant le travail de réorganisation des données pour passer de l'ensemble des points sismiques d'un fichier UKOOA à la représentation de la ligne de navigation sismique 27
Figure 9 : Exemple de buffer avec lignes de navigation entières 30
Figure 10 : Exemple de buffer avec lignes de navigation à raccorder entre différents segments de fichier 31
Figure 11 : Exemple de l'en-tête d'un fichier issu de SISMAGE 32
Figure 12 : Interface de réorganisation de fichier 32
Figure 13 : Modélisation d'une ligne de navigation sismique 34
Figure 14 : Exemple des lignes de navigation sismiques ré-échantillonnées 37
Figure 15 : Passage du modèle ancien de la ligne sismique au modèle AMR 39
Figure 16 : Synthèse HBDS de l'AMR 41
Figure 17 : Diagramme des Faits et des Programmes pour la mesure de ressemblance. 43
Figure 18 : Classification des mesures de similarité textuelle 46
Figure 19 : Extrait d'un échantillon de toponymes 47
Figure 20 : Extrait d'un résultat de mesure de similarité sur l'échantillon de toponymes de Cardiff, avec prise de décision manuelle sur l'identification d'un doublon. 48
Figure 21 : Résultat de la comparaison des$(Mesuresk)k\epsilon 17$ mesures de similarité textuelles par Exhaustivité-Précision-Variance sur les échantillons de TO (Ontologie de Toponymes) ... 57
Figure 22 : Arbre de décision des règles de croisement multi critères pour les lignes de navigation 2D 66
Figure 23 : Imbrication des différentes échelles de similarité 69
Figure 24 : Illustration d'un dendrogramme théorique représentant les différentes phases d'agrégation des données (les données étant placées sur l'axe horizontal). 73
Figure 25 : Interface homme-machine du logiciel de comparaisons automatiques (LAC). Cette interface permet de paramétrer les informations nécessaires aux comparaisons et de gérer les attributs lacunaires. 75
Figure 26 : Résultats de comparaison de données concernant la documentation technique associée à des puits de forage, sans gestion des attributs lacunaires (mention « NON RENSEIGNE ») 76
Figure 27 : Résultats de comparaison de données concernant la documentation technique associée à des puits de forage, avec gestion des attributs lacunaires (mention « NON RENSEIGNE ») 77

Figure 28 : Exemple de données à harmoniser. Les lignes surlignées d'une même couleur sont des exemples de doublons. 78
Figure 29 : Exemple de données à rattacher, sur deux sources différentes, l'une correspondant aux données du Siège et l'autre aux données d'une filiale. Le document surligné en violet sera rattaché aux deux puits BAOBAP-WP-11 et BELIER-AO-11. 80
Figure 30 : Exemple de fusion entre deux clusters. Phénomène de contagion. 84
Figure 31 : Représentation de données dans un graphique de similarité, avec les contours de clusters formés selon différents vecteurs de similarité. 85
Figure 32 : Schématisation de mesure automatisée de ressemblance avec une classification en structure de couples. 86
Figure 33 : Schématisation de mesure automatisée de ressemblance avec une classification en structure de groupes asymétriques 87
Figure 34 : Schématisation de mesure automatisée de ressemblance avec une classification en structure de clusters 88
Figure 35 : Schématisation de mesure automatisée de ressemblance avec une classification en structure adaptative 89
Figure 36 : Principe du workflow – Etape 1, un outil de diagnostic 97
Figure 37 : Principe du workflow – Etape 2, un outil de correction 98
Figure 38 : Schéma d'articulation des modules Réorganisation et Comparaisons de LAC 99
Figure 39 : Structure de l'information depuis un point de vue système expert 102
Figure 40 : Architecture de LAC par abstractions concentriques et processus engendrés 106
Figure 41 : Représentation cartographique des données du Brésil sur lesquelles ont été faits les tests de performance de LAC vs ILX. 108
Figure 42: Histogramme des attributs numériques pour l'ensemble des données du Brésil 110
Figure 43 : Relation des attributs numériques entre eux (corrélation attributaire deux à deux), pour l'ensemble des données du Brésil 112
Figure 44 : Valeurs propres en pourcentage d'inertie 114
Figure 45 : Représentation complète des données du Brésil, sur les deux axes de synthèse d'information. 119
Figure 46 : Représentation des vecteurs attributaires constituant des deux axes principaux de représentation. 120
Figure 47 : Représentation graphique du graphe éclaté du jeu de données du Brésil. En bleu les données et en rouge les représentants de clusters. 125
Figure 48 : Tables utilisées dans les algorithmes de calcul du graphe éclaté. 126
Figure 49 : Triangulation de Delaunay de l'ensemble des données du Brésil en incluant les représentants et à échelle réduite sur 20 unités représentées via leurs coordonnées polaires basées sur la mesure de similarité. 127
Figure 50 : Triangulation de Delaunay de l'ensemble des données du Brésil sans les représentants, sur une échelle de 100%, représentées via leurs coordonnées polaires basées sur la mesure de similarité. 128
Figure 51 : Tables utilisées dans les algorithmes de calcul de l'image colorée. 130
Figure 52 : Image 1000*1000 pixels de du jeu de données du Brésil sans inclure les représentants, et à échelle sur 100 unités. 131
Figure 53 : Quelques informations à tirer de l'image d'une base de données. 132
Figure 54 : Exemple de suivi de données depuis le fichier d'entrée jusque dans l'image. Sur le graphe éclaté, en bleu on marque les données très similaires, appartement ici au même cluster. En jaune les autres données. 134
Figure 1 : Schéma HBDS de la méthodologie LAC 149
Figure 2 : Exemple de fusion entre deux clusters. Phénomène de « contagion » 154

Table des tableaux

Tableau 1 : Tableau des algorithmes en fonction des objectifs de mesure textuelle 53
Tableau 2 : Les combinaisons de lettres modifiées par Soundex 55
Tableau 3 : Résultats des tests pour la phase 1 91
Tableau 4 : Résultats des tests pour la phase 2 92
Tableau 5 : Résultats des tests pour la phase 3 92
Tableau 6 : Résultats des tests pour la phase 4 93
Tableau 7 : Performances temporelles de LAC 94
Tableau 8 : Performances temporelles de ILX 94
Tableau 9 : Les mécanismes liés aux pratiques d'intelligence artificielle dans LAC 104
Tableau 10 : Contribution de chaque attribut à la construction de deux axes de composantes principales 115
Tableau 11 : Contribution de chaque attribut au premier axe de représentation, tri en contribution croissante 116
Tableau 12 : Contribution de chaque attribut au second axe de représentation, tri en contribution croissante 116
Tableau 13 : Contribution de chaque attribut à la construction de deux axes de composantes principales 117
Tableau 14 : Contribution de chaque attribut au premier axe de représentation, tri en contribution croissante 117
Tableau 15 : Contribution de chaque attribut au second axe de représentation, tri en contribution croissante 118

Fuga Alba – Thèse de doctorat - 2016

Résumé :

Pour automatiser l'harmonisation des bases de données industrielles de navigation sismique, une méthodologie et un logiciel ont été mis en place.

La méthodologie d'Automatisation des Mesures de Ressemblance (AMR), permet de modéliser et hiérarchiser les critères de comparaison servant de repères pour l'automatisation. Accompagné d'un ensemble de seuils de tolérance, le modèle hiérarchisé a été utilisé comme filtre à tamis dans le processus de classification automatique permettant de trouver rapidement les données fortement similaires. La similarité est mesurée par un ensemble de métriques élémentaires, aboutissant à des scores numériques, puis elle est mesurée de manière plus globale et contextuelle, notamment suivant plusieurs échelles : entre les attributs, entre les données, et entre les groupes. Ces évaluations de la similarité permettent à la fois au système expert de présenter des analyses précises automatisées et à l'expert géophysicien de réaliser des interprétations multicritères en faisant en environ deux jours le travail qu'il faisait en trois semaines.

Les stratégies de classification automatique sont quant à elles adaptables à différentes problématiques, à l'harmonisation des données, mais aussi à la réconciliation des données ou au géo-référencement de documents techniques.

Le Logiciel Automatique de Comparaisons (LAC) est une implantation de l'AMR réalisée pour les services de Data Management et de Documentation Technique de TOTAL. L'outil industrialisé est utilisé depuis trois ans, mais n'est plus en maintenance informatique aujourd'hui malgré son usage. Les nouvelles fonctionnalités d'imagerie de base de données qui ont été développées dans cette thèse n'y sont pas encore intégrées, mais devraient permettre une meilleure visualisation des phénomènes.

Cette dernière manière de représenter les données, fondée sur la mesure de similarité, permet d'avoir une image assez claire de données lourdes car complexes tout en permettant de lire des informations nécessaires à l'harmonisation et à l'évaluation de la qualité des bases.

Ne pourrait-on pas chercher à caractériser, comparer, analyser, gérer les flux entrants et sortants des bases de données, suivre leurs évolutions et tirer des modes d'apprentissage automatique à partir du développement de cette imagerie ?

Mots clés : bases de données industrielles, méthodologie, mesures de similarité, système de filtrage à tamis, classification automatique, harmonisation, imagerie de bases de données, système expert, automatisation, optimisation

Fuga Alba – Thèse de doctorat - 2016

Harmonization of geo-scientific information in industrial data bases, thanks to automatic similarity metrics

Abstract:

In order to harmonize industrial seismic navigation data bases, a methodology and a software have been developed.

The methodology of Similarity Measurement Automation provides protocols to build a model and a hierarchy for the comparison criteria that shall be used as points of reference for the automation. With its tolerance set of thresholds, the model has been used as a scaled filter within the automatic classification process which aim is to find as quickly as possible very similar data. Similarity is measured by combinations of elementary metrics giving scores, and also by a global and contextual procedure, giving access to three levels of results: similarity between attributes, between individuals, and between groups.
Accurate automated analyses of the expert system as well as human interpretations on multiple criteria are now possible thanks to these similarity estimations, reducing to two days instead of three weeks the work of a geophysicist.

Classification strategies have been designed to suit the different data management issues, as well as harmonization, reconciliation or geo-referencing.

The methodology has been implemented in software for automatic comparisons named LAC, and developed for Data Management and Technical Documentation services in TOTAL. The software has been industrialized and has been used for three years, even if now there is no technical maintenance anymore. The last data base visualization functionalities that have been developed have not been integrated yet to the software, but shall provide a better visualization of the phenomena.

This latest way to visualize data is based on similarity measurement and obtains an image of complex and voluminous data clear enough. It also puts into relief information useful for harmonization and data quality evaluation.

Would it be possible to characterize, compare, analyze and manage data flows, to monitor their evolution and figure out new machine learning methods by developing further this kind of data base imaging?

Keywords: industrial data bases, methodology, similarity measures, scaled filter, automatic classification, harmonization, data base imaging, expert system, automation, optimization