Imperfection, temps et espace : modélisation, analyse et visualisation dans un SIG archéologique

THÈSE

présentée et soutenue publiquement le 25/11/2008

pour l'obtention du

Doctorat de l'Université de Reims Champagne-Ardenne (URCA)

(Spécialité Informatique)

par

Cyril de Runz

Composition du jury

Président :	Mme Thérèse Libourel	Professeur des Universités (Informatique) Université Montpellier II
Rapporteurs :	M. Patrice Boursier	Professeur des Universités (Informatique) Université de La Rochelle
	Mme Sophie de Ruffray	Professeur des Universités (Géographie) Université de Rouen
Examinateurs :	Mme Anne Ruas	Directrice du laboratoire COGIT (HDR) Institut Géographique National
	M. Ricardo Gonzalez Villaescusa	Professeur des Universités (Archéologie) Université de Reims Champagne-Ardenne
	M. Eric Desjardin	Maître de Conférences (Informatique) Université de Reims Champagne-Ardenne
Directeurs :	M. Michel Herbin	Professeur des Universités (Informatique) Université de Reims Champagne-Ardenne
	M. Frédéric Piantoni	Maître de conférences (Géographie) Université de Reims Champagne-Ardenne

« This is the Discworld, which travels through space on the back of four elephants which themselves stand on the shell of Great A'Tuin, the sky turtle.
Once upon a time such a universe was considered unusual and, possibly, impossible.
But then . . . it used to be so simple, once upon a time.
Because the universe was full of ignorance all around and the scientist panned through it like a prospector crouched over a mountain stream, looking for the gold of knowledge among the gravel of unreason, the sand of uncertainty and the little whiskery aight-legged swimming things of superstition.
Occasionally he would straighten up and say things like "Hurrah, I've discovered Boyle's Third Law." And every one knew where they stood.

Terry Pratchett, *Annals of the Discworld — Witches Abroad* (1991).

Remerciements

Je souhaite exprimer ma plus grande reconnaissance aux Professeurs Patrice Boursier et Sophie de Ruffray pour avoir accepté de rapporter cette thèse. Je remercie également les Professeurs Thérèse Libourel et Ricardo Gonzalez-Villaescusa ainsi que Madame Anne Ruas pour leur examen attentif de ce travail.

Je souhaite adresser mes plus sincères remerciements à Michel Herbin, Frédéric Piantoni et Eric Desjardin qui m'ont encadré durant ces trois années. Je tiens à témoigner de mon admiration pour Michel Herbin qui par son exigence, sa rigueur scientifique et son ouverture d'esprit a su m'aider à donner corps à ce mémoire. Je tiens de même à manifester mon plus grand respect pour Eric Desjardin qui a accordé énormément de son temps à mon travail et qui par sa patience et son sens didactique, m'a aidé à progresser et à tirer le meilleur de moi-même durant ces trois années de thèse. Je voudrais faire part de mon estime pour Frédéric Piantoni qui a toujours répondu à mes sollicitations lorsque le besoin s'en faisait sentir et qui par son sens de l'interdisciplinarité a été l'initiateur des collaborations au cœur de ce travail de thèse. J'espère que cette thèse sera un témoignage digne du soutien et de la confiance sans cesse renouvelée dont ils ont, tous trois, fait preuve à mon égard.

Je voudrais également signifier aux membres du laboratoire Habiter et aux participants au projet SIGRem mes plus vifs remerciements. Je tiens à témoigner ma reconnaissance à François Berthelot pour m'avoir permis de travailler sur les données de *BDRues*. Je remercie Marcel Bazin et Vincent Barbin pour leur considération envers mon travail ainsi que Dominique Pargny pour le soutien logistique, pour la confiance et pour les longues heures passées ensemble (vive le train et ses pannes). Je souhaite de même exprimer ma plus sincère amitié à Daouda Dianka.

Je tiens à remercier les gens du groupe SIC du laboratoire CReSTIC pour m'avoir accueilli, intégré et ainsi avoir plus que participé à la réussite de ce travail. Je commencerai par les « patrons », Herman Akdag, Yannick Rémion et Laurent Lucas, sans qui je n'aurai pu faire cette thèse. Ils ont tous trois contribués grandement à mon arrivée à Reims, à l'obtention du financement de mon travail de thèse et à mon monitorat. Ils jouent aussi un rôle prépondérant dans la bonne ambiance au sein du laboratoire. Cette ambiance chaleureuse me fut d'une grande aide surtout durant ces derniers six mois.

Pour cette ambiance et le soutien qu'ils m'ont apporté durant ces trois années, je tiens à témoigner de ma gratitude à Stéphanie Prevost, Barbara Romaniuk, Nora Faci, Zahia Guessoum, Eric Bittar, Jérôme Cutrona, Amine Aït Younes, Jean-Michel Nourrit, Laurent Hussenet, Didier Gillard, Jason Mahdjoub et Stéphane Cormier. Un grand merci à Philippe Vautrot pour les longs échanges et les conseils qu'il m'a prodigué. Merci à Gilles Valette, pour m'avoir supporté pendant le stage d'anglais et pour m'avoir soutenu à IPMU (ma première conf). Je ne puis oublier de remercier pour leur amitié Sylvia Piotin, Aassif Benassarou, Loïc Nolin, Olivier Nocent, Antoine Jonquet sans oublier les deux néoSICiens Jessica Prévoteau et Cédric Niquin.

Ne pouvant trouver les mots pour exprimer tout ce qu'a pu m'apporter l'amitié et la collaboration scientifique entretenues depuis trois ans avec Frédéric Blanchard, je lui ai réservé une bouteille :-)

Je voudrais aussi remercier tous les membres de Student Branch IEEE Reims Champagne pour m'avoir témoigné leur confiance en m'élisant deux années de suite comme président. j'ai ainsi pu découvrir l'intérêt et le fonctionnement d'une « société scientifique » majeure en science de l'information. Merci à Francis Rousseau et Janan Zaytoon pour leur soutien permanent aux activités de la branche.

Cela va de soi, je remercie évidemment ma famille pour son irremplaçable et inconditionnel soutien. Ils ont été présents pour écarter les doutes, soigner les blessures et partager les joies. Cette thèse est un peu la leur. Merci à mon père, à ma soeur, à mon frère et à leurs conjoints. Merci à mes grand-mères, à mes tantes, à mes oncles et à mes cousins et cousines. Merci aux petits anges : Alessandro, Lorenzo, Dandéan, Aria, Eliott et Ingrid. Merci à mes potes Steff, Nico, Benoït, Aurélien, Bérengère, Adam, Elodie, Adil, Ouassif, Samira, Younes et Youssef. Merci à Rosalia et Abdelah.

Enfin, Je remercie la région Champagne-Ardenne pour avoir financé ce travail.

Je dédie cet ouvrage à ma mère...

Table des matières

-

Introduction générale

« Every so often, you have to unlearn what you thought you already knew, and replace it by something more subtle. This process is what science is all about, and it never stops. »

Ian Stewart, Jack Cohen et TerryPratchett, *The Science of Discworld* (1999).

Préambule

Au cours des dernières décennies, l'avènement et la démocratisation de l'informatique ont provoqué de profonds changements dans les sociétés modernes. Durant la dernière décade, de par le déploiement à coût modéré d'internet à haut débit, l'influence et l'emploi de l'informatique se sont étendus de manière fulgurante. La plupart des entreprises des sociétés occidentales ont maintenant leurs sites *web* et exploitent des services *web*.

À l'heure où la correspondance « papier » est de plus en plus remplacée par le courriel, où les systèmes de navigation par GPS nous guident dans nos déplacements et où les réunions se tiennent à distance par visioconférences, de nouvelles thématiques, à l'instar de la géomatique, émergent des collaborations entre l'informatique et d'autres domaines tel que la géographie.

Dans ce contexte, la géomatique [1], ou science de l'information géographique, connaît un développement croissant depuis les années 1980. L'apparition de revues thématiques scientifiques à caractère international (*International Journal of Geographical Information Science*, *GeoInformatica*) et francophone (*Revue Internationale de Géomatique*) mais aussi des revues visant un plus large public (*SIG la Lettre*, *GeoMag*) le démontre. La géomatique est aussi un sujet d'intérêt vivant pour la communauté informatique où de nombreuses revues intègrent cette thématique dans leur politique éditoriale (*Soft Computing*, *Knowledge and Information Systems*).

En regroupant l'ensemble des outils et des méthodes informatiques permettant de représenter, d'analyser et de visualiser des données géographiques, la géomatique implique donc au moins trois activités distinctes : collecte, traitement et diffusion des données.

Dès 1994, pour Claude Ecobichon, « les outils que propose [...] la géomatique peuvent améliorer très sensiblement l'analyse et la gestion des territoires, espaces régionaux et mondes urbains » [Ecobichon, 1994]. Depuis, ce domaine des sciences a beaucoup progressé, et il constitue désormais une thématique de recherche pluridisciplinaire orientée vers l'acquisition, l'intégration, la modélisation, l'analyse et la visualisation de données spatialisées. L'information étudiée est gérée via les Systèmes d'Information Géographique (SIG), aussi appelés au Canada Systèmes d'Information à Références Spatiales.

Les SIG proposent des méthodes toujours plus performantes pour des applications dans des domaines intégrant des impératifs de localisation spatiale et d'évaluation territoriale. Une vision globale des outils, des enjeux et des problématiques de la géomatique et des SIG est présentée dans [Longley *et al.*, 2005b], [Burrough et McDonnell, 1998]. Une approche à but pédagogique est proposée dans [Longley *et al.*, 2005a].

La géomatique est fortement liée au développement des thématiques informatiques de la représentation, de la modélisation et de la visualisation de données [Longley *et al.*, 2005c]. Parmi les domaines concernés par ces thématiques, l'intelligence artificielle et le traitement d'images occupent une place essentielle ; les journées *Information Géographique et*

1. Géomatique : en anglais *Geomatics*, *Geoinformatics* ou *GIScience*.

Observation de la Terre organisées conjointement par les groupes d'animations du CNRS, le GdR SIGMA [2] et le GdR I3 [3] montrent l'intérêt des liens disciplinaires transversaux. De même le GdR ISIS [4] organise régulièrement des journées dédiées à la télédétection montrant l'intérêt de la communauté informatique pour la manipulation de l'information spatialisée.

Depuis les années 1980, la volonté d'utiliser la puissance des SIG au sein d'équipes en archéologie se fait jour (voir [Lock et Stancic, 1995] et [Mehrer et Wescott, 2005]). James Conolly et Mark Lake exposent l'intérêt d'un tel outil pour l'archéologie [Conolly et Lake, 2006]. En France, notamment en archéologie urbaine, le recours aux SIG connaît un développement croissant comme l'illustre la création en 2001 du réseau Information Spatiale et Archéologie [5], dévolu à l'exploitation et à la diffusion de la géomatique appliquée à l'archéologie.

En effet, pour gérer les données spatiales, pour associer le traitement des données aux plans, et pour arriver à de futurs logiciels de saisie de terrain, le besoin d'outils permettant aux archéologues de lier la cartographie, la topologie et les bases de données s'avère impératif. Ainsi, en intégrant les bases de données géographiques [6] (BDG), les SIG s'imposent comme une méthode classique d'archivage, de structuration, de visualisation et de prospective de données issues des fouilles archéologiques permettant ainsi la constitution et l'exploitation du patrimoine numérique.

Le besoin d'informations archéologiques spatialisées s'inscrit dans le contexte des mutations des métropoles européennes. En effet, face aux restructurations des espaces urbains (réhabilitation, nouveaux modes de transports), la valorisation du patrimoine ressort comme un enjeu majeur pour les villes (Bordeaux, Toulouse, Reims) qui l'associent à leurs politiques de développement (appropriation citoyenne, identité, image). Aussi l'archéologie urbaine, en termes scientifiques et institutionnels (INRAP [7], SRA [8]), est aujourd'hui sollicitée par les collectivités territoriales et, plus largement, par les opérateurs d'aménagement du territoire.

Dans la problématique de la valorisation et de la gestion du patrimoine archéologique champardennais, la démarche développée conjointement par l'URCA (Université de Reims Champagne-Ardenne), l'INRAP et Ministère de la Culture et de la Commu-

2. GdR SIGMA : Groupe de Recherche Systèmes d'Information Géographique — Méthodologies et Applications dit « Cassini ».

3. GdR I3 : Groupe de Recherche Information — Interaction — Intelligence.

4. GDR ISIS : Groupe de Recherche Information, Signal, Images et viSion.

5. Réseau ISA : Réseau ayant pour objet de réunir des géographes et des archéologues et qui se définit autour de l'application à l'archéologie des outils de la géomatique, nouveaux outils de cartographie, systèmes d'information géographique, télédétection.

6. Base de données géographiques ou à références spatiales.

7. INRAP : Institut National de Recherches Archéologiques Préventives.

8. SRA : Service Régional de l'Archéologie.

nication dans le CIRAR [9] peut être considérée comme novatrice par l'intégration de la géomatique au cœur de l'analyse urbaine et régionale.

Au-delà de l'élaboration de la cartographie archéologique de la cité des Rèmes [10], le projet *SIGRem* [Pargny et Piantoni, 2005b], soutenu par la région Champagne Ardenne, l'état et la ville de Reims, s'est appuyé sur la mise en place d'un SIG pluridisciplinaire [Piantoni, 2005]. Il relève d'une ambition scientifique puisant ses outils conceptuels dans la recherche fondamentale, ses méthodes opérationnelles dans les outils informatiques en matière d'analyse spatiale et son application pratique dans la mise en valeur des données archéologiques recueillies durant les vingt dernières années [Piantoni *et al.*, 2005]. Il constitue le cadre applicatif de ce travail de thèse.

Ainsi, l'émergence du projet *SIGRem* et la création du *Centre Image* [11] ont naturellement conduit l'URCA à la naissance d'une collaboration autour des SIG entre son laboratoire d'informatique (CReSTIC [12]) et son laboratoire de géographie (Habiter [13]). Cette thèse en interface financée par la région Champagne-Ardenne se positionne dans la démarche scientifique de cette collaboration.

Le projet *SIGRem* affiche les objectifs suivants : la conception du SIG, le développement d'une interface interactive de saisie et d'analyse, la valorisation du patrimoine archéologique rémois, la prévention des risques archéologiques, la modélisation, l'analyse et la visualisation des données archéologiques dans le SIG en vue de processus de simulation et d'élaboration de scénarios prospectifs [Pargny et Piantoni, 2005a]. La modélisation, l'analyse et la visualisation des données font l'objet de ce travail.

Les données archéologiques, à l'instar des données géographiques, sont complexes : spatiotemporelles [14], multimodales [15] et aussi par nature imparfaites [16]. La modélisation des données avec leurs imperfections est un des objectifs de ce mémoire.

Afin d'être un outil opérationnel pour les archéologues, les SIG dédiés à l'archéologique doivent comporter des outils d'analyse des données entreposées. Ainsi, une fois les données archéologiques modélisées avec leurs imperfections, il est nécessaire de proposer des processus d'analyse spatiotemporelle de ces données. Ces outils peuvent être des processus d'interrogation ou des processus d'exploration des données stockées. Les méthodes

9. CIRAR : Centre Interinstitutionnel de Recherches Archéologiques de Reims auquel participent l'INRAP Reims, le SRA Champagne et les laboratoires GEGENA, CERHIC, Habiter et CReSTIC.

10. Cité des Rèmes : Reims et ses environs à l'époque romaine.

11. Centre Image : projet pluridisciplinaire ayant pour objectif de développer la recherche autour de l'imagerie numérique à l'URCA.

12. CReSTIC : Centre de Recherche en Sciences et Technologies de l'Information et de la Communication — EA 3804, URCA.

13. Habiter — EA 2076, URCA.

14. Données spatiotemporelles : données qui ont sont définies dans l'espace et dans le temps.

15. Données multimodales : données issues de différents supports.

16. Données imparfaites : données sujettes à de l'incertitude, de l'imprécision, de l'ambiguïté et/ou de l'incomplétude.

d'analyse proposées dans cette thèse portent sur les deux types de processus.

Enfin, les SIG forment des outils permettant l'exploitation visuelle de données. Dans le cadre de l'archéologie, obtenir des visualisations permettant une lecture intuitive des données modélisées est un défi pour lequel je proposerai une approche logicielle intégrable dans les SIG.

Ainsi, en interface entre l'informatique et la géographie, cette thèse participe au projet pluridisciplinaire *SIGRem* — archéologie, histoire, géographie et informatique — afin de proposer une démarche scientifique permettant la modélisation, l'analyse et la visualisation de l'information archéologique dans sa complexité au sein d'un système d'information géographique.

Contexte : l'information et son traitement

La géographie et l'archéologie utilisent des informations localisées sur la Terre et sont attachées à l'identification d'objets matériels.Dans une approche SIG, ces informations sont stockées dans une base de données et sont visualisées dans le système à l'aide de méthodes de cartographie [17].

La problématique de l'acquisition et de la production de cartes thématiques est à l'origine de la création de logiciels de cartographie numérique, précurseurs des SIG. Cependant, à la logique de représentation thématique de la cartographie numérique, s'ajoute, dans les SIG, la logique de l'analyse spatiale. En effet, dans la compréhension des structures de l'organisation des territoires, les corrélations entre objets géographiques doivent être mises en évidence et codées numériquement. Par ailleurs, dans l'espace, les liens entre objets (ou classe d'objets) impliquent un référentiel géographique commun. La localisation des informations est essentielle : à la différence d'une approche en cartographie numérique, elle répond à des requêtes en proposant, de fait, une aide à la décision.

Ainsi, les SIG permettent, en plus de l'aspect cartographique, de stocker, de traiter, d'analyser des données à références spatiales et forment des outils d'aide à la décision. C'est dans ce contexte que les travaux de recherche présentés dans ce mémoire ont été effectués.

Les Systèmes d'Information Géographique

La définition d'un système d'information géographique varie selon l'objectif de son utilisation comme l'indiquent à la fois [Bordin, 2002] et [Burrough et McDonnell, 1998]. Selon Peter Burrough et Rachael McDonnell un SIG peut être défini comme :

17. « la carte est une représentation géométrique conventionnelle, généralement plane, en position relatives, de phénomènes concrets ou abstraits, localisables dans l'espace ; c'est aussi un document portant cette représentation ou une partie de cette représentation sous forme d'une figure manuscrite, imprimée ou réalisée par tout autre moyen » Comité français de Cartographie, 1990.

– une boîte à outils permettant la collecte, le stockage, le traitement, la transformation et l'affichage de données spatiales issues du monde réel dans un objectif particulier [18],
– un Système de Gestion de Base de Données (SGBD) où la plupart des données sont indexées spatialement et où un ensemble de procédures est disponible afin de répondre à des requêtes sur l'aspect spatial (***i.e.*** localisation, géométrie, relation topologique) des données [19],
– un service d'aide à la décision permettant l'intégration de données à références spatiales dans un environnement pour la résolution de problèmes [20].

Des approches proposant des définitions plus larges d'un SIG existent. Par exemple, « un Système d'Information Géographique, ou SIG [...] est un système informatisé qui comprend une base de données sur un ensemble d'unités géographiques, et un logiciel ou un ensemble de logiciels permettant de gérer le stockage, la mise à jour, un accès efficace (facile, rapide, et sûr) aux informations, le traitement et la représentation visuelle de ces données » [Béguin et Pumain, 2003]. Selon [Bordin, 2002], un SIG est « un outil informatique permettant d'effectuer des traitements divers sur des données à références spatiales ».

Dans le contexte d'application des SIG en géographie tout comme en archéologie, deux fonctions essentielles sont développées :

– la gestion d'une base de données spatiotemporelles permettant la corrélation pertinente entre couches d'informations (classification et information thématique), cette fonctionnalité étant très utilisée dans les collectivités territoriales ;
– l'évaluation prospective, qui, dans le système, peut passer par une étude topologique ; cet aspect, qui nécessite une méthodologie orientée vers l'analyse spatiale, est d'une grande richesse dans l'évaluation des phénomènes spatiaux, bien qu'il soit, toutefois, moins exploité.

Dans ce mémoire, les travaux de recherche portent sur la modélisation, l'analyse et la visualisation des données spatialement géo-référencées. La conceptualisation des SIG est plutôt celle de [Bordin, 2002] voire celle de [McDonnell et Kemp, 1996] pour lesquels « un SIG est un système informatique permettant l'acquisition, la gestion, l'intégration, la manipulation, l'analyse et la visualisation de données spatialement référencées sur la Terre » [21].

Les SIG n'y seront pas abordés selon la problématique scientifique de la gestion de bases de données spatiales, ni sur l'optimisation du traitement des données. Ils ne sont

18. *"a powerful set of tools for collecting, storing, retrieving at will, transforming and displaying spatial data from the real world for a particular set of purposes"* [Burrough et McDonnell, 1998].

19. *"a database system in which most of the data are spatially indexed, and upon which a set of procedures operated in order to answer queries about spatial entities in the database"* [Smith *et al.*, 1987].

20. *"a decision support system involving the integration of spatially referenced data in a problem-solving environment"* [Cowen, 1998].

21. Traduction de *A computer system for capturing, managing, integrating, manipulating, analysing, and displaying which is spatially referenced to the Earth.*

pas envisagés selon les définitions de [Smith *et al.*, 1987] et de [Béguin et Pumain, 2003]. La démarche adoptée dans cette thèse est donc davantage de contribuer à l'amélioration des SIG, que de proposer un nouveau cadre d'utilisation comme service en archéologie. Les concepts manipulés et les méthodes présentées dans ce mémoire visent à contribuer aux possibilités offertes par un SIG à l'archéologue dans sa démarche d'expertise. Les ressources de données se fondent sur de l'information localisée en archéologie, traitant de l'espace et des objets dans les sites de fouilles.

Ainsi, l'information manipulée dans un SIG est généralement géographique et, dans ce mémoire, elle est de plus archéologique.

L'information géographique

Pour Michel Béguin et Denise Pumain, « une information est dite géographique lorsqu'elle se rapporte à un ou plusieurs lieux de la surface terrestre. C'est une information localisée, repérée, ou encore « géocodée » [Béguin et Pumain, 2003]. On peut la formaliser de la façon suivante : un objet i est localisé en un lieu, situé dans le système de coordonnées terrestres x et y, dans lequel on repère la latitude x_i et la longitude y_i de l'objet. La définition de l'objet est complétée par un ou plusieurs attributs z_i qui le définissent ou qui le caractérisent. » Ils ajoutent que ces attributs peuvent non seulement consister dans une simple définition de la nature d'objets concrets mais aussi dans une représentation des données plus abstraites.

De plus, « l'information géographique [...] peut porter plusieurs noms : information géographique, information localisée ou encore information à référence spatiale. La composante spatiale est leur point commun »[Bordin, 2002]. Selon [Denègre et Salgé, 1996], l'information géographique peut être définie comme un ensemble reliant :

- une composante relative à un objet ou un phénomène terrestre décrit par sa nature, son aspect et ses attributs ; cette description peut inclure des relations avec d'autres objets ou phénomènes ;
- et une composante spatiale relative à la Terre, exposée dans un système de référence explicite et portant sur la localisation et la forme géométrique (point, ligne, polygone) de l'objet ou du phénomène.

Ainsi, la composante spatiale est la spécificité de l'information géographique, car elle permet des traitements particuliers par le raisonnement sur les liens entretenus entre les objets, voire la structure des objets eux-mêmes, relativement à l'échelle choisie.

Cependant, deux composantes peuvent être dégagées de la composante relative précédente : la ***composante sémantique*** (aussi appelée descriptive), qui donne une signification aux objets et du sens aux données, et la ***composante topologique***, qui relie les données entre elles et définit leurs imbrications (*c.f.* [Bordin, 2002]).

L'information géographique est donc structurée en trois composantes :

- la ***composante géométrique*** relative à la forme géométrique (point, ligne, polygone) et à la localisation de l'objet,

– la ***composante descriptive*** ou ***sémantique*** relative à la qualification de l'objet,
– la ***composante topologique*** : réseaux, voisinage spatiaux ; elle est relative aux liens qu'entretiennent les objets entre eux et de fait, elle intègre aussi les évolutions de la structure de l'espace en fonction de l'évolution des relations entre les objets.

Les données sont structurées uniquement selon les deux premières composantes, la troisième résulte d'une analyse spatiale ou spatiotemporelle.

En géographie, l'information temporelle est généralement considérée comme un attribut de l'information portée par la composante descriptive de l'information. Cette représentation du temps correspond aux motivations géographiques que sont l'étude des dynamiques et l'analyse des changements d'état dans des séries d'informations temporelles.

En effet, les dynamiques des objets spatiaux, et, à travers eux, des phénomènes, puis l'analyse des changements dans le temps sont des processus largement étudiés par les SIG. Le temps est alors associé au changement d'état ou au mouvement entre des instants donnés. La gestion de l'historique des phénomènes restitue alors la dynamique de ces derniers [Thériault et Claramunt, 1999]. Cette approche repose sur la mise à jour des données [Cheylan *et al.*, 1999].

Cette approche du temps dans l'information géographique diffère de celle nécessaire à la définition de l'information archéologique.

L'information archéologique

L'ambition des archéologues, en mobilisant les SIG, est de « travailler sur les héritages, les inerties, les trajectoires, les dynamiques [des objets] dans la longue durée » [Rodier et Saligny, 2007]. Les bases de données géo-historiques associées à cette approche ont pour objectifs :

– d'offrir une vision spatiotemporelle des phénomènes, c'est-à-dire déterminer la situation à une époque donnée d'une part et l'évolution d'un lieu en fonction du temps d'autre part,
– de conserver les mutations fonctionnelles, temporelles et spatiales d'un lieu,
– d'éviter la redondance des informations.

Dans cet objectif, Xavier Rodier et Laure Saligny proposent de distinguer trois composantes [22] pour modéliser l'information archéologique : la ***fonction***, l'***espace*** et le ***temps***. Dans cette approche, l'information archéologique et l'information géographique n'ont pas la même structure.

En effet, la fonction d'un objet résulte de la description de l'objet par l'expert (mur, rue, etc.). Cette composante est donc descriptive, ce qui correspond à la composante sémantique de l'information géographique.

22. La composante topologique résulte de l'analyse et n'est pas prise en compte dans la modélisation des données

De plus, l'espace concerne toutes les informations géométriques et de localisation de l'objet sur la surface terrestre. Cette composante spatiale est donc analogue à celle de l'information géographique.

La composante temporelle représente la période d'activité [23] de l'objet et non sa date de découverte ou sa date de numérisation ; ces deux dernières sont portées par la composante descriptive comme pour l'information géographique. Cette composante est déterminante dans l'approche de l'objet archéologique.

Ainsi, l'information archéologique et l'information géographique présentent de fortes similitudes ouvrant l'archéologie aux nombreux outils de la géomatique. Cependant, les données archéologiques dictent des modes de représentation adaptés à la nature même de celles-ci. Une étude préliminaire de la nature de ces données et des représentations qui en découlent est un préalable à l'analyse de l'information archéologique dans un SIG.

La nature des données

Dans ce mémoire, les recherches se basent sur une modélisation de l'information fournie dans sa complexité en évitant « les modes simplificateurs de connaissances [qui] mutilent plus qu'ils n'expriment les réalités ou les phénomènes dont ils rendent comptent » comme l'a écrit Edgar Morin dans [Morin, 2005].

Données complexes

Considérer les informations géographiques et archéologiques dans leur complexité conduit à prendre en compte les données dans leurs multiples formes.

La première particularité de l'information archéologique est d'être spatiotemporelle. Cette particularité a pour impact de pouvoir regarder les données soit selon leurs composantes spatiales, soit selon leurs composantes temporelles, soit selon les deux conjuguées. Ceci ajoute donc en complexité de modélisation et de traitement.

De plus, pour récolter les données archéologiques ou géographiques, plusieurs procédés d'acquisition et plusieurs sources d'information sont utilisés. Par exemple, les composantes descriptives/fonctionnelles peuvent être issues d'expertises différentes, et les données spatiales peuvent avoir été acquises à partir d'instruments différents. Il faut donc considérer l'information au regard de ces modalités. Les données archéologiques et géographiques sont donc par essence multimodales.

Enfin, à l'image de l'information géographique, l'information archéologique est de qualité diverse. Les données stockées dans le SIG ne donnent pas une vision parfaite de la réalité. En effet, l'étude de la qualité des données est déterminante pour l'obtention d'analyses de confiance dans les SIG. Pour celles-ci, la précision et la fiabilité du résultat sont à

23. La période d'activité d'un objet est la période durant laquelle l'objet a été utilisé selon la description qui lui est associée.

examiner selon les instruments d'observations utilisés et selon les modes d'expertise tant fonctionnelles que temporelles.

Dans ce mémoire, afin de refuser la simplification des données et de réduire les erreurs d'analyse, l'étude de la qualité des données est un préalable nécessaire à la modélisation de celles-ci.

Qualité des données

La qualité des données géographiques est une information relative à la différence entre les données et la réalité qu'elles représentent. Elle s'appauvrit à mesure que les données et la réalité correspondante divergent. Ainsi, plus la qualité des données est faible, moins ces données nous apprennent sur le monde géographique et moins ce qu'elles nous apprennent a de valeur [Goodchild, 2006]. Il en va évidemment de même pour les données archéologiques.

La problématique de la qualité de l'information est régulièrement associée à la gestion de l'imperfection des données, c'est à dire à la gestion de l'imprécision, de l'ambiguïté, de l'incomplétude et/ou de l'incertitude. En effet, obtenir une représentation des données tenant compte de leurs imperfections permet de réduire l'incertitude des décisions et donc d'augmenter la qualité de ces dernières. Dans cette démarche, la qualité des données est prise en considération dans leur représentation.

De plus, comme « on mesure l'intelligence d'un individu à la quantité d'incertitudes qu'il est capable de supporter » (Emmanuel Kant [24]), la représentation de l'incertitude et son exploitation dans des systèmes d'information forment une thématique majeure de l'intelligence artificielle mais aussi un gage de fiabilité des dits systèmes. De fait, la gestion de l'imperfection (l'incertitude est une forme d'imperfection) des données fait partie de la gestion de la qualité.

Ainsi, l'ouvrage [Devillers et Jeansoulin, 2005], dirigé par Rodolphe Devillers et Robert Jeansoulin, propose une vision de la gestion de la qualité des données géographiques dans laquelle une attention particulière est portée sur le traitement de l'imperfection des données.

L'imperfection

Depuis « le doute est le commencement de la sagesse » (Aristote [25]), la question de l'incertitude et de l'imprécision des phénomènes et des pensées n'a cessé d'être posée.

Comme le mentionne Bernadette Bouchon-Meunier dans [Bouchon-Meunier, 1993], « les connaissances disponibles sur une situation quelconque sont généralement imparfaites, soit parce qu'un doute peut être émis sur leur validité, elles sont alors *incertaines*,

24. Emmanuel Kant (1724, 1804)
25. Aristote (-384, -322) Éthique à Eudème

soit parce que nous éprouvons une difficulté à les exprimer clairement, elles sont alors *imprécises* [...] ainsi, le monde réel apparaît-il à la fois imprécis et incertain ». A cela s'ajoute le fait que « les observations que nous recueillons peuvent donc être incertaines, mais également approximatives ou vagues ».

Dans un deuxième ouvrage [Bouchon-Meunier, 1995], l'auteur distingue un troisième facteur d'imperfection des données : l'incomplétude. Ainsi, on peut définir l'imperfection comme l'incertitude, l'imprécision et/ou l'incomplétude.

Comme énoncé précédemment, l'incertitude est liée à la validité d'une connaissance. Le doute sur cette validité peut avoir plusieurs origines comme la fiabilité de la source, ou les erreurs.

L'imprécision est due au caractère vague ou approximatif de la sémantique utilisée. Le premier cas est dû à l'utilisation de connaissances subjectives. Le second cas est la conséquence de catégories aux limites mal définies.

L'incomplétude peut être issue d'une absence de connaissances ou d'une connaissance partielle sur certaines informations du système. Cependant, pour les données archéologiques, une autre forme d'incomplétude existe : la lacune. En effet, dans certains cas, les objets archéologiques trouvés ne représentent que des fragments des objets auxquels ils appartenaient, ils décrivent donc ces derniers de manière lacunaire.

L'ensemble des théories suivantes permet de formaliser l'imperfection des données : la théorie des possibilités et les ensembles flous, la théorie des fonctions de croyance, les ensembles approximatifs, etc. Elles se regroupent sous la terminologie ***théories de l'imparfait***[26]. Cette thèse se consacre en partie à ces théories.

Les types d'imperfections sont interdépendants. En effet, « les incomplétudes entraînent des incertitudes [...], les imprécisions peuvent de même être associées à des incomplétudes, et elles engendrent des incertitudes au cours de leurs manipulations » [Bouchon-Meunier, 1995].

De plus, de ces dépendances peut naître de l'ambiguïté. L'ambiguïté est issue de problèmes mélangeant à la fois de l'incertitude et de l'imprécision et elle peut être l'effet d'un conflit (ou d'un désaccord) dans l'information, ou encore d'un problème de spécificité des informations utilisées.

En utilisant ces dépendances, l'appellation *théories de l'incertain* a été suggérée pour parler des théories précédentes. Dans ce mémoire, je préférerai la dénomination *théories de l'imparfait.*

Ces théories proposent des représentations mathématiques prenant en compte les imperfections des données, car « la connaissance progresse en intégrant en elle l'incertitude, non en l'exorcisant »(Edgar Morin [27]).

26. L'imparfait est à l'imperfection ce que l'incertain est à l'incertitude
27. Edgar Morin, La Méthode, Seuil, 1977

Les représentations de l'information imparfaites

Les mathématiciens se sont depuis longtemps attachés à la prise en compte de l'imperfection des données. La plus ancienne approche est la théorie des probabilités (objectives, subjectives, etc.). Celle-ci est pertinente pour la représentation de tout phénomène aléatoire que l'on peut approcher de manière quantitative. Elle portera plus sur la représentation de l'incertitude.

Avec l'avènement de l'informatique et l'émergence de l'Intelligence Artificielle, de nouvelles thématiques sont apparues, notamment la prise en compte dans des systèmes complexes de l'imperfection des données dites *réelles*.

C'est dans ce but que Lotfi Zadeh a introduit en 1965 la théorie des ensembles flous [Zadeh, 1965] pour représenter l'imprécision, puis plus tard, dans [Zadeh, 1978], la théorie des possibilités, qui se développa sous l'impulsion de Didier Dubois et Henri Prade et de leurs ouvrages [Dubois et Prade, 1985] et [Dubois et Prade, 1988], pour permettre la transition entre imprécision et incertitude.

Dans une approche différente, Arthur Dempster [Dempster, 1968] puis Glenn Shafer [Shafer, 1976] ont introduit la théorie de l'évidence, aussi appelée théorie des fonctions de croyance ou encore théorie de Dempster-Shafer, qui permet notamment de modéliser l'ignorance.

Plus récemment, la théorie des ensembles approximatifs a été formalisée par Zdzislaw Pawlak en 1991 dans [Pawlak, 1991]. Toutes ces théories permettent de modéliser les données et leurs imperfections de manières complémentaires.

Au cours de ces dernières années, plusieurs typologies autour de l'imperfection et de ses théories ont été formulées [Klir et Yuan, 1995]. Certaines, à l'instar de celle proposée par Peter Fisher dans [Fisher, 2005] [28], se penchent sur les problèmes des données spatiales ou des données spatiotemporelles [Bejaoui *et al.*, 2007] dans le cadre des Systèmes d'Information Géographique.

Les différentes théories de l'imparfait ont amené la recherche à développer plusieurs modèles de représentations de données imparfaites dans les SIG. Ainsi, le programme scientifique GISDATA s'est penché en 1996 sur la question des frontières vagues dans les SIG [Burrough et Frank, 1996], et de nombreux modèles spatiaux ont vu le jour : modèles probabilistes, flous [Fisher, 1996], approximatifs [Clementini et Di Felice, 1996].

A l'instar de la composante spatiale, la composante descriptive (fonctionnelle en archéologie) peut de même être formalisée à l'aide de ces théories, l'imperfection portant alors sur l'information associée [Dupin de Saint-Cyr et Prade, 2006]. La composante temporelle est aussi à formaliser car soumise à de l'imperfection. La datation en archéologie est une composante par essence imparfaite.

Cependant, bien que la modélisation de l'information soit un des objectifs des SIG, l'analyse des données issues de cette modélisation est un axe majeur de la géomatique.

28. La version en français se trouve dans [Fisher *et al.*, 2005]

L'analyse de données spatiotemporelles

L'analyse de l'information spatiale ou spatiotemporelle est un champ en plein essor en science de l'information géographique [Longley *et al.*, 2005a]. Elle étudie les relations et les positionnements des objets.

L'un des objectifs de l'analyse de données spatiotemporelles via des SIG en géographie est la généralisation des données. Celle-ci consiste à étudier la mise en perspective de l'information par le changement d'une échelle fine vers une échelle globale comme, par exemple, l'analyse au niveau d'une ville des relations entres données collectées à l'échelle des quartiers. En archéologie urbaine, cela consiste à analyser des données au niveau d'une ville à partir de celles acquises sur les chantiers de fouilles.

Les méthodes et outils de l'analyse de données spatiotemporelles sont principalement de deux types dans les SIG :

- les processus qui utilisent des informations fournies en entrée ; ces processus peuvent être simples (requête) ou complexes (traitement des données) ;
- les processus exploratoires qui cherchent à extraire des relations entre objets dans un ensemble d'objets ; ces processus ne demandent pas à l'utilisateur de fournir de l'information externe au système.

Dans ce travail, je proposerai des approches adaptées à la spécificité des données archéologiques, cas particuliers en informatique de données spatiotemporelles imparfaites. L'objectif de ce travail est le développement de méthodes et d'outils permettant l'analyse des données archéologiques. Ce travail s'appuie sur le programme *SIGRem* dont les données fournissent les bases de l'étude de cas.

La visualisation de données spatiotemporelles dans un SIG

La visualisation des données rassemble toutes les techniques pour créer des images, des diagrammes ou des animations afin de représenter de l'information. Les techniques de visualisation proposent de faciliter la compréhension humaine des données par l'intermédiaire de sélections, de transformations et de représentations d'un résumé de celles-ci. Ce sont des techniques devant permettre à l'utilisateur de percevoir des formes ou/et des structures liant les données.

Dans les SIG, le principal outil de visualisation est la production de cartes thématiques en fonction d'une analyse. Dans ces cartes, les représentations spatiales des objets étudiés sont visualisées selon une légende adaptée à l'objectif thématique. Elles permettent donc un rendu de l'organisation spatiale des territoires en fonction de cet objectif.

Problématique : l'information imparfaite dans un SIG

Pour permettre la généralisation d'objets archéologiques issus de fouilles urbaines et pour dégager la structure d'une ville antique (Reims), la prise en compte de la nature de

l'information archéologique issue de plusieurs sites de fouilles ainsi que son analyse et sa visualisation dans un SIG sont nécessaires.

La prise en compte du caractère complexe (temporel, spatial et imparfait) des objets archéologiques pose des problèmes notamment d'interprétations et d'utilisations de l'information archéologique dans un SIG. Il faut donc étudier les modes de représentation et de modélisation des objets archéologiques dans leurs complexités afin de pouvoir les analyser et les visualiser.

Le questionnement porte sur la manière d'approcher, de modéliser, d'analyser et de visualiser les données spatiotemporelles imparfaites à travers le cas de données archéologiques. Pour cela, je regarde l'information disponible tant spatialement que temporellement tout en tenant compte de son imperfection. Je propose donc des réponses aux questions suivantes :

- Quelle théorie de représentation et quelle modélisation choisir pour l'information archéologique dont je dispose ?
- Quels outils sont nécessaires à l'analyse de données spatiotemporelles imparfaites en vue de la généralisation de l'information archéologique ?
- Quelle visualisation choisir pour faciliter la compréhension humaine de l'information ?

La première question porte sur l'étude du caractère imparfait des données archéologiques afin d'expliciter la modélisation des données dans le SIG. La seconde question porte sur le positionnement temporel et spatiotemporel des objets à une échelle différente de celle de la collecte mais aussi sur les relations spatiotemporelles entre objets afin de faciliter la mise en perspective de l'information disponible. La troisième question porte sur le mode de visualisation de l'information facilitant l'approche humaine de l'information disponible en tenant compte de son imperfection.

Concepts utilisés

Ces trois questions amènent à la mobilisation de concepts variés.

Répondre à la première question nécessite la catégorisation du caractère imparfait des données à travers une typologie mais aussi l'association de ces catégories avec des théories de représentation de l'information imparfaite. Pour cela je propose d'utiliser une taxonomie liant les formes d'imperfection aux différents formalismes permettant la représentation de données imparfaites.

Répondre à la seconde question implique, pour l'étude du positionnement temporel des objets, la définition d'un indice d'antériorité et l'utilisation d'un graphe orienté pondéré. Pour l'étude du positionnement spatiotemporel, l'utilisation de critères de similarité entre objets est essentielle de même que les rangs induits par ces critères. Pour la mise en relation spatiotemporelle des données, l'exploitation de la transformée floue de Hough [Han *et al.*, 1994] et de la méthode d'appariement de [Olteanu, 2007a] permettent de proposer respectivement des scénarios de structuration de la cité à une époque donnée et

d'associer des données issues de plusieurs sources différentes dans une même structure.

Répondre à la troisième et dernière question correspond à rechercher une technique de visualisation permettant d'appréhender l'ensemble de l'information afin d'en dégager des structures de manière intuitive. Pour cela, l'adaptation au contexte de données spatiotemporelles imparfaites d'une technique de visualisation de grands volumes de données quantitatives paraît pertinente. Je choisis dans ce but d'adapter la technique développée dans [Blanchard *et al.*, 2005a] en lui fournissant en entrée des quantifieurs sur les représentations de l'information archéologique ou encore des critères de similarité.

Ainsi, mon travail porte à la fois sur le choix de la représentation et de la modélisation des données et de leurs imperfections mais aussi sur leurs analyses spatiotemporelles ainsi que sur leurs visualisations.

Les données utilisées

Le cadre applicatif de mon travail est une base de données à références spatiales issues de fouilles archéologiques et contenant l'information relative aux rues romaines à Reims : *BDRues*. Cette base de données, mise à disposition par François Berthelot (SRA Champagne), est accessible grâce au *SIGRem*, programme multidisciplinaire utilisant les données portant sur la période antique et collectées durant les 20 dernières années à Reims [29].

Classiquement, selon [Rodier et Saligny, 2007], il existe sept échelles de lecture des données issues de fouilles archéologiques :

- l'*unité stratigraphique* (p.ex. première couche d'occupation à l'intérieur d'un bâtiment),
- le *fait* (p.ex. un mur),
- la *structure* (p.ex. une pièce),
- l'*élément constitutif* (p.ex. un bâtiment),
- l'*objet urbain* (p.ex. un secteur artisanal),
- l'*espace urbanisé ancien* (p.ex. une ville) et
- l'*aire chrono-culturelle* (p.ex. un réseau de villes).

Dans *BDRues*, les données sont décrites à l'échelle du *fait* et concernent les tronçons de rues de Durocortorum [30].

Les tronçons sont stockés sous la forme d'objets ponctuels localisés de manières planaires qui ont pour principales caractéristiques, en plus de la localisation (partie de l'information géométrique), l'orientation de la rue (une autre partie de l'information géométrique), et la période d'activité du tronçon de rue (information temporelle). De plus, par construction de la base, la composante fonctionnelle de l'information disponible dans *BDRues* est décrite à l'échelle du *fait*, et tous les objets ont pour fonction d'être des tronçons de rues. Ainsi on retiendra qu'un tronçon de rue est représenté dans *BDRues* par un

29. Une localisation des sites de fouilles est présentée en annexe D.
30. Durocortorum : ville de Reims à l'époque romaine, capitale de la province de *Belgique*.

point ayant une orientation (angle) et une période d'activité.

Une information complémentaire à ces données est connue mais non stockée dans la base : à l'échelle de la *structure* les rues romaines de Durocortorum sont linéaires et à l'échelle de l'*espace urbanisé ancien* les rues antiques de Reims forment une trame plus ou moins régulière.

Plan du mémoire

La problématique est la modélisation, l'analyse et la visualisation de données spatio-temporelles imparfaites particulières : les données archéologiques. Cette thèse comprend trois parties de trois chapitres qui sont suivis de l'exposé des conclusions et des perspectives. Les annexes illustrent certains aspects complémentaires aux recherches exposées dans ce mémoire et présentent l'ensemble de mes publications, la bibliographie utilisée, la table des figures ainsi que les listes des tableaux et des algorithmes.

La partie I porte sur la représentation et la modélisation de données spatiotemporelles imparfaites, notamment les données archéologiques de Reims.

Le chapitre 1 est dédié à la description de l'imprécision des données archéologiques. Cette étude m'amène à la proposition d'une typologie de l'imperfection dans le contexte de l'information archéologique dans laquelle l'imperfection est due à l'incertitude, à l'imprécision, à l'ambiguïté et/ou à l'incomplétude. Je distingue ensuite deux formes d'imprécision (le vague et l'approximatif), deux formes d'ambiguïté (le conflit et la non-spécificité) et enfin deux formes d'incomplétude (l'absence et la lacune).

Le chapitre 2 porte sur la représentation formelle de l'information imparfaite. Cette représentation peut se faire à l'aide de plusieurs théories : la théorie des probabilités, la théorie de l'évidence, la théorie des ensembles flous, la théorie des possibilités et la théorie des ensembles approximatifs. Dans ce chapitre, j'expose les concepts et les notions de bases nécessaires à l'utilisation de ces théories.

Le chapitre 3 expose les taxonomies de l'imperfection et les différentes modélisations des données spatiotemporelles imparfaites dans un SIG. La taxonomie présentée dans [Fisher *et al.*, 2005] propose de lier le choix du formalisme à la forme d'imperfection dans le contexte de l'information géographique. Je propose dans ce chapitre d'adapter ce schéma à celui de l'information archéologique en fonction des différents types d'imperfections définis dans le chapitre 1. Ensuite, après avoir étudié les différents modèles spatiaux et temporels associés aux différentes représentations et en suivant la taxonomie proposée, je représente les composantes temporelles (les périodes d'activité) et géométriques (les localisations et les orientations) des objets de *BDRues* par des ensembles flous selon une modélisation faisant correspondre pour chaque objet et à chaque type d'information (la période d'activité, la localisation et l'orientation) un ensemble flou convexe et normalisé. L'issue de cette modélisation est la production d'une nouvelle base de données dénommée

BDFRues.

La partie II s'intéresse à l'analyse de données spatiotemporelles imparfaites dans un SIG archéologique selon des informations externes aux données. Pour cela deux types de méthode d'analyse sont utilisés : l'interrogation et l'appariement. Les chapitres 4 et 5 présentent des processus d'interrogation tandis que le chapitre 6 porte sur l'analyse multi-sources via l'étude de l'appariement de données archéologiques.

Le chapitre 4 porte sur l'interrogation temporelle des données selon l'antériorité à une période en entrée du processus. L'antériorité d'un objet à une période en entrée est difficile à déterminer dans le cadre d'informations représentées par des nombres flous (comme les périodes d'activités). En effet, bien que plusieurs méthodes existent, elles ne répondent pas totalement aux critères nécessaires à l'interrogation recherchée. Dans ce contexte, je propose de définir un indice quantifiant l'antériorité entre deux nombres flous appelé indice d'antériorité. La définition de cet indice se situe dans une démarche analogue au Fuzzy Max Order [Ramik et Rimanek, 1985] et utilise l'indice de Kerre [Kerre, 1982]. Je présente un processus d'interrogation permettant une vision générale en couleurs de la position relative des objets de la base selon la quantification de leur antériorité à la période en entrée du processus obtenue à l'aide de l'indice d'antériorité.

Le chapitre 5 présente l'interrogation spatiotemporelle sur critères de formes prenant une période en entrée. L'objectif est de proposer des scénarios estimant les présences possibles des objets à reconstituer en fonction d'une période en entrée de processus. Pour cela, l'analyse se base sur l'aspect lacunaire de l'information archéologique : les objets présents dans la base ne représentent que des fragments d'objets plus importants ayant une forme géométrique particulière. En donnant en paramètre du processus cette forme, l'interrogation va utiliser le principe de la construction d'accumulateurs selon la transformée floue de Hough [31]. Dans le processus exposé, un accumulateur est construit pour chaque caractéristique considérée. Ces accumulateurs sont ensuite fusionnés afin d'en extraire par une sélection les estimations de présences des objets à l'échelle de la ville. Ainsi, dans cette démarche analytique, une méthode de reconnaissance de forme est utilisée afin de tenter de compléter l'information archéologique.

Le chapitre 6 expose le problème de l'appariement nécessaire à l'analyse de données multi-sources. Pour cela, après avoir décrit le contexte de l'analyse multi-source en géomatique, je présente une étude sur les critères à utiliser dans le contexte de l'information archéologique. J'expose ensuite la méthode d'appariement de [Olteanu, 2007a] qui utilise les critères simultanément et prend en considération l'imperfection de l'information en utilisant les fonctions de croyance. Cette approche me semble pertinente pour l'appariement de données archéologiques. Je propose cependant de calculer les rangs des données à apparier avant la définition des fonctions de croyances afin d'augmenter la fiabilité des

31. La transformée floue de Hough introduite dans [Han *et al.*, 1994] est une adaptation de la transformée de Hough présentée dans [Hough, 1962] pour les lignes et généralisée aux formes géométriques simples dans [Duda et Hart, 1972].

appariements. L'appariement des données est nécessaire à toutes démarches de généralisation utilisant différentes sources d'information pour un même territoire.

La partie III est consacrée à l'exploration et à la visualisation de données spatiotemporelles imparfaites dans un SIG.

Dans le chaptire 7, je dégage les positions temporelles des objets de *BDFRues* afin d'extraire les objets les plus antérieurs et les plus postérieurs. Dans ce but, en calculant la valeur de l'indice d'antériorité pour chaque couple de données de *BDFRues*, je construis un graphe orienté pondéré dont les sommets sont les tronçons de rues et dont les arcs auront pour coût la valeur de l'indice d'antériorité de l'origine de l'arc par rapport à la destination. L'indice temporel d'un objet est égal à la somme des coûts des arcs arrivant dans le sommet associé à cet objet moins celle des coûts des arcs en sortant. L'objet de plus faible indice temporel correspond au tronçon le plus antérieur, tandis que celui de plus fort indice temporel sera le plus postérieur. J'explorerai ainsi ce graphe afin d'en dégager les positions temporelles des objets et des classes de positions temporelles (« plutôt antérieur », « plutôt postérieur », « anté-postérieur ») sur ces objets. L'information temporelle issue de la construction de ce graphe est donc la base de cette analyse exploratoire.

Le chapitre 8 propose une exploration statistique des données dans le but d'extraire les objets les plus représentatifs d'un ensemble de données. Pour cela, je présente une analyse exploratoire des données s'orientant vers la recherche d'éléments représentatifs car déterminer l'élément qui représente le mieux en termes de localisation, date et orientation un ensemble d'objets issus de fouilles archéologiques est important pour la classification des groupes de données. Cette recherche se base sur une statistique de représentativité, le Vecteur de Meilleur Rang Moyen ($VMRM$), basée sur la définition de la représentativité proposée dans [Blanchard, 2005] ; la notion -quantitative- de représentativité est basée sur la transformation par rangs des dissimilarités entre éléments. Cette statistique permet de sélectionner l'élément le plus représentatif d'un échantillon. J'appliquerai cette méthode statistique à la recherche des éléments les plus représentatifs des objets de *BDFRues* selon leurs localisations, leurs orientations, leurs périodes d'activités et les trois conjuguées. Pour *BDFRues*, les objets sélectionnés caractérisent l'ensemble des objets et donc les rues gallo-romaines trouvées à Reims.

Le dernier chapitre (chapitre 9) présente une approche pour la visualisation des objets. L'objectif est de pouvoir visualiser les objets archéologiques à travers une image en couleurs via des quantifications. Cette approche se base sur une méthode de visualisation de données par une image couleur [Blanchard *et al.*, 2005a, Blanchard *et al.*, 2005b] qui associe à chaque donnée multi-composante un pixel d'une image. Pour cela, elle réduit les données grâce à une Analyse en Composantes Principales (ACP) [Rao, 1964] [32] à trois composantes. Enfin, par la transformée inverse de celle de [Ohta *et al.*, 1980], la méthode affecte une couleur à chaque donnée. La spatialisation des données dans l'image est obte-

32. ACP : des revues des techniques d'ACP sont proposées dans [Jolliffe, 1986], [Cardoso et Comon, 1996] et [Hyvärinen, 1999].

nue en utilisant une courbe de remplissage de Peano-Hilbert [33]. L'intérêt de cette méthode est que les données sont organisées dans l'image et que de manière intuitive on arrive à appréhender les différentes classes de données. Pour les données de *BDFRues*, l'ambition est de permettre par ce type d'image une exploration temporelle de la base mais aussi une exploration selon la similarité des objets à un objet sélectionné. Dans l'exploration temporelle, pour chaque objet, on défuzzifie sa représentation temporelle par de multiples techniques afin d'avoir des quantifieurs le décrivant temporellement, et on applique la technique de visualisation sur celles-ci. Dans l'exploration des similarités, on calcule les dissimilarités en fonction des localisations, des orientations et des périodes d'activité de l'ensemble des objets de la base à l'objet en entrée. La méthode de visualisation est ensuite appliquée sur ces dissimilarités. L'image résultant de la visualisation fournit une carte synthétique des objets.

33. Courbe de remplissage d'une zone étudiée par Guiseppe Peano (1858-1932) en 1890 et David Hilbert (1862-1943) en 1891

Première partie

Représentation et modélisation des données imparfaites dans un SIG archéologique

« WITH HIM HERE, EVEN UNCERTAINTY IS UNCERTAIN. AND I'M NOT SURE EVEN ABOUT THAT. »

Terry Pratchett, *Annals of the Discworld — Interesting Times* (1994).

« “MILLION-TO-ONE CHANCES," she said, “crop up nine times out of ten." »

Terry Pratchett, *Annals of the Discworld - Equal Rites* (1983).

Introduction

Dans l'objectif de la généralisation de l'information archéologique, l'étude de la nature des données est capitale à leur modélisation en vue de leur analyse. L'information archéologique est non seulement spatiotemporelle mais aussi souvent imparfaite. La modélisation des objets archéologiques en tenant compte de la complexité de cette information est un préalable à leur analyse.

L'objectif de cette partie est de proposer une démarche pour la modélisation des données archéologiques. Elle est composée de trois étapes. La première étape porte sur la nature de l'imperfection des données. La seconde est consacrée au choix d'une théorie de représentation de l'imparfait. La troisième étape est dédiée à la modélisation des données suivant le formalisme de représentation choisi. Cette démarche doit permettre de préparer les données afin d'en appréhender les complexités durant leur analyse.

On peut observer différents types d'imperfections dans les données archéologiques. À l'instar de l'information géographique, l'information archéologique peut être imprécise, incertaine, incomplète et/ou ambiguë. Mais si les données archéologiques et géographiques sont globalement structurées de la même manière, la qualité des données archéologiques est altérée notamment par le temps séparant l'observation de l'usage. Ainsi, la granularité de la description de leurs imperfections doit être plus fine. Pour cela, on peut distinguer deux sous-types d'imprécision et deux d'incomplétude. Ceux-ci sont, pour l'incomplétude, la lacune des données archéologiques et l'absence d'information dans la base, et pour l'imprécision, le vague et l'approximatif. Afin de tenir compte du caractère imparfait de l'information, je vais, dans le chapitre 1, expliciter ce qu'est l'imperfection de l'information à la lumière de ses différents types.

Les mathématiciens ont élaboré plusieurs théories de représentation des données imparfaites. La plus ancienne, la plus étudiée et la plus répandue est la théorie des probabilités. Elle fut pendant longtemps en France orientée vers la représentation du hasard, c'est-à-dire les probabilités objectives ou statistiques. Le monde anglophone s'était orienté plutôt vers les probabilités subjectives (à l'instar de Bayes). Depuis la fin de la Seconde Guerre Mondiale, l'approche subjective est devenu prédominante (d'après [Shafer, 1976]) dans le monde. D'un autre côté, Zadeh en 1965 a introduit dans [Zadeh, 1965] les ensembles flous qui s'avérèrent une alternative importante aux probabilités pour l'imprécis. Les théories plus récentes qui représentent les données à l'aide de fonctions de croyances, de possibilités ou d'ensembles approximatifs complètent les deux premières dans l'optique de la forma-

lisation de l'imprécis et de l'incertain. Grâce à la théorie des fonctions de croyances ou des possibilités, on peut notamment représenter l'ignorance, tandis qu'avec les ensembles approximatifs on s'abstrait de certaines contraintes de graduation de l'imprécision. J'exposerai ces différentes formalisations dans le chapitre 2.

Enfin le chapitre 3 porte sur le problème du choix de la théorie de représentation des données dans le cadre de données archéologique dans un SIG et les modèles spatiaux et temporels associés à ces choix. Partant de la taxonomie de l'incertain des objets spatiaux proposée par Fisher ([Fisher *et al.*, 2005] et [Fisher, 2005]), j'en propose une extension à l'imperfection adaptée au contexte de l'information archéologique. J'appliquerai cette démarche aux données de *BDRues* dans le cadre du projet *SIGRem*.

Ainsi cette partie se structure en trois chapitres. Le premier porte sur la description de l'imperfection des données, le second sur les différentes solutions théoriques de formalisation de l'imparfait, le dernier sur la modélisation des données dans un SIG archéologique et plus spécifiquement celle des données sur les rues de l'époque romaine à Reims. Cette partie expose une démarche pour la représentation et la modélisation des données archéologiques guidée par l'observation de l'information dans sa complexité.

Chapitre 1

Nature des imperfections de l'information

Sommaire

Les systèmes d'information géographique dédiés à la gestion de l'information archéologique permettent la gestion de grands volumes d'information sur le bâti et l'environnement naturel présents et passés. Ces connaissances sont sujettes à différentes formes d'imperfection. Si cette imperfection est minorée dans la représentation des données, la validité des résultats des processus de généralisation et de mise en relation des objets archéologiques peut être remise en question. Ainsi, l'imperfection de l'information impacte sur la qualité des données et des décisions.

Il est donc essentiel que la qualité des données spatiotemporelles soit étudiée et intégrée au processus d'analyse en fonction de la nature de l'information disponible. Dans ce cadre, la compréhension des différentes formes d'imperfection est fondamentale [Rolland-May, 2000].

Ce chapitre examine la nature conceptuelle des différentes sortes d'imperfection pouvant être attachées à l'information archéologique dans ses aspects spatiaux et temporels. En se référant à [Fisher *et al.*, 2005], on peut distinguer, en étudiant les données, si les

classes d'objets ou les objets suivant sont bien ou mal définis [34]. Dans le cas où l'objet et la classe sont bien définis, on sera soumis à de l'incertitude. Dans les autres cas, l'objet ou la classe sont mal définis et l'imperfection des données sera due à de l'imprécision, à de l'ambiguïté et/ou à de l'incomplétude [35].

Dans ce chapitre, je décrirai chacun des aspects de l'imperfection et en étudierai les différents facteurs possibles. Je conlurai par la présentation d'une typologie de l'imperfection des données archéologiques.

1.1 Incertitude

L'incertitude est liée à la validité d'une connaissance.

Si les dimensions tant spatiale, que temporelle, fonctionnelle ou descriptive d'un objet archéologique sont définissables par des sous-ensembles de l'ensemble des cas possibles, alors chaque cas possible soit appartient, soit n'appartient pas à l'ensemble décrivant l'objet. Ainsi, si l'on considère une position spatiale (un cas spatialement possible), à cette position on se situe ou non dans l'objet. Si l'on considère une année (une position temporelle), l'objet est présent ou non. Les correspondances des objets et des positions peuvent donc être obtenues de manière booléenne par l'intermédiaire de requêtes dans le SIG.

Toutefois, pour une position spatiale ou temporelle, il peut y avoir doute sur la validité du résultat booléen. Le doute sur cette validité peut être de plusieurs origines : fiabilité de la source ou erreurs. De plus, en archéologie, l'aspect fonctionnel de l'objet est issu d'une expertise dont le résultat peut à tout moment être remis en cause. Dans ce cas, la validité de la fonction de l'objet archéologique est en question. Ainsi l'information descriptive de l'information archéologique peut être incertaine.

Pour l'information géographique, [Fisher, 2005] propose l'ensemble, présenté dans le tableau 1.1, des types et causes classiques d'erreurs qui sont sources d'incertitudes dans les données. En suivant la présentation de Fisher, l'incertitude à laquelle est sujette l'information peut être issue d'erreurs de mesure, d'enregistrement, de classement, de regroupement de classes, de généralisation, temporelle et de traitement. Ces erreurs sont pour certaines constituantes de l'information contenue dans le système (mesure, enregistrement, temporelle, classement) et pour les autres de l'exploitation du système (regroupement, analyse, traitement). L'incertitude est alors attachée à l'information se dégageant de l'utilisation du système (données + résultats d'analyses).

Le tableau 1.1 est adaptable aux incertitudes auxquelles l'information archéologique peut être soumise. En effet, on peut regarder la composante temporelle d'un objet archéologique comme la date de validité de l'objet ; cette composante correspondant souvent

34. Un objet bien défini est un objet défini de manière complète, précise et unique. Un objet mal défini est un objet dont la définition est soit imprécise, soit multiple, soit incomplète.

35. Les définitions de l'incertitude, l'imprécision et de l'incomplétude sont extraites de [Bouchon-Meunier, 1993] et de [Bouchon-Meunier, 1995].

Tableau 1.1 – Types et causes d'erreurs qui sont sources d'incertitudes dans une base de données géographiques — d'après [Fisher, 2005]

Types d'erreurs	**Causes de l'erreur**
De mesure	la mesure d'un objet est erronée (localisation)
De classement	l'objet est affecté à une mauvaise classe d'objets à cause d'une erreur dans l'expertise
De regroupement de classes d'objets	l'objet est regroupé avec des objets aux propriétés différentes
De généralisation spatiale	généralisation de la représentation cartographique de l'objet avant numérisation (déplacement, simplification, échelles d'approches)
D'enregistrement	mauvais codage en entrée du SIG
Temporelle	l'objet a changé de caractéristique entre la date d'enregistrement dans la base et celle de son utilisation
De traitement	une erreur est causée pendant le traitement par un arrondi ou une erreur algorithmique

à la période d'activité de l'objet. Ainsi, l'erreur temporelle peut venir d'une utilisation de l'objet (par exemple un mur en activité durant le 1er siècle ap. J.C.) à une époque où il n'était pas en activité (au 3ème siècle ap. J.C.). Cependant l'information contenue dans ces périodes est proche dans sa modélisation de celle contenue dans l'information spatiale [Rodier et Saligny, 2007]. Dans ce contexte archéologique, les incertitudes dues aux erreurs de mesure, de généralisation, d'enregistrement et de traitement peuvent être aussi bien spatiales que temporelles (voir le tableau 1.2).

Ainsi, les données archéologiques peuvent être sujettes à de l'incertitude par l'intermédiaire de plusieurs types d'erreurs. Ces types, qui sont fonction des causes d'erreurs, diffèrent légèrement de ceux proposés par [Fisher *et al.*, 2005] pour l'information géographique. Ces différences prennent essentiellement naissance dans la divergence de l'aspect temporel entre l'information géographique et l'information achéologique.

Ce n'est toutefois pas la seule source d'imperfection dans les données archéologiques. En effet, la période d'activité d'un objet peut être dépendante de la sémantique utilisée pour sa définition. Cette sémantique peut ne pas avoir la même définition en fonction des experts, même si elles sont proches. Ainsi l'objet dont la composante temporelle est exprimée à l'aide d'une telle période d'activité sera mal défini. La proximité des différentes définitions induit de l'imprécision dans la définition de la période d'activité. L'imprécision peut donc aussi être un facteur d'imperfections.

Tableau 1.2 – Types et causes d'erreurs sources d'incertitudes dans une base de données spatiales contenant des objets archéologiques

Types d'erreurs	**Cause de l'erreurs**
De mesure	la mesure spatiale (localisation) ou temporelle (période d'activité) d'un objet est erronée
De classement	l'objet est affecté à une mauvaise classe à cause d'une erreur dans l'expertise
De regroupement de classes	l'objet est regroupé avec des objets aux propriétés quelque peu différentes
De généralisation spatiale ou temporelle	généralisation de la représentation cartographique ou temporelle de l'objet avant numérisation (déplacement, simplification, échelle d'analyse)
D'enregistrement	mauvais codage en entrée du SIG
De traitement	une erreur est causée pendant le traitement par un arrondi ou une erreur algorithmique

1.2 Imprécision

L'imprécision de l'information est due au caractère vague ou approximatif de la sémantique utilisée. Le caractère vague de l'information est la conséquence d'une insuffisance des instruments d'observation et de l'utilisation de connaissances subjectives ou flexibles. L'approximatif est issu de catégories aux limites mal-définies. Contrairement à l'incertitude qui porte sur la validité de l'information et que l'on peut voir d'un point de vue statistique ou probabiliste, l'imprécision est un des défis du domaine de la logique et de la cognition.

L'imprécision des frontières territoriales est de plus en plus étudiée dans l'analyse par la mise en évidence du phénomène de marge. On peut à ce propos citer les recherches menées par Sophie de Ruffray sur la mise en évidence d'un espace de marge dans l'organisation spatiale du Grand Est français [de Ruffray, 1999, de Ruffray, 2004].

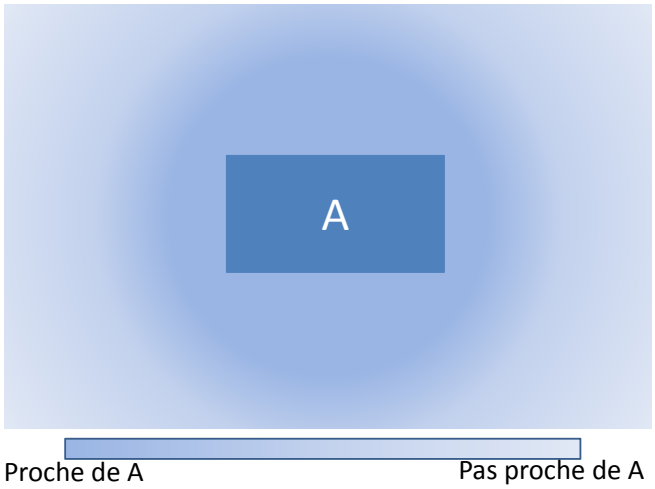
L'imprécision d'une information est facile à mettre en évidence. Par exemple, considérons la catégorie d'humain définie selon la relation « être vieux ». On peut dire qu'un homme est vieux s'il a 100 ans. Est-il vieux s'il n'a que 99 ans ou 98 ans ? Selon l'usage normal, la réponse sera oui. Ainsi, le rajeunissement d'une année de la personne ne semble pas changer un homme vieux en homme jeune. Cependant, est-il pensable de dire qu'un homme de 20 ans est vieux ? Évidemment non. Par contre, peut-on définir une limite différenciant les ages pour lesquels on est vieux ? Est-on vieux à partir de 50 ans, 55 ans, 60 ans, 65 ans, 70 ans ? Aucune limite précise et admise par tous ne peut être émise. La catégorie regroupant les humains vieux ne peut donc pas être définie précisément. L'information associée à cette catégorie est donc imprécise.

1.2.1 Vague

L'information est vague si elle est définie à l'aide de connaissances subjectives ou flexibles, ou si elle est la conséquence d'une insuffisance des instruments d'observation.

L'utilisation de la relation spatiale « proche de » pour le positionnement d'un objet implique du vague. Qu'entend-on par « B est proche de A » ? L'objet B est-il « proche de » A s'il est situé à moins 5 de mètres de A ? Si la réponse est oui, l'est-elle encore s'il est situé à 500 mètres ? Cela dépend du contexte et de l'échelle d'analyse. Cependant quel que soit le contexte ou l'échelle d'analyse, B est plus proche de A s'il est situé à moins de 5 mètres que s'il est situé à 500 mètres. La figure 1.1 propose une zone illustrant cette relation. Cette notion est donc dépendante du contexte et de la personne qui l'utilise. Elle est donc subjective et flexible. Cependant, l'utilisation d'une telle notion par des requêtes dans un SIG est évidement souhaitable. Il faut donc tenir compte de cette caractéristique de l'information.

Figure 1.1 – Exemple de relation spatiale vague : « proche de A »



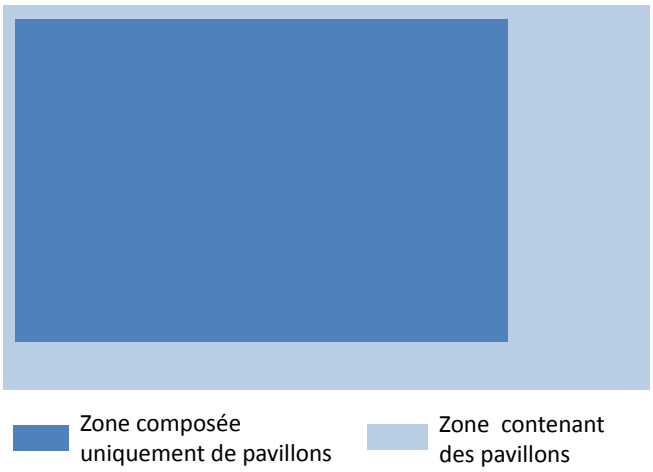
Dans le contexte de l'information archéologique, l'estimation des périodes d'activités des objets archéologiques est soumise à la fois à des connaissances subjectives (estimation issues d'expertises) et à des représentations flexibles (sémantiques à signification variable). Par exemple, prenons un objet dont la période d'activité serait qualifiée par « $17^{\text{ème}}$ siècle », l'activité de cet objet a-t-elle commencé dès la date de début de cette période ou quelques années avant ou après ?

Ainsi, les périodes d'activité des objets archéologiques ne peuvent souvent être définies que d'une manière vague.

1.2.2 Approximation

Le caractère approximatif de l'information provient de catégories aux limites mal-définies, c'est-à-dire d'objets dont la représentation ne peut être délimitée clairement.

Figure 1.2 – Exemple d'objet spatial sujet à de l'approximation : zones d'un quartier pavillonnaire



Sur l'exemple de la figure 1.2, la définition du quartier pavillonnaire englobe nécessairement la zone ne contenant que des pavillons (zone en bleu foncé). Cependant, la zone contenant des pavillons (zone en bleu clair incluant la zone en bleu foncé) est aussi liée au quartier pavillonnaire. Ces deux zones approximent donc le quartier pavillonnaire. Dans ce cas, l'objet quartier pavillonnaire est sujet à de l'approximatif.

En archéologie, il peut arriver que l'on trouve un objet (par exemple un mur) que l'on va définir à la fois grâce à des objets le composant (des pierres) et à des traces qui indiquent la présence de l'objet mais n'appartiennent pas vraiment à l'objet. L'ensemble formé des traces et des pierres et celui formé des pierres forment donc des limites différentes à l'objet. Ces différentes limites ne permettent pas de définir l'objet de manière précise. La définition de l'objet est alors sujette à de l'approximation.

Si l'on reprend l'exemple précédent de la vieillesse, on peut considérer que la cause de l'imprécision de la sémantique « être vieux » est due soit au fait qu'« être vieux » est une connaissance flexible et subjective, soit au fait que la catégorie « est vieux » a des limites mal définies. Cette connaissance est donc imprécise au sens large. Ainsi, la structure de l'imprécision proposée ici ne sépare pas les types d'imprécision de manière disjonctive (vague ou bien approximatif).

À l'imprécision et l'incertitude, on peut ajouter l'ambiguïté et l'incomplétude aux

formes possibles de l'imperfection de l'information. L'ambiguïté est à considérer dans le cas de choix entre de multiples identifications possibles d'un même objet.

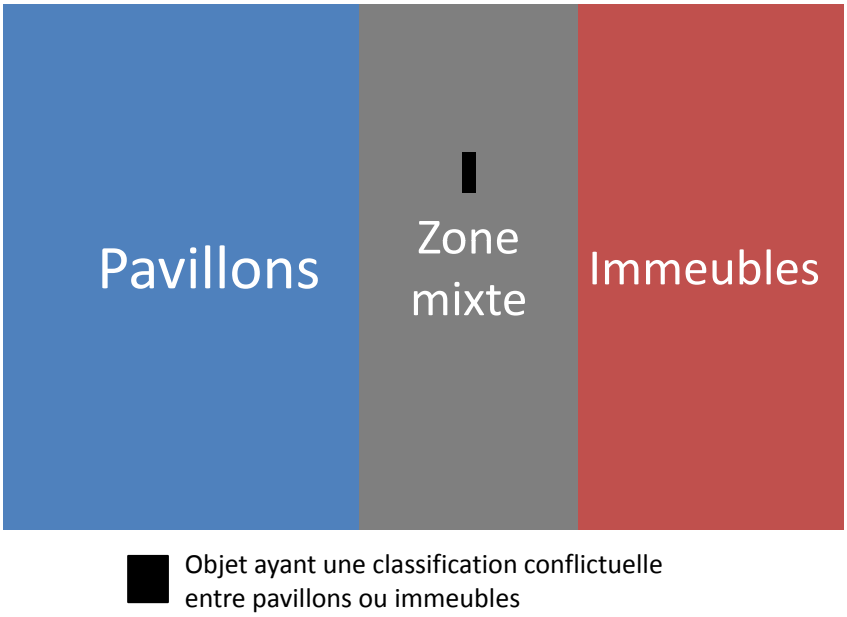
1.3 Ambiguïté

L'ambiguïté survient lorsqu'il y a un doute sur la manière de définir un objet ou un phénomène, c'est-à-dire quand un élément peut appartenir à plusieurs catégories disjointes ou d'échelles différentes, ou encore quand la description de l'élément peut donner lieu à plusieurs sens. Peter Fisher, dans [Fisher *et al.*, 2005], distingue les deux cas d'ambiguïté suivants : le désaccord ou conflit, le manque de spécificité ou non-spécificité.

1.3.1 Conflit

Il y a conflit si au minimum deux classifications contradictoires pour un unique objet sont possibles.

Figure 1.3 – Zones urbaines comme exemple de conflit de données spatiales



L'exemple de la figure 1.3 illustre le cas d'informations spatiales conflictuelles. Dans ce cadre, le choix porte sur la classe (pavillon ou immeuble) à l'aide de laquelle l'objet doit être décrit. Or cet objet se situe sur une zone tampon mélangeant des objets des deux classes. Les informations sont dans ce cas conflictuelles car la définition d'un pavillon est distincte de celle d'un immeuble. La zone tampon est la zone définissant l'imprécision de

chacune des deux classes. Le choix de la classe se fait sur des définitions imprécises des classes d'objets.

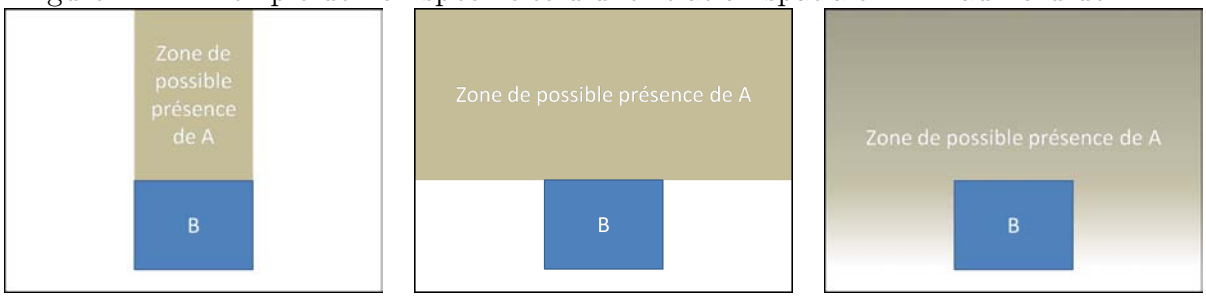
En archéologie, le cas d'un objet, auquel on pourrait attribuer deux représentations temporelles, spatiales ou fonctionnelles différentes et non incluses, illustre une situation conflictuelle. Par exemple, si deux expertises décrivent un même objet comme étant pour l'une un mur, pour l'autre un pavage de rues, la description de l'objet de manière unique amènera à choisir l'une ou l'autre.

La résolution du conflit ou désaccord revient à choisir une approche (à gérer l'incertitude) sur des objets (ou relations) imprécis (ayant plusieurs représentations disjointes possibles). Cependant les sémantiques peuvent ne pas être disjointes. C'est le cas d'informations non-spécifiques.

1.3.2 Non-spécificité

Lorsqu'une définition d'une relation ou d'un objet peut amener à plusieurs sens, ou lorsque l'échelle de l'analyse est susceptible d'amener à de multiples interprétations, on parle alors de non-spécificité.

Figure 1.4 – Exemple de non-spécificité d'une relation spatiale : A « au nord de » B



a b c

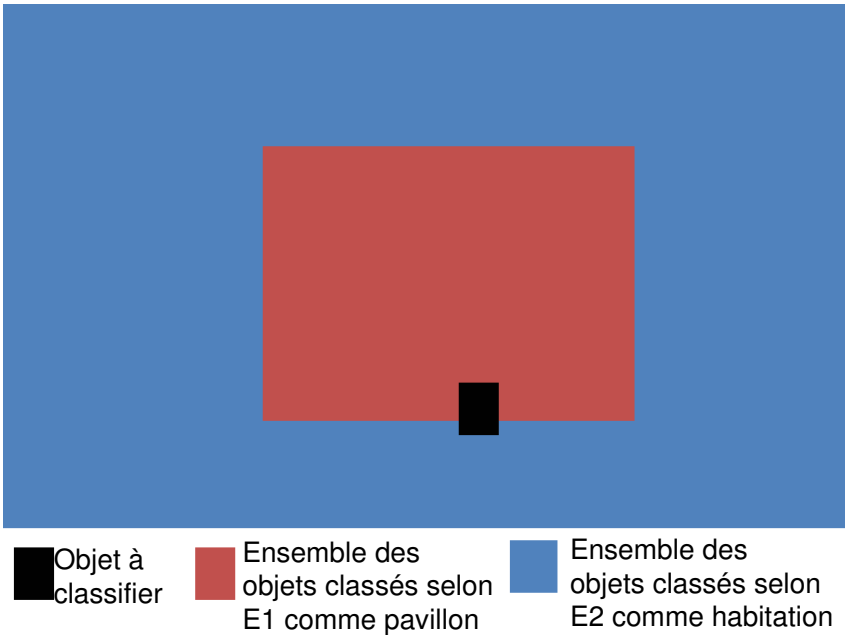
Considérons l'exemple extrait de [Fisher, 2005] illustrant la non-spécificité spatiale par la relation « au nord de ». La relation A est « au nord de » B veut-elle dire :

- que les objets sont sur la même longitude mais que A est plus au nord que B sur cette ligne (voir figure 1.4a) ou
- que A est quelque part au nord du parallèle qui passe par B (voir figure 1.4b) ou
- que A se situe entre le nord-nord-est et le nord-nord-ouest de B (voir figure 1.4c).

Les deux premières définitions sont précises et valides, la troisième est vague.

L'exemple de la figure 1.5 illustre le problème de l'utilisation de plusieurs échelles d'analyse. L'expert $E1$ prend une échelle de description (zone pavillonnaire) plus fine que l'expert $E2$ (zone d'habitation). Bien que les deux classifications n'aient pas des sémantiques contradictoires (l'une peut être incluse dans l'autre), la classification de l'objet en noir est difficile : doit-on considérer l'objet selon l'échelle d'analyse de $E1$ ou selon celle

Figure 1.5 – Exemple de non-spécificité de la description d'objet spatial



de $E2$? De plus les notions de zone pavillonnaire ou de zone d'habitation sont approximatives.

Ainsi, en archéologie, il y a non-spécificité quand on doit choisir la description d'un objet soit selon une analyse à l'échelle du ***fait***, soit selon une analyse à l'échelle de la ***structure***. Il peut encore y avoir non-spécificité quand un expert considère qu'un objet est un mur et qu'un autre dit que l'objet est inclus dans la description spatiale d'une maison.

Les différentes sémantiques associées à la relation (exemple de la figure 1.4) (ou à l'objet, voir figure 1.5) impliquent l'imprécision de la relation (ou de l'objet), et le choix de la définition est incertain. Ainsi la non-spécificité d'une relation ou d'un phénomène correspond à un traitement incertain de données imprécises.

L'ambiguïté est une forme d'imperfection combinant à la fois de l'incertitude et de l'imprécision. En effet, qu'elle soit due à un conflit ou à un problème de non-spécificité, l'imperfection résulte d'une imprécision sur les définitions des classes qui implique une incertitude sur la classification.

L'ambiguïté est un trait majeur de l'archéologie urbaine, marquée par la superposition de fragments et leur réutilisation fréquente.

Cependant, l'incertitude, l'imprécision et l'ambiguïté ne permettent pas de décrire toutes les formes d'imperfections. En archéologie, comme en géographie, et dans toutes les bases de données en général, l'incomplétude de l'information est très répandue.

1.4 Incomplétude

Les incomplétudes sont des absences de connaissances ou des connaissances lacunaires sur des informations du système.

1.4.1 Absence

L'absence est le phénomène qui survient quand dans une base de données, des valeurs manquent à la description de certains objets. On parle d'absence quand, pour un objet particulier, il manque une information nécessaire à sa complète définition. En géographie, par exemple, si l'on souhaite stocker dans un SIG un pays dont on a la représentation spatiale mais pas le nom, la description du pays est incomplète du fait de l'absence de l'information descriptive.

Pour l'information archéologique, toutes les informations ne sont pas forcément disponibles. En effet, si l'on prend par exemple le cas de *BDRues*, les objets inscrits dans la base peuvent être issus de fouilles anciennes pour lesquelles on ne dispose pas de l'orientation ou de la période d'activité des rues qu'ils représentent. L'information archéologique contenue dans *BDRues* est donc sujette à de l'absence.

1.4.2 Lacune

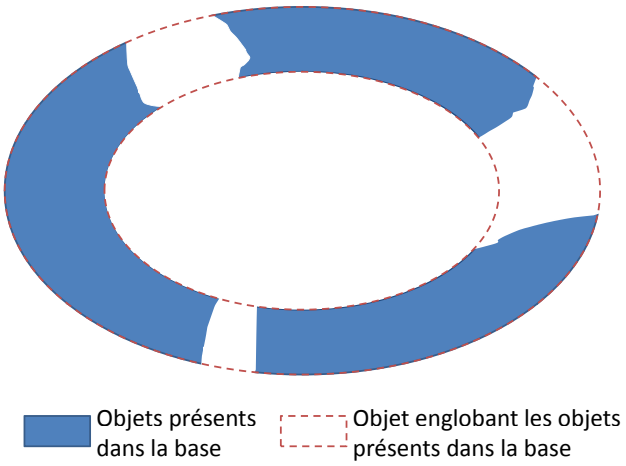
La lacune est le fait qu'un ou plusieurs objets de la base de données ne décrivent que de manière partielle une structure les englobant. Les connaissances expertes sur la structure étudiée permettent de mettre en évidence l'aspect fragmentaire de la description du phénomène obtenue selon les données contenues dans la BDG associée au SIG.

Dans l'exemple de la figure 1.6, les objets de la base de données (en bleu) ne décrivent que partiellement l'objet englobant délimité par des pointillés rouges. L'information spatiale associée aux objets de la base de données est donc lacunaire.

Ceci est généralement le cas en archéologie urbaine, car la structure est partiellement détruite durant le temps séparant son activité et la découverte de fragments de celle-ci. En fait, arriver à reconstituer un objet archéologique par regroupement des vestiges est souvent difficile. Dans le cas des données de *BDRues* et en reprenant la terminologie des échelles de lecture présentée dans l'introduction, les objets archéologiques sont des tronçons de rues, définis à l'échelle du *fait*, ne représentant que très partiellement les rues auxquelles ils appartenaient au regard d'une analyse à l'échelle de la *structure*.

On peut remarquer que bien qu'à une échelle de description archéologique donnée (comme par exemple l'échelle du *fait*), les objets puissent être bien définis (des tronçons de rue peuvent théoriquement être bien définis), par le changement d'échelle (le passage à l'échelle de la *structure*), les objets à la nouvelle échelle (les rues) sont mal-définis car l'information qu'ils comportent apparait alors comme lacunaire (les rue ne sont que partiellement définies).

Figure 1.6 – Exemple d'information spatiale lacunaire



Remarquons que l'absence et la lacune de l'information nécessitent généralement un traitement particulier des données en présence afin de compléter cette information. Ce traitement peut, par exemple, être une analyse spatiotemporelle ou une généralisation.

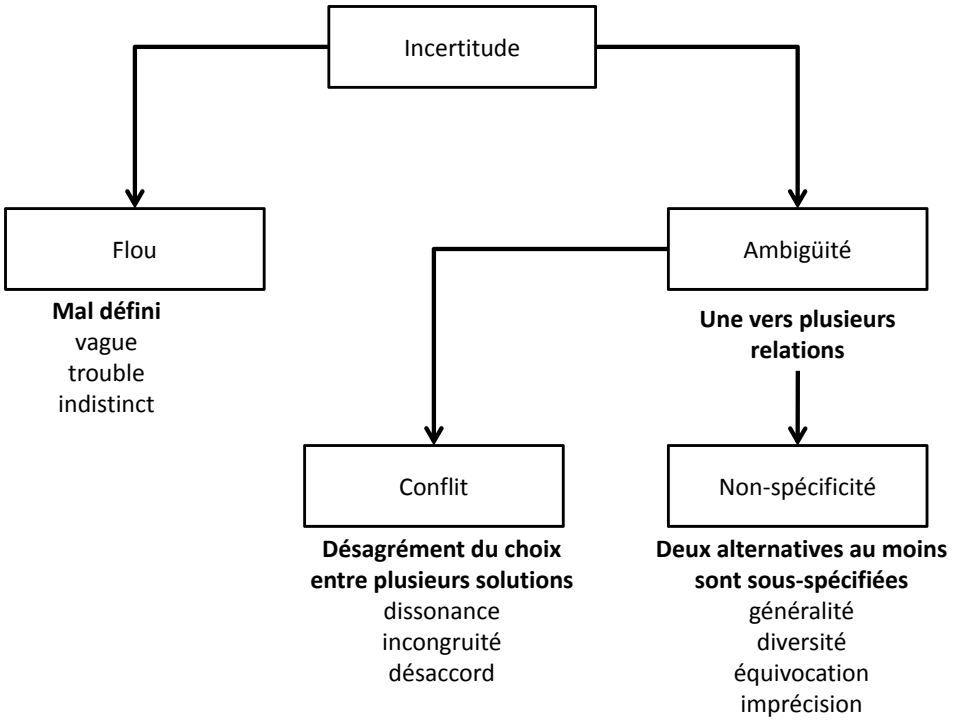
◇◇◇◇◇

[Klir et Yuan, 1995] proposent la classification des incertitudes présentée dans la figure 1.7. À partir de cette structuration, plusieurs approches spécifiques à l'information géographique ont été proposées, comme celle proposée dans [Fisher *et al.*, 2005] et dans [Fisher, 2005].

Dans ce chapitre, la typologie de l'imperfection présentée, qui se base sur celle proposée par Fisher, considère l'imperfection et ses formes. Elle est illustrée dans la figure 1.8. Elle se différencie de l'approche de [Klir et Yuan, 1995] qui observe l'incertitude. Par l'utilisation de cette typologie, l'étude de la nature des imperfections auxquelles sont sujettes les données archéologiques est facilitée. L'information peut alors être analysée en tenant compte de ses imperfections, et ainsi de sa qualité et par la même d'une partie de sa complexité.

En décrivant, selon leur qualité, les données stockées dans la base de données associée à un système d'information à références spatiales dédié à un usage archéologique, je guide le choix de la théorie de représentation pour la modélisation de l'information ; ce choix est nécessaire à la préparation des données en vue de les analyser.

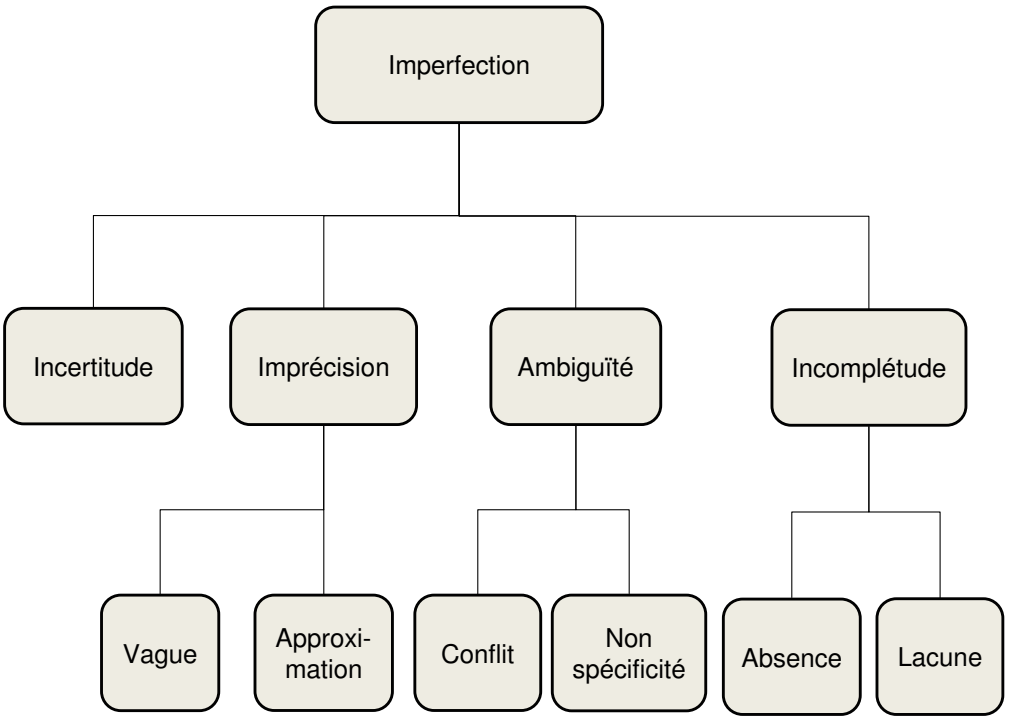
Figure 1.7 – Typologie de l'incertitude selon [Klir et Yuan, 1995]



Le choix de la théorie dépend de la nature de l'imperfection. Pour faire ce choix, on dispose, entre autres, des différentes théories présentées dans le chapitre 2 : la théorie des probabilités, des ensembles flous, des possibilités, des fonctions de croyances et des ensembles approximatifs.

Ces théories ont toutes pour objectif de permettre la modélisation de l'information selon leurs imperfections. Elles sont complémentaires et permettent la représentation des données imparfaites, qu'elles soient de nature incertaine, imprécise ou ambiguë. L'incomplétude doit elle être comblée par des traitements de données.

Figure 1.8 – Typologie de l'imperfection des données archéologiques



Chapitre 2

Représentations formelles de l'information imparfaite

Sommaire

L'information archéologique présente des imperfections dont la nature a été décrite au chapitre 1. Pour modéliser l'information imparfaite, plusieurs approches ont été proposées dans la littérature [Bloch, 1996]. Ces approches proposent des formalisations différentes et complémentaires pour la représentation de l'imparfait [36]. L'objectif de ce chapitre est de présenter les formalismes de l'imparfait les plus classiques.

36. L'imparfait est à l'imperfection ce que l'incertain est à l'incertitude.

Le développement des théories de représentation de l'information imparfaite est lié à l'émergence, depuis le milieu du vingtième siècle, de la théorie de la décision et de l'intelligence artificielle. Ces deux domaines ont abordé le problème de manières différentes : la théorie de la décision fonde ses représentations sur l'observation des choix d'un individu tandis que l'intelligence artificielle opte pour une approche plus introspective et cherche à formaliser des modes de raisonnement. La première s'est fortement basée sur la théorie des probabilités ; la seconde s'est intéressée à des approches plus qualitatives.

Au vu des nombreuses formes d'imperfection de l'information, les mathématiques et les sciences de l'information ont développé de nombreuses théories permettant la représentation de l'information imparfaite. Généralement, la théorie des probabilités est considérée comme une théorie de représentation de l'incertitude, alors que la théorie des ensembles flous et la théorie des ensembles approximatifs s'intéressent à la représentation de l'imprécis. Cette distinction entre les champs d'application (incertitude, imprécision) de ces théories (probabilités et ensembles flous ou approximatifs) est formelle [37]. Les théories des possibilités et de l'évidence peuvent être vues comme des approches mixtes tentant de faire le pont entre imprécision et incertitude. Elles permettent une bonne modélisation pour la résolution de l'ambiguïté.

J'aborderai dans ce chapitre ces différentes théories qui sont toutes mobilisables pour représenter des données imparfaites dans un SIG. Je commencerai par présenter la plus ancienne et la plus étudiée — la théorie des probabilités — par l'intermédiaire de la description du cadre classique, de l'approche fréquentiste et de l'approche subjective. Ensuite, en se basant sur cette dernière, j'introduirai la théorie de l'évidence. Puis, j'exposerai la démarche générale de la théorie des ensembles flous et celle des possibilités. Enfin, j'expliciterai la théorie des ensembles approximatifs.

2.1 Théorie des probabilités

La théorie des probabilités est la plus ancienne et la plus étudiée des théories de l'imparfait. Elle trouve son origine dans l'étude des jeux de hasard [38], le mot *probabilité* étant alors associé à la chance [39].

Depuis, d'autres interprétations furent proposées : une probabilité mesure-t-elle le réel, la tendance physique d'un événement à se réaliser, ou est-ce plutôt une mesure de la croyance de quelqu'un en l'apparition de l'événement ?

Quelle que soit la réponse, elle fournit une interprétation des valeurs de la probabilité. Il existe deux larges catégories d'interprétation des probabilités : les ***probabilités***

37. En effet, dans les usages et pratiques, elle n'est pas évidente.

38. Le véritable début de la théorie des probabilités date de la correspondance entre Pierre de Fermat (1601-1665) et Blaise Pascal (1623-1662) en 1654. Ceux-ci commencent à élaborer les bases du traitement mathématique des probabilités autour de l'étude de jeux de hasard.

39. Dans son ouvrage *Théorie analytique des probabilités*, Pierre-Simon Laplace (1749-1827) parle de *théorie de la chance*.

objectives et les ***probabilités subjectives***.

Les probabilités objectives, ou approche fréquentiste des probabilités, sont associées aux systèmes ayant un fonctionnement aléatoire à l'image du jeu de la roulette, le lancer de dé ou la durée de vie d'un atome radioactif. Dans ces systèmes, la probabilité d'un événement (par exemple « le résultat du lancer d'un dé est 6 ») correspond à la fréquence de réalisation de l'événement lorsque l'on répète un nombre important de fois l'expérience.

Les probabilités subjectives aussi appelées bayésiennes peuvent être affectées à n'importe quel fait (même issu de phénomène non aléatoire). Elles représentent la plausibilité subjective, ou le degré de confiance que l'on donne à la validité du fait. Le plus généralement, on considère que la valeur d'une probabilité subjective est un degré de croyance. Par exemple, à la question « quelle est la probabilité qu'il neige demain ? », un agent n'ayant pas accès aux modélisations météorologiques donnera une réponse subjective.

On peut remarquer que, quelle que soit l'interprétation, la probabilité est une représentation de la validité de la réalisation d'un événement (ou de la véracité d'une hypothèse). Ainsi, elle permet ainsi de modéliser l'incertitude associée à cet événement (cette hypothèse).

La théorie des probabilités a un cadre mathématique et deux principales interprétations. C'est pourquoi, je présenterai d'abord le cadre mathématique de la théorie des probabilités, puis l'approche fréquentiste pour finir par l'approche des subjectivistes.

2.1.1 Cadre classique

Que l'univers [40] Ω soit discret ou continu, la probabilité $P(A)$ d'un événement [41] $A \subseteq \Omega$ prend sa valeur dans $[0;1]$. Ainsi, soit 2^Ω l'ensemble des parties de Ω, la probabilité est donc :

$$\begin{array}{rccc} P: & 2^\Omega & \rightarrow & [0;1] \\ & A & \rightarrow & P(A). \end{array}$$

La probabilité de l'événement Ω est de 1, celle de l'événement vide ($\emptyset$) vaut 0.

Par définition, les probabilités sont additives, *i.e.* la probabilité de deux événements A et B disjoints ($A \cap B = \emptyset$) est égale à la somme de la probabilité de A et de celle de B. Cette propriété se généralise à la réunion d'un ensemble dénombrable d'évènements $\bigcup_{i\in I} A_i$ ($I \subseteq \mathbb{N}$) deux à deux disjoints. Dans ce cas, la probabilité de la réunion est égale à la somme des probabilités des événements :

$$P(\bigcup_{i\in I} A_i) = \sum_{i\in I} P(A_i), \quad I \subseteq \mathbb{N}.$$

Cette contrainte d'additivité pour des événements disjoints est l'une des caractéris-

40. Univers : ensemble des issues ou événements élémentaires possibles à une expérience.

41. Événement : ensemble de valeurs appartenant à l'univers. Soit l'univers Ω et A un événement alors $A \subseteq \Omega$.

tiques des probabilités [42].

Dans le cas discret, et lorsque les issues sont équivalentes, la probabilité est alors considérée comme étant le nombre de issues favorables sur le nombre d'issues possibles. Ainsi, pour revenir à l'exemple du lancer d'un dé, on peut modéliser cette expérience en se donnant un univers $\Omega = \{1; 2; 3; 4; 5; 6\}$ correspondant aux valeurs possibles du dé, et une fonction p qui à chaque $i \in \Omega$ associe $p(i) = \frac{1}{6}$.

Dans le cadre classique, la probabilité d'un événement est donc associée à la valeur résultant d'une expérience aléatoire. La probabilité d'un événement pour une expérience aléatoire résulte de la plausibilité de réalisation de cet événement. Par exemple, pour le lancer d'un dé à six faces numérotées de 1 à 6, le résultat est 6 avec une probabilité de 1/6.

Il existe une autre façon de déterminer la probabilité associée à un événement de cette expérience. Cette autre méthode repose sur la répétition à l'identique de l'expérience et c'est l'objet de l'approche fréquentiste.

2.1.2 Approche fréquentiste

L'approche fréquentiste ou objectiviste des probabilités définit la probabilité d'un événement en terme de fréquence relative. Supposons qu'une expérience puisse être effectuée plusieurs fois dans des conditions identiques. Pour chaque événement A, on définit $N(A)$ comme étant le nombre d'apparitions du résultat A lors des N premières expériences. La probabilité de l'événement A est la limite de la fréquence relative $N(A)/N$ c'est-à-dire :

$$P(A) = \lim_{N \to \infty} N(A)/N \tag{2.1}$$

La probabilité est ainsi relative à une série d'événements semblables. Il s'agit d'une probabilité statistique. Les probabilités objectives permettent notamment la représentation des phénomènes de hasard.

Le lancer multiple d'une pièce pour obtenir *pile ou face* est un exemple de phénomène étudié par le prisme des probabilités fréquentistes. Quand la pièce n'est pas truquée, selon les fréquentistes, la probabilité de tirer *pile* est de 1/2 non pas du fait que cet événement est aussi plausible que de tirer *face* mais du fait que des séries répétées d'un grand nombre d'essais démontre que la fréquence de tirage *pile* converge empiriquement vers la limite de 1/2 quand le nombre d'essais tend vers l'infini.

Toutefois, dans ce cadre, un nombre suffisant d'observations (idéalement infini) du phénomène observé est nécessaire. Ceci interdit d'attribuer des probabilités à des événements dans le cadre d'expériences non renouvelables. De plus, une expérience est-elle réellement

42. Voir pour une présentation plus complète des probabilités l'ouvrage de P. Lazar et D. Schwartz *Eléments De Probabilité Et Statistiques*, Flammarion 1967 Paris.

renouvelable dans des conditions identiques ? Dans le cas de phénomènes non clairement reproductibles à l'identique, une vision subjective des probabilités est souhaitée.

2.1.3 Approche subjective

Dans cette approche, les probabilités subjectives (ou bayésiennes) désignent la classe d'estimations, de jugements et de croyances qui guident les agents parfois indépendamment de toute rigueur ou cohérence logique. Les probabilités sont alors perçues comme des degrés de confiance d'un agent dans une déclaration arbitraire (ou proposition pertinente pour le problème appréhendé).

Par exemple [43], on peut considérer que ce qui joue le rôle des fréquences pour les phénomènes non-reproductibles, ce sont des sommes d'argent misées sur l'occurrence ou la non-occurrence d'événements pertinents pour le phénomène. Le degré de confiance d'un agent en l'événement A est alors défini comme le prix $P(A)$ que cet agent accepterait de payer pour acheter un billet de loterie qui lui fait gagner 1 euro si l'événement A se produit. Plus l'agent croit en l'occurence A, plus il acceptera facilement d'acheter un billet pour un prix proche de 1 euro et donc plus $P(A)$ sera élevé. Moins il croit en l'occurence A, plus le montant qu'il acceptera de payer sera faible et plus $P(A)$ sera faible.

Ainsi, dans le cadre classique la probabilité d'un événement correspond à l'aléa de cet événement à apparaitre après une expérience aléatoire ; dans l'approche fréquentiste, elle est égale à la fréquence d'apparition d'un événement au cours d'un grand nombre d'expériences similaires alors que dans l'approche des subjectivisites, elle équivaut au degré de croyance d'un agent dans l'apparition de l'événement pour toute expérience. La théorie des probabilités permet de représenter l'incertitude quelle que soit l'approche choisie.

Cependant, à l'aide du modèle probabiliste, il est difficile de modéliser l'absence de connaissances, des connaissances imprécises (au contraire des connaissances incertaines qui sont naturellement représentées par des probabilités), ou encore l'ignorance que l'on peut avoir sur un phénomène.

En général, les probabilités posent deux problèmes. En effet, l'interprétation et l'additivité des probabilités ne permettent pas de modéliser certains phénomènes comme l'imprécision, l'ignorance ou l'ambiguïté par exemple.

Pour ce faire, Arthur Dempster a proposé en 1968 un nouveau cadre mathématique [44] qui est à la base de la théorie des fonctions de croyance. Cette théorie, aussi appelée théorie de l'évidence, permettant de considérer à la fois de l'imprécision et de l'incertitude est adéquate pour résoudre les problèmes où il y a conflit entre différentes sources.

43. Exemple extrait d'un complément de cours de DEA de l'Université Toulouse 3 datant de 2002 proposé par Didier Dubois et Henri Prade

44. Ce cadre a été proposé dans [Dempster, 1968] et fut la base d'un séminaire suivi par Glenn Shafer. Ce dernier a développé, en 1976, ce cadre mathématique et formalisé les concepts de la théorie de l'évidence [Shafer, 1976].

2.2 Théorie de l'évidence

La théorie de l'évidence, aussi appelée théorie des fonctions de croyance ou encore théorie de Dempster-Shafer, fut introduite par Arthur Dempster dans [Dempster, 1968] puis formalisée par Glenn Shafer dans [Shafer, 1976].

Dans la théorie des croyances de Dempster-Shafer [27], les raisonnements sont effectués à partir de deux mesures et non une seule (la crédibilité et la plausibilité), qui permettent de représenter à la fois l'incertitude et l'imprécision, ce qui constitue donc une généralisation de l'approche bayésienne des probabilités.

La théorie de l'évidence provient des tentatives d'adaptation, par les chercheurs en intelligence artificielle, de la théorie des probabilités aux systèmes experts. Dans le modèle des fonctions de croyance, on probabilise l'approche ensembliste de l'imprécis. On passe d'une représentation de la forme $x \in A$ où A est un ensemble d'événements possibles, à une distribution de probabilité discrète sur les divers énoncés possibles de la forme $x \in A$.

2.2.1 Cadre de discernement

Soit Ω l'univers, c'est à dire l'ensemble fini contenant tous les éléments, ou encore l'ensemble fini des issues ou événements élémentaires possibles.

On note m une distribution de probabilité sur l'ensemble 2^Ω des parties de Ω. m est alors une *fonction de masse* et $m(A)$ est la masse de croyance affectée à A. A est appelé ensemble focal si $m(A) > 0$.

La valeur de $m(A)$ concerne donc seulement l'événement A et n'apporte aucun crédit aux sous-ensembles de A chacun ayant, par définition, sa propre masse.

En général, l'ensemble vide n'est pas un ensemble focal, on suppose donc que $m(\emptyset) = 0$. $m(\emptyset) = 0$ peut être vu comme une forme de normalisation, Ω est alors dit exhaustif. La fonction de masse est une distribution de probabilité sur 2^Ω, ainsi on a la condition :

$$\sum_{A \subseteq \Omega} m(A) = 1.$$

$m(\Omega)$ représente l'ignorance, ainsi lorsque $m(\Omega) = 1$ alors on se situe dans l'ignorance totale. Si on a $m(A) = 1$, avec $A \subset \Omega$ et $A \neq \emptyset$, alors la connaissance est parfaite.

La croyance $\text{bel}(A)$ d'un événement A est définie comme la somme des masses de tous ses sous-ensembles :

$$\text{bel}(A) = \sum_{B | B \subseteq A} m(B).$$

La plausibilité $\text{pl}(A)$ est définie comme la somme des masses de tous les ensembles B qui intersectent A :

$$\text{pl}(A) = \sum_{B | B \cap A \neq \varnothing} m(B)$$

Ces deux mesures sont liées : $\mathrm{pl}(A) = 1 - \mathrm{bel}(\overline{A})$. De ce fait, la connaissance d'une seule de ces distributions (masse, croyance ou plausibilité) suffit à déduire les deux autres.

De plus, à partir de la valeur de la masse d'un état, on peut définir un intervalle de probabilité. Cet intervalle contient la valeur précise de la probabilité de l'état, et est borné par la croyance et la plausibilité de l'état :

$$\mathrm{bel}(A) \leq P(A) \leq \mathrm{pl}(A).$$

Cet intervalle permet la gestion à la fois des imprécisions et de l'incertitude.

2.2.2 Règle de combinaison de Dempster

Dans le cadre de fonctions de masse multiples, cas des conflits entre les données, on peut avoir à combiner les masses pour prendre une décision. Dans ce cadre, Dempster a proposé la règle de combinaison $\otimes$ suivante : soit m_1 et m_2 deux fonctions de masses issues de sources d'information distinctes, alors

$$(m1 \otimes m_2)(A) = \frac{\sum_{B \cap C = A} m1(B) \times m_2(C)}{1 - \sum_{B \cap C = \emptyset} m1(B) \times m_2(C)}, \forall A \subseteq \Omega$$

Cette règle se veut une généralisation du théorème de Bayes aux fonctions de croyances. D'autres opérateurs de combinaison ont été définis dans la littérature ; ils ne seront pas présentés dans ce mémoire.

Ainsi, avec les fonctions de croyance et l'opérateur de combinaison, la théorie de l'évidence donne un cadre permettant l'obtention des plausibilités et croyances des événements étudiés dans le cadre de sources multiples. Il reste ensuite, dans ce cadre, à choisir la meilleure masse. Pour cela, Philippe Smets dans [Smets, 1990] a proposé le modèle des croyances transférables (MCT). Ce modèle a pour but de permettre la prise de décision.

2.2.3 Modèle des croyances transférables

Le modèle de croyances transférables, introduit dans [Smets, 1990], est un modèle de raisonnement incertain sur de l'imprécis reposant sur la théorie des fonctions de croyance. Il a pour objet de donner un environnement permettant la prise de décision [Smets, 1994].

Le MCT comporte deux niveaux : le niveau crédal pour représenter et combiner les informations, et le niveau pignistique pour prendre une décision. Généralement, au niveau crédal, on se place dans le cadre de la théorie de l'évidence.

De nombreux outils ont initialement proposés dans la théorie des probabilités ont été adapté au cadre du MCT comme les réseaux de croyances et les arbres de décisions.

Pour cela, si l'on considère un problème dont la modélisation au niveau crédal s'est faite à l'aide de la théorie de l'évidence et de l'opérateur de combinaison de Dempster,

alors la probabilité pignistique betP(X) de $X \subseteq \Omega$ est définie ainsi :

$$\text{betP}(X) = \sum_{Y \subseteq \Omega, Y \neq \emptyset} \frac{card(X \cap Y) \times m(Y)}{card(Y) * (1 - m(\emptyset))}$$

Afin de prendre une décision, une solution habituelle consiste à prendre le maximum des probabilités pignistiques c'est à dire :

$$\max_{X \in \Omega}(\text{betP(X)}).$$

Il existe d'autres opérateurs, mais celui-ci est un compromis entre une vision optimiste (maximum de plausibilité) et pessimiste (maximum de crédibilité).

Ainsi, le MCT donne un cadre formel à la décision dans le contexte de problème de fusion de sources, c'est-à-dire de résolution d'ambiguïtés et plus particulièrement de conflits.

À l'aide de la théorie des fonctions de croyance on peut modéliser des problèmes de choix incertain sur de l'imprécis. Elle peut être considérée comme une théorie dérivée ou comme une généralisation de la théorie des probabilités bayésiennes. En utilisant en plus le modèle de croyances transférables, on obtient un cadre complet pour la représentation et la décision.

Quand l'approche bayésienne fait appel à des probabilités pour chaque question, les fonctions de croyance permettent de baser l'initialisation des degrés de confiance pour une question, sur les probabilités d'une question afférante. Les degrés de confiance associés aux fonctions peuvent avoir ou non les propriétés mathématiques des probabilités. La divergence de ces degrés avec les probabilités dépend de la proximité entre les deux questions.

Cependant, on ne décrit pas réellement les ensembles avec leurs imperfections. On ne peut pas évaluer l'appartenance d'un individu à une classe. Pour cela, Lotfi Zadeh introduisit dans [Zadeh, 1965], les ensembles flous et la théorie associée.

2.3 Théorie des ensembles flous

Afin de pouvoir représenter une notion imprécise, le postulat de la théorie des ensembles flous est de considérer une gradation, définie sur $[0, 1]$, sur l'adéquation à la notion. Ainsi, on peut *être vieux* avec une graduation de, par exemple, 1, 0.6, 0.4 ou encore 0. Par cette graduation, le souhait est de relâcher les contraintes classiques de l'appartenance à un ensemble.

Ainsi, la théorie des ensembles flous [45] [Zadeh, 1965] peut être considérée comme une adaptation de la théorie des ensembles où, d'après [Bouchon-Meunier, 1993], « la notion

45. Les ensembles flous sont aussi nommés sous-ensembles flous [Bouchon-Meunier, 1995]

de sous-ensemble flou a pour but de permettre des graduations dans l'appartenance d'un élément à une classe ».

2.3.1 Des ensembles classiques aux ensembles flous

On peut classiquement définir un ensemble classique dans E par les trois méthodes suivantes :

- Un ensemble est défini, par extension, par l'énumération de tous ses membres. Ce listage ne peut être utilisé que pour les ensembles finis. Ainsi l'ensemble A, dont les membres sont $a_1, a_2, ..., a_n$ est noté $A = \{a_1, a_2, ..., a_n\}$.
- Un ensemble est défini par une propriété satisfaite par ses membres. Une notation classique pour ces ensembles est : $A = \{x|P(x)\}$, où le symbole | signifie « tel que » et $P(x)$ désigne l'hypothèse « x vérifie P ». Il est nécessaire que P soit une propriété telle que, pour tout x, $P(x)$ est soit vrai soit faux.
- Un ensemble peut être défini par une fonction caractéristique indiquant les éléments qui sont membres de l'ensemble et ceux qui ne le sont pas. Un ensemble A est défini par sa fonction caractéristique a comme suit :

$$\begin{cases} a(x) = 1 & \text{si } x \in A, \\ a(x) = 0 & \text{si } x \notin A. \end{cases}$$

Ainsi, l'ensemble défini par l'intervalle $[2, 8]$ sur l'axe des réels peut être représenté par la Figure 2.1.

Les ***ensembles flous*** sont définis comme des ensembles pouvant contenir des éléments de façon partielle, c'est-à-dire que la fonction caractéristique (sa ***fonction d'appartenance*** dans la théorie des ensembles flous) a d'un ensemble flou A peut prendre des valeurs entre 0 et 1 :

$$\begin{cases} a(x) = 1 & \text{si } x \text{ appartient } A, \\ a(x) = 0 & \text{si } x \text{ n'appartient pas } A, \\ a(x) \in]0, 1[& \text{quand } x \text{ appartient partiellement à } A \end{cases}$$

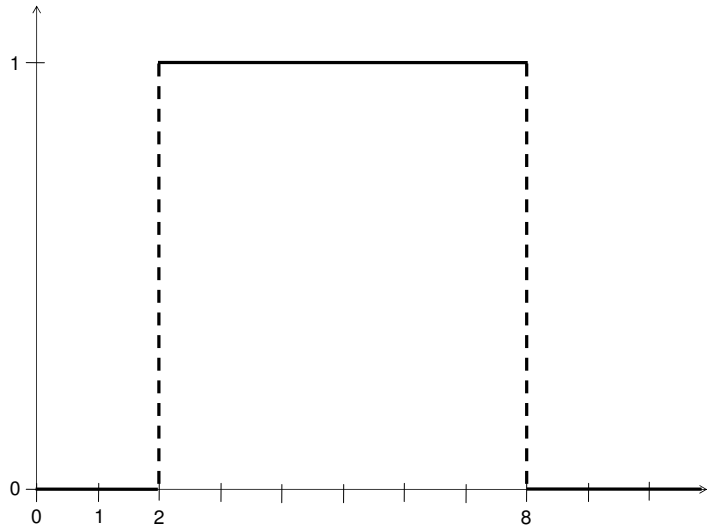
Les ensembles flous permettent donc de graduer la notion d'appartenance d'un élément à un ensemble. Cette graduation permet de modéliser plus facilement le vague.

2.3.2 Définitions

Des définitions supplémentaires sont nécessaires pour la compréhension de ce mémoire de thèse. Soit un ensemble flou F défini sur E et de fonction d'appartenance f.

Le ***support*** de F ($Support(F)$) est l'ensemble des éléments x tels que $f(x)$ est non

Figure 2.1 – Représentation de la fonction caractéristique de l'ensemble $[2, 8]$ défini sur l'axe des réels



nulle :

$$Support(F) = \{x \in E | f(x) > 0\}$$

La ***hauteur*** de F est la valeur maximale de la fonction d'appartenance f :

$$Hauteur(F) = \sup_{x \in E}(f(x)).$$

Le ***cœur*** de F est l'ensemble des x pour lesquels la valeur de la fonction d'appartenance $f(x)$ est égale à la hauteur de F :

$$Coeur(F) = \{x \in E | f(x) = Hauteur(F)\}.$$

Le ***noyau*** de F est l'ensemble des x tels que $f(x)$ est égale à 1 :

$$Ker(F) = \{x \in E | f(x) = 1\}.$$

Lorsque la hauteur de F est égale à 1, le cœur de F est son noyau.

Un ensemble flou est dit normalisé si et seulement si sa hauteur est égale à 1.

Bien qu'il existe d'autres définitions d'une quantité floue, celle choisie dans ce mémoire est celle de [VanLeekwijck et Kerre, 1999] : une quantité floue est un ensemble flou défini sur $\mathbb{R}$.

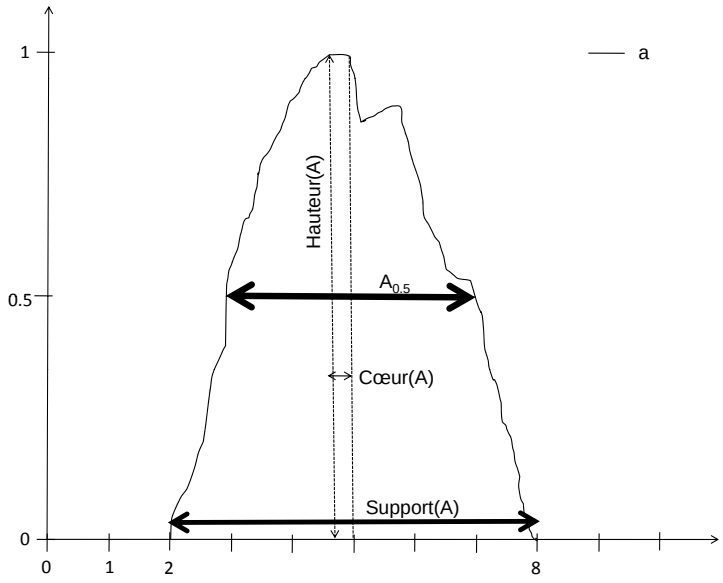
Pour tout E, une α***-coupe*** F_α (α positif) d'un ensemble flou F défini sur E est l'en-

semble des éléments pour lesquels la valeur de la fonction d'appartenance f de F est supérieure à α c'est-à-dire

$$F_\alpha = \{x \in E, f(x) \geq \alpha\} \ \ avec\ \alpha > 0.$$

Par construction, quand $\alpha = 0$, F_α est le support de l'ensemble flou.

Figure 2.2 – Illustration des définitions sur l'ensemble flou A de fonction d'appartenance a défini sur $\mathbb{R}$



Par exemple, un ensemble flou dont le support est l'intervalle $[2, 8]$ défini sur l'axe des réels peut avoir une fonction d'appartenance a telle que présentée dans la Figure 2.2. De plus, cette figure illustre les notions de support, de cœur, de hauteur et d'α-coupe (avec α=0.5) sur un ensemble flou A défini sur $\mathbb{R}$ et de fonction d'appartenance a. Dans cette figure, puisque la hauteur de A vaut 1, on a $Ker(A) = Coeur(A)$.

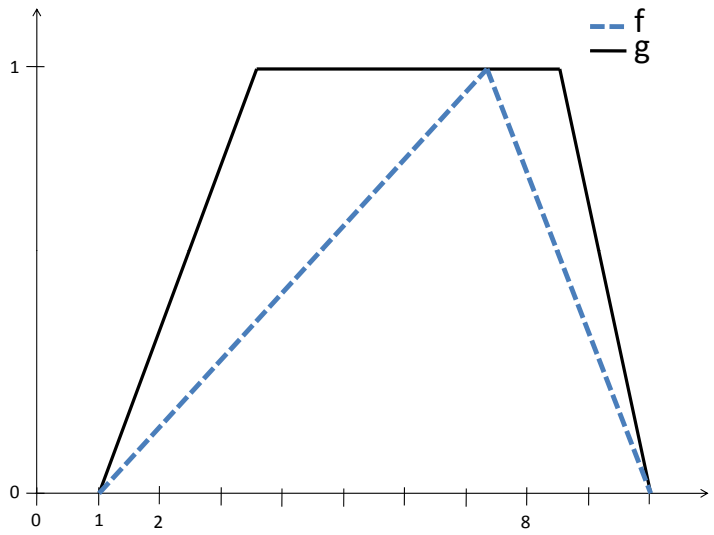
Un ensemble flou F défini sur l'ensemble des réels $\mathbb{R}$ est convexe si et seulement si pour tout α $(0 < \alpha < 1)$ F_α est convexe, c'est-à-dire un intervalle fermé. Ainsi, un ensemble flou F défini sur l'ensemble des réels $\mathbb{R}$ de fonction d'appartenance f est convexe si :

$$\forall (x, y) \in \mathbb{R} \times \mathbb{R},\ \forall z \in [x, y],\ f(z) \geq \min(f(x), f(y)).$$

Dans le cas général, pour tout E, un ensemble flou F défini sur E est dit ***convexe*** si et seulement si toutes ses α-coupes sont des sous-ensembles convexes de E.

Traditionnellement, un ***nombre flou*** est un ensemble flou convexe et normalisé dont la fonction d'appartenance est définie sur $\mathbb{R}$ et semi-continue supérieurement et de support compact (par exemple une forme triangulaire). Dans la figure 2.3, F ayant pour fonction

Figure 2.3 – Ensembles flous F et G de fonction d'appartenance f et g



d'appartenance f est un nombre flou.

Un ***intervalle flou*** est un ensemble flou convexe et normalisé dont la fonction d'appartenance est définie sur $\mathbb{R}$. Dans la figure 2.3, G ayant pour fonction d'appartenance g est un intervalle flou.

Dans ce mémoire, à l'image de [Klir et Yuan, 1995], je ne ferai pas de distinction entre un nombre flou et un intervalle flou. Je les nommerai, tous les deux, nombres flous. Alors, un nombre flou est un ensemble flou convexe et normalisé défini sur $\mathbb{R}$. Ainsi, dans la figure 2.3, F et G seront considérés comme des nombres flous.

De plus, je considère que deux ensembles flous F et G définis sur E sont égaux s'ils ont des fonctions f et g d'appartenance identiques, c'est à dire :

$$\forall x \in Support(F) \cup Support(G), f(x) - g(x) = 0.$$

Plus simplement, soient F et G deux ensembles flous de E,

$$F = G \Leftrightarrow \forall x \in E, f(x) = g(x).$$

Un nombre flou est donc une quantité floue convexe et normalisée.

2.3.3 Logique floue

Comme la théorie des ensembles flous est une théorie ensembliste qui diffère de la théorie des ensembles, une logique particulière a été développée afin de pouvoir raisonner

sur ces ensembles. Cette logique s'appelle la logique floue. Dans celle-ci plusieurs règles d'inférence ont été proposées et le modus ponens fut généralisé.

Des *t-normes* et *t-conormes* ont été proposées pour définir le *ET* et le *OU* logique. Je retiens celles proposées par Zadeh en 1965. Pour Zadeh, si on considère deux ensembles flous F et G définis sur E, de fonctions d'appartenance f et g, alors $F\ ET\ G$ est un ensemble flou $F \wedge G$ dont la fonction d'appartenance $f \cap g$ est définie comme étant le minimum des deux fonctions d'appartenance, soit :

$$\forall x \in E, (f \cap g)(x) = \min(f(x), g(x)).$$

De manière analogue, $F\ OU\ G$, noté $F \vee G$, est défini par le maximum des fonctions d'appartenance :

$$\forall x \in E, (f \cup g)(x) = \max(f(x), g(x)).$$

De nombreux opérateurs flous ont aussi été définis. Dans les chapitres 4 et 7 ainsi que dans l'annexe A, je présente certains opérateurs pour la comparaison d'ensembles flous.

Le concept d'ensemble flou est une adaptation de celui d'ensemble classique. Dans cette approche, à chaque notion est affecté un ensemble flou qui value chaque élément du domaine sur lequel est définie la notion à représenter. Ces valuations prennent valeurs dans $[0, 1]$. Les ensembles flous permettent donc par une certaine granularité de représenter des notions imprécises.

Cette représentation permet de représenter une réalité imprécise, voire vague, dans un formalisme mathématique puissant permettant des traitements par la logique floue. Cependant, les valuations n'ont que peu de potentiel d'interprétation, et ne permettent pas de gérer l'incertitude.

Pour le permettre, la théorie des possibilités introduite dans [Zadeh, 1978] propose d'utiliser deux mesures afin de permettre cette gestion conjointe de l'incertitude et de l'imprécis. Seulement, contrairement à l'approche de Dempster-Shafer, la théorie des possibilités se base sur des représentations ensemblistes floues.

2.4 Théorie des possibilités

Lotfi Zadeh a introduit la théorie des possibilités dans [Zadeh, 1978]. Elle propose un cadre dans lequel connaissances imprécises et incertaines peuvent coexister et être traitées conjointement. Elle permet donc de manipuler des informations imprécises et dont l'incertitude associée est de nature non probabiliste. C'est donc une théorie pratique pour la manipulation de l'information ambiguë. L es ouvrages [Dubois et Prade, 1985] et [Dubois et Prade, 1988] présentent plus longuement la théorie des possibilités.

La théorie des possibilités fournit une méthode pour formaliser des incertitudes subjectives sur des événements. Elle examine dans quelle mesure un événement est possible

et dans quelle mesure on est certain de la réalisation d'un événement. La théorie des possibilités fait intervenir une mesure de ***possibilité*** Π et une mesure de ***nécessité*** N afin de formaliser ces deux évaluations subjectives.

2.4.1 Mesure de possibilités

Soit un ensemble de référence fini Ω, on attribue à chaque sous-ensemble de Ω un coefficient entre 0 et 1 évaluant à quel point cet événement est possible. Une mesure de possibilité Π est une fonction définie sur l'ensemble 2^Ω des parties de Ω, prenant ses valeurs dans $[0, 1]$ telle que :

- $\Pi(\emptyset) = 0$
- $\Pi(\Omega) = 1$
- $\Pi(\bigcup_{i=1,2...,n} A_i) = \sup_{i=1,2...,n}(\Pi(A_i))$, $\forall A_1, A_2, \ldots, A_n \in \Omega$

2.4.2 Mesure de nécessités

Soit un ensemble de référence fini Ω, on attribue à chaque sous-ensemble de Ω un coefficient entre 0 et 1 évaluant à quel point cet événement est nécessaire. Une mesure de nécessité Ω est une fonction définie sur l'ensemble 2^Ω des parties de Ω, prenant ses valeurs dans $[0, 1]$ telle que :

- $N(\emptyset) = 0$
- $N(\Omega) = 1$
- $N(\bigcap_{i=1,2...,n} A_i) = \min_{i=1,2...,n}(N(A_i))$, $\forall A_1, A_2, \ldots, n \in \Omega$

De plus, on a :

$$\forall A \subseteq \Omega, N(A) = 1 - \Pi(\overline{A})$$

et :

$$N(A) \leq \Pi(A).$$

Pour ce qui est de l'interprétation des possibilités :

- si $N(A) = 1$ alors A est certainement vrai (nécessaire) et $\Pi(A) = 1$,
- si $\Pi(A) = 0$ alors A est certainement faux (impossible) et $N(A) = 0$,
- si $\Pi(A) = 1$ alors il ne serait pas surprenant que A arrive ; $N(A)$ est indéterminé, A est donc possible mais pas forcément nécessaire,
- si $N(A) = 0$, il ne serait pas surprenant que A n'arrive pas, $\Pi(A)$ est indéterminé, A n'est pas nécessaire mais n'est pas pour autant impossible.

Les propriétés dans la théorie des possibilités sont soumises à moins de contraintes que leurs homologues dans la théorie des probabilités ou que dans les fonctions de croyances. La théorie des possibilités est très pratique pour la gestion d'information imprécise dans un contexte incertain, et donc particulièrement intéressante pour représenter des données dans le cas de problème de non-spécificité.

Cependant, les mesures de possibilité, les mesures de nécessité, les ensembles flous impliquent l'utilisation de graduations de l'information. Toutefois dans le cas d'information approximative, il peut être difficile de déterminer ces graduations. Dans ce contexte, la théorie des ensembles approximatifs a été proposée par Zdzislaw Pawlak dans [Pawlak, 1982] et présente une vision ensembliste de l'imprécis.

2.5 Théorie des ensembles approximatifs

La théorie des ensembles approximatifs a été introduite par Zdzislaw Pawlak dans [Pawlak, 1982]. Le même auteur proposa en 1991 un ouvrage, [Pawlak, 1991], présentant de manière plus approfondie cette théorie. « Cette approche semble avoir une importance fondamentale pour l'intelligence artificielle et les sciences cognitives » [Pawlak *et al.*, 1995].

Un ensemble approximatif est une approximation formelle d'un ensemble classique par un ensemble appelé approximation haute et un autre ensemble appelé approximation basse. Les ensembles approximatifs se structurent autour de la possibilité et la nécessité d'appartenance d'un élément à un ensemble.

2.5.1 Approximation haute et basse

Soit $I = (\Omega, \mathbb{A})$ un système d'information dans lequel Ω est un ensemble fini non vide d'objets (*i.e.* l'univers), $\mathbb{A}$ l'ensemble des attributs des objets.

Soit $X \subseteq \Omega$ un ensemble cible que l'on souhaite représenter à l'aide d'un sous-ensemble d'attributs P de $\mathbb{A}$. P induit une relation d'équivalence sur les objets de Ω. Soit x un objet de Ω, $[x]_P$ représente la classe d'équivalence contenant x obtenue selon la relation d'équivalence induite par P regroupant les objets ayant les mêmes valeurs pour les attributs de P :

$$[x]_p = \{y \in \Omega | \forall p \in P, p(x) = p(y)\}$$

X peut être approximé en utilisant uniquement l'information contenue dans P par la construction des approximations *P-bas* ($\underline{P}$) et *P-haut* ($\overline{P}$) de X :

$$\underline{P}X = \{x \in \Omega \mid [x]_P \subseteq X\}$$

$$\overline{P}X = \{x \in \Omega \mid [x]_P \cap X \neq \emptyset\}$$

P-bas présente l'ensemble des $x \in \Omega$ tels que la classe d'appartenance de x induite par P est un sous-ensemble de X, c'est-à-dire que l'ensemble des valeurs de x pour les attributs P fait que x appartient à X.

P-haut présente l'ensemble des $x \in \Omega$ tels que l'intersection entre la classe d'appartenance de x induite par P et X est non nulle, c'est-à-dire que l'ensemble des valeurs de x pour les attributs P fait que x est peut-être membre de X.

La région frontière, donnée par l'ensemble issu de la différence $\overline{P}X - \underline{P}X$, est constituée des objets ne pouvant être considérés comme membres ou non-membres de X pour le sous-ensemble P de $\mathbb{A}$.

2.5.2 Interprétation

L'approximation basse d'un ensemble X est une approximation prudente ne contenant que les objets pouvant être considérés comme membres de X pour P. Elle contient donc les objets dont on est certain qu'il seront membres de X ; ils sont donc nécessairement membres de X.

L'approximation haute de X est une approximation généreuse contenant tous les objets susceptibles d'appartenir à X pour P. Elle contient donc les objets dont on peut penser qu'il seront membres de X ; ils sont donc possiblement membres de X.

La frontière contient les objets possiblement membres de X mais non nécessairement membres de X au vu des attributs de P.

Cette approche permet de représenter l'appartenance des éléments à des catégories aux limites mal définies. Ainsi la théorie des ensembles approximatifs permet de représenter des données imprécises et plus précisément des données approximatives.

Dans cette approche, selon un ensemble d'attributs, on définit la possibilité et la nécessité d'appartenance de chaque élément de l'univers à l'ensemble à définir.

Un avantage majeur de cette approche est qu'elle ne nécessite pas de connaissances préliminaires ou additionnelles sur les données, comme pourraient l'exiger par exemple les distributions de probabilités en statistiques, voire comme les graduations des fonctions d'appartenance en théorie des ensembles flous.

◇◇◇◇◇

Dans ce chapitre, j'ai exposé les principaux formalismes de représentation de l'imparfait. Ainsi, j'ai, tout d'abord, présenté la théorie des probabilités selon que l'on se situe dans une approche fréquentiste ou subjective. Puis, j'ai évoqué la théorie de l'évidence. Ensuite j'ai exposé la théorie des ensembles flous très utilisée dans les SIG pour la formalisation de l'imprécision. J'ai, de même introduit la théorie des possibilités et enfin celle des ensembles approximatifs.

Ces théories sont fortement liées et complémentaires. Bien qu'il existe des passerelles entre elles, chacune des théories apporte une possibilité de modélisation différente qui donne un éclairage particulier. En effet, le changement de théorie modifie notamment la qualité de l'interprétation de l'information étudiée. Par exemple, l'interprétabilité des

probabilités étant plus forte que celle des possibilités, le passage des probabilités vers les possibilités réduit l'impact sémantique, tandis que celui des possibilités vers les probabilités l'augmente artificiellement.

Le choix de la théorie de formalisation de l'information imparfaite, capital à la gestion de l'information, dépend du contexte de l'utilisation. Il doit se baser sur la nature de l'imperfection des informations. À partir des différentes approches de représentation de l'imparfait et de la description des imperfections, on peut déduire des contextes d'utilisation.

Dans le chapitre 3, après avoir décrypté la nature des données dans le chapitre 1 et étudié, dans ce chapitre, les théories de formalisation de l'imparfait, je proposerai une taxonomie associant aux types d'imperfections les théories de l'imparfait les plus adéquates dans le contexte de l'information archéologique manipulée dans un SIG. Ensuite j'exposerai les différentes modélisations de l'information spatiotemporelle associées aux théories de représentation de l'information imparfaite. Enfin, à partir de la taxonomie et des possibles modélisations, j'expliciterai les choix de représentation faits pour la représentation des tronçons de rues romaines dans le projet *SIGRem*.

Chapitre 3

Taxonomie et modélisation de l'imperfection dans un SIG archéologique

Sommaire

Dans les chapitres précédents, la nature de l'imperfection de l'information archéologique et les théories de représentation de l'imparfait ont été étudiées. L'objectif de ce chapitre est de lier le choix de la théorie à la nature de l'information et de modéliser l'information en fonction de ce choix.

Peter Fisher propose de relier la nature de l'imperfection au choix d'une théorie de représentation dans le cas de l'information géographique [Fisher, 2005]. Pour cela, il propose l'utilisation d'une taxonomie de l'imperfection, basée sur la typologie proposée dans [Klir et Yuan, 1995] (voir fin du chapitre 1), dans laquelle il associe à chaque taxon [46] n'ayant pas de descendant le choix d'au moins une théorie de représentation. Je propose

46. Taxon : entité conceptuelle de la taxonomie qui est censée regrouper tous les éléments possédant en commun une certaine sémantique.

une adaptation de la taxonomie de [Fisher *et al.*, 2005] en vue de considérer les spécificités de l'information archéologique.

L'utilisation de modèles pour la représentation des informations en tenant compte de la nature de l'imperfection de celles-ci est une pratique classique en SIG. En effet que ce soit pour la représentation spatiale (voir [Shi, 2007]) ou pour la représentation temporelle (*c.f.* [Dragicevic et Marceau, 2000b]), l'utilisation de modèles permet une uniformisation du traitement des données.

Malheureusement, si les composantes spatiales ont un comportement similaire en géographie et en archéologie, ce n'est pas le cas des composantes temporelles. Ainsi, en géographie, les modélisations temporelles sont explorées de manières multiples et ont généralement pour but la simulation des dynamiques géographiques entre les objets identifiées dans des séries temporelles d'observation du sol. Ces modélisations ne sont pas les plus pertinentes pour l'information archéologique qui évolue plus sur le temps long [Rodier et Saligny, 2007] où les objets sont définis sur des périodes [Accary *et al.*, 2003]. Mais bien que les objectifs de la modélisation temporelle dans les SIG diffèrent entre utilisation géographique et archéologique, les modèles géographiques de l'information temporelle imparfaite peuvent être adaptés au contexte archéologique.

Le travail proposé dans ce mémoire a été effectué à partir d'une base de données archéologiques relatives aux tronçons de rues datant de l'époque romaine trouvés à Reims (*BDRues*). Je propose dans ce chapitre de représenter et modéliser l'information contenue dans cette base en tenant compte de son imperfection. Les choix de représentation et de modélisation sont guidés par la taxonomie adaptée au contexte archéologique exposée dans ce chapitre, et ont porté sur des représentations des composantes par des ensembles flous suivant une modélisation floue propre à chaque composante. Ces choix m'ont amené à créer une nouvelle base *BDFRues* contenant les objets ainsi modélisés.

Après avoir étudié certaines taxonomies de l'imperfection existantes en géomatique, je proposerai une adaptation de celles-ci pour le cas des données archéologiques. J'exposerai de même les différents modèles spatiaux et temporels associés aux différentes théories de représentation de l'imperfection. Enfin, j'utiliserai la taxonomie proposée et certains modèles présentés afin de modéliser les données de *BDRues*.

3.1 Taxonomie de l'imperfection dans les SIG

Depuis la typologie de [Klir et Yuan, 1995] (voir figure 1.7 dans le chapitre 1, de nombreuses approches visant à structurer l'incertitude ou l'imperfection de l'information dans les SIG ont été proposées. La plupart d'entre elles proposent des typologies et ont donc un but ontologique : donner une sémantique à chaque cas d'imperfection. Elles n'ont pas toutes la même finesse de description et ne s'adaptent pas forcément à l'ensemble de l'information géographique. Elles dépendent de l'objectif visé. Par exemple, [Bejaoui *et al.*, 2007] explicitent la modélisation du vague en terme spatial et temporel

par une structuration du fonctionnement de la modélisation dans les SIG.

Mais ces catégorisations ne considèrent pas les mêmes descriptions de la nature de l'imperfection que celles proposées dans le chapitre 1. Elles ne sont pas forcément les plus adaptées à la gestion de l'information spatiotemporelle imparfaite voire archéologique. Cependant, la typologie de l'imperfection proposée dans la figure 1.8 est proche de la taxonomie proposée dans [Fisher *et al.*, 2005], seule la structuration pour l'imprécision et l'incomplétude diffèrent. La taxonomie de Fisher présente l'avantage d'associer à chaque taxon de plus bas niveau une théorie de représentation. C'est le choix fait pour la taxonomie proposée dans ce chapitre.

Ainsi après avoir présenté l'approche de [Fisher, 2005] pour les données géographiques, je présenterai et expliciterai les choix faits pour la taxonomie proposée.

3.1.1 Cas de l'information géographique

Dans [Fisher, 2005] et[Fisher *et al.*, 2005], Peter Fisher propose d'associer à la nature de l'incertitude de l'information spatiale une théorie de représentation. Cette approche apporte une solution pratique pour le choix de la représentation dans la construction méthodologique du système d'information. La taxonomie de Peter Fisher est une adaptation de celle présentée dans la figure 1.7 aux problèmes d'incertitudes spécifiques à l'information spatiale. Cette taxonomie est présentée dans la Figure 3.1.

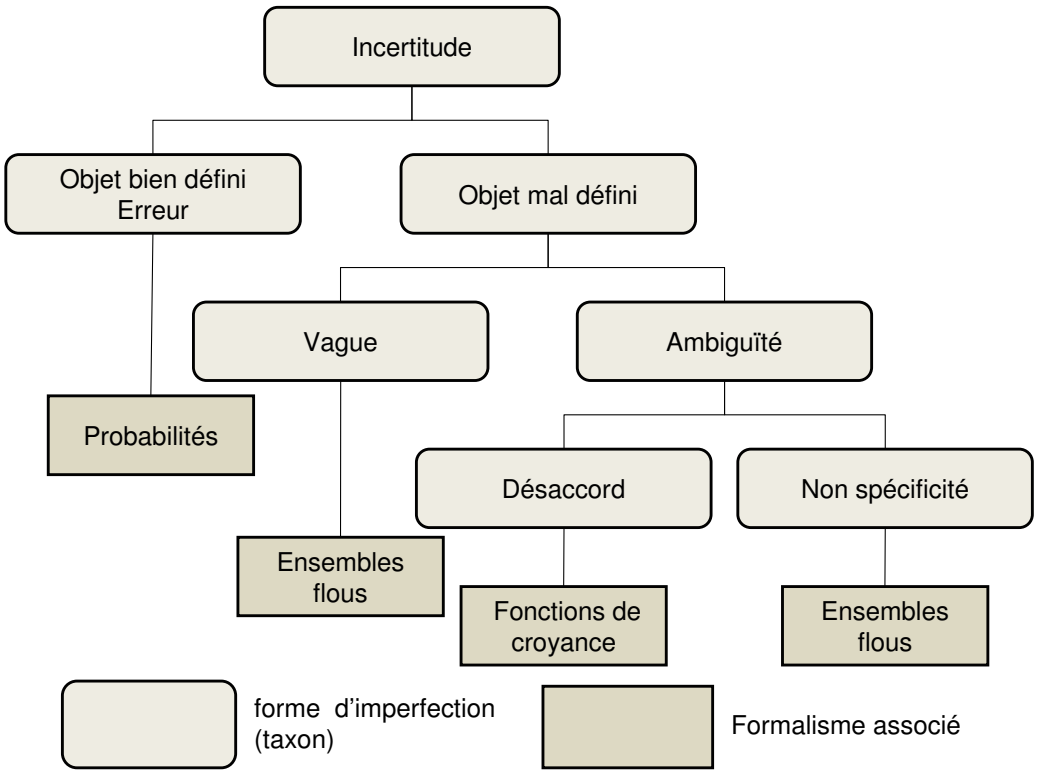
Dans son approche, le sens donné à l'incertitude n'est pas exactement le même que celui proposé dans ce mémoire. Il est plus proche de la notion d'imperfection que d'incertitude. En adaptant la taxonomie de Fisher au vocabulaire explicité dans le chapitre 1, on obtient celle présentée dans la Figure 3.2.

Fisher propose de distinguer les objets bien définis pour lesquels seule la validité de l'information peut être mise en cause (incertitude), des objets mal-définis pour lesquels il sépare l'imprécision et l'ambiguïté. L'ambiguïté est spécialisée en non-spécificité et conflit. Il propose de modéliser les objets incertains par la théorie des probabilités, les objets imprécis par des ensembles flous, les informations conflictuelles par la théorie de l'évidence, et la non-spécificité par la théorie des ensembles flous. Ces choix permettent une lecture synthétique en vue du choix de la représentation. Ces associations imperfections-théories semblent naturelles.

En effet, la théorie des probabilités est un outil puissant pour modéliser la fiabilité d'une source ou encore évaluer les erreurs. La probabilité qu'un instrument d'observation soit défectueux est par exemple une bonne indication sur la fiabilité de l'observation faite avec cet instrument. La probabilité subjective en la fiabilité d'une source est aussi à prendre en considération. Le modèle probabiliste est donc un bon outil pour la représentation de l'incertitude des objets bien formés.

La théorie des ensembles flous a été définie pour la représentation de notions imprécises ; la choisir pour la formalisation d'objets imprécis semble logique. La relation spatiale « proche de » ou la catégorie « est vieux » sont facilement représentables à l'aide

Figure 3.1 – Taxonomie de l'incertitude de l'information géographique selon [Fisher, 2005] et [Fisher *et al.*, 2005]

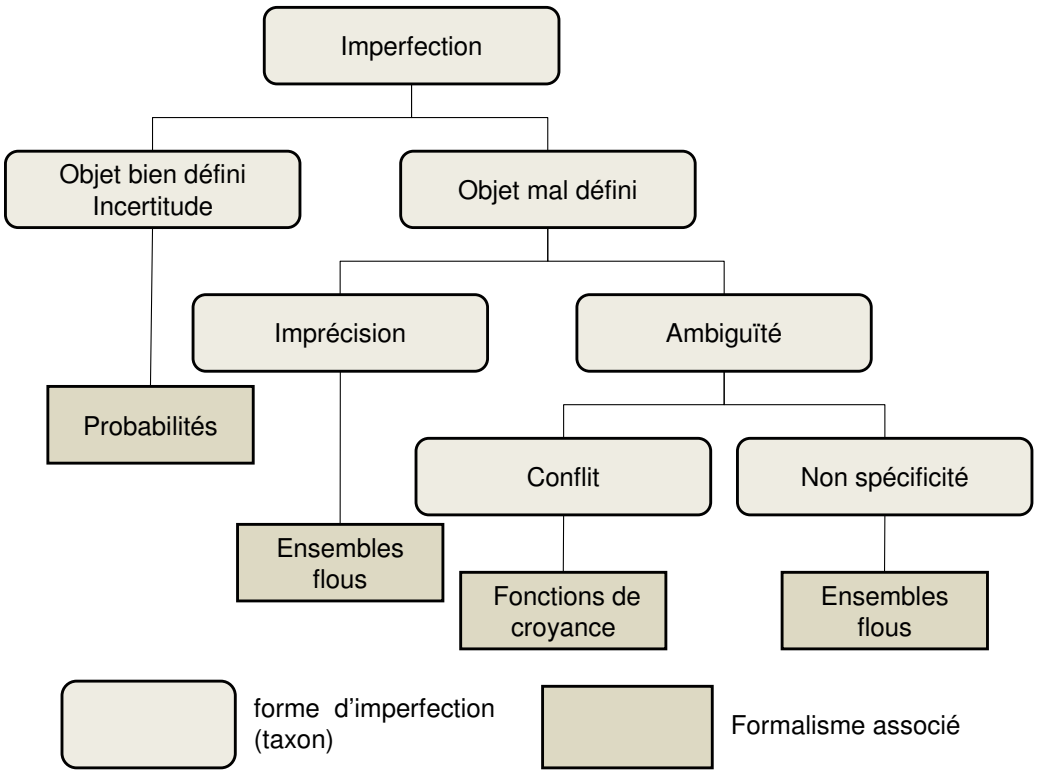


d'ensembles flous. Les représentations et modélisations qui en sont issues permettent de considérer l'objet ou la relation plus finement. La quantification de l'appartenance permet de mieux appréhender un phénomène. On peut, pour la catégorie « proche de », à l'aide d'une représentation en ensemble flou à la fois associer aux objets distants de « moins de 5 mètres » une appartenance maximale (de valeur 1) à la catégorie issue de la relation « proche de » et nuancer l'appartenance des objets plus distants à cette catégorie.

Lorsqu'il y a ambiguïté, les théories à utiliser doivent combiner modélisation de l'incertitude et de l'imprécision. Parmi les différentes théories possibles, deux (théorie de l'évidence, théorie des possibilités) permettent de représenter directement ces deux types d'imperfections, et deux (théorie des ensembles flous, théorie des ensembles approximatifs) permettent de les représenter à l'aide des logiques associées aux représentations (logique floue, logique approximative, voire logique modale).

Si un conflit existe pour la définition d'un objet, le choix se fait entre plusieurs entités disjonctives ; en affectant des masses de croyances à ces entités, le choix est guidé à la fois par l'imprécision de l'information et la validité de chaque entité. Considérons l'exemple de la figure 1.3, en construisant deux fonctions de croyance (une pour pavillon, une pour

Figure 3.2 – Taxonomie de l'imperfection de l'information géographique adaptée de [Fisher, 2005] et [Fisher *et al.*, 2005]



immeuble), on peut en fonction de la position de la zone mixte attribuer à chacune des possibilités une masse de croyance, ainsi qu'une masse de croyance à l'ignorance sur la catégorisation de l'objet. Ainsi, toutes les hypothèses sont considérées et le choix de la catégorie est justifié mais nuancé par la croyance en ce choix.

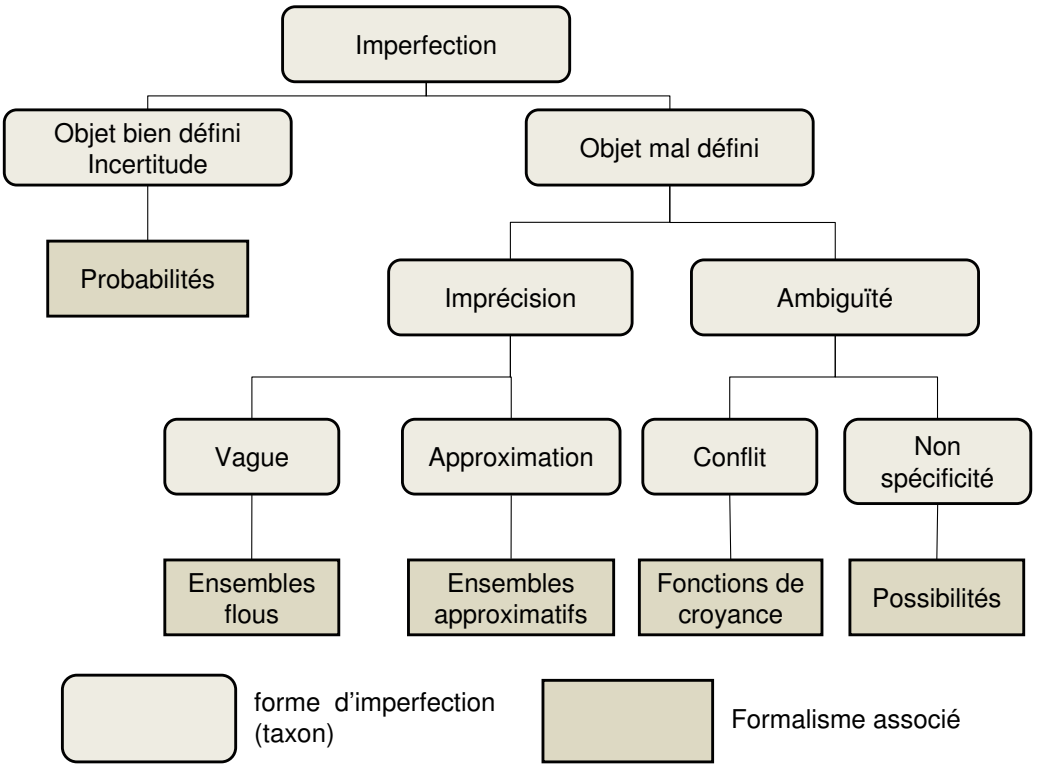
Le choix de la théorie des ensembles flous pour la non-spécificité se justifie par la représentation d'objets imprécis. La logique floue utilisée ensuite permet de lever l'ambiguïté associée à ces cas. Par exemple, pour la relation « au nord de », en associant à chaque définition un ensemble flou, on représente la pluralité des définitions mais aussi leurs imprécisions. En utilisant la logique floue, on peut en fonction du contexte choisir la meilleure.

Par cette taxonomie, le choix du formalisme de représentation est explicité par la nature des données. Je propose d'adapter cette approche pertinente pour l'information géographique au contexte de l'information archéologique et des formes d'imperfections associées (voir chapitre 1). J'utiliserai ensuite la taxonomie résultante de cette adaptation pour modéliser les données archéologiques contenues dans *BDRues*.

3.1.2 Cas de l'information archéologique

Les différences entre information géographique et archéologique impliquent de définir une taxonomie propre aux imperfections des données archéologiques. Dans ce mémoire, je propose la taxonomie présentée dans la figure 3.3. Elle a pour but de déterminer en fonction des types d'imperfection, le meilleur choix de représentation des données archéologiques dans le SIG. Elle constitue une adaptation de celle de Fisher au contexte de l'information issue de fouilles archéologiques. Je propose dans celle-ci de spécialiser l'imprécision au travers des deux sous-types : le vague et l'approximation (voir 1.2).

Figure 3.3 – Taxonomie de l'imperfection de l'information archéologique



Pour le cas où l'information est vague, je propose d'utiliser la théorie des ensembles flous (à l'instar du choix associé à l'imprécis dans [Fisher *et al.*, 2005]). Ce choix permet de représenter la flexibilité des connaissances. Ainsi pour le Bas-Empire, le support de l'ensemble flou le représentant commencera à la date 193 et son noyau débutera à la date 284. Par cette représentation, le vague de l'information est pris en considération. Grâce à cette spécialisation de l'imprécision, le choix de la théorie de représentation est clarifié.

Si l'information est approximative, je propose de la modéliser à l'aide de la théorie des ensembles approximatifs. Prenons le cas d'un mur à la fois défini à l'aide de ses traces et

de pierres le composant, l'ensemble des pierres le composant est considéré comme l'approximation haute du mur, tandis que l'approximation basse du mur englobe à la fois les traces et les pierres. Pour le mur ainsi représenté, les pierres appartiendront nécessairement au mur, et les traces seront possiblement dans le mur. Dans ce cas, la quantification de l'appartenance apparaît difficile à déterminer, les ensembles approximatifs semblent donc plus appropriés.

Une seconde différence entre la taxonomie proposée et celle de Fisher porte sur le choix de la représentation de la non-spécificité. Fisher propose, dans ce cas, d'utiliser les ensembles flous et la logique floue tandis que je propose de modéliser les objets à l'aide de la théorie des possibilités. La théorie des possibilités est régulièrement utilisée pour résoudre les problèmes d'expertises multiples comme le sont les cas de non-spécificité. Prenons l'exemple d'une expertise donnée à l'échelle du *fait* qui associe l'objet à un mur et d'une autre à l'échelle de la *structure* qui inclut l'objet dans une maison. À l'aide de distributions de nécessité et de possibilité pour chacune des expertises, la modélisation de l'objet considèrera à la fois la nécessité et la possibilité que l'objet soit un mur et qu'il appartienne à une maison, afin de définir l'objet en fonction de l'échelle d'analyse.

Ainsi, l'archéologue peut choisir de manière plus affinée le formalisme, tant pour le spatial que pour le temporel et le descriptif. On peut donc prendre en compte les imprécisions auxquelles sont soumises les données archéologiques dues à leur distanciation temporelle avec le temps de l'enregistrement, aux possibles variations géologiques, ou encore à la lacune de l'information.

3.1.3 Vision générale

On peut se rendre compte que les différentes taxonomies que je viens de présenter ont chacune pour but de donner une grille de lecture pour la représentation de l'information imparfaite. Elles proposent de représenter l'information par le prisme d'un formalisme de représentation. Ces grilles présentent donc un intérêt méthodologique essentiel pour la gestion de l'information dans les SIG.

Ainsi, le contributeur (le fournisseur de données) dispose d'une panoplie d'outils pour la modélisation de l'information imparfaite dans un SIG. Celle-ci s'appuie sur des modèles de représentation adaptables aux données.

3.2 Modélisation des données spatiotemporelles imparfaites

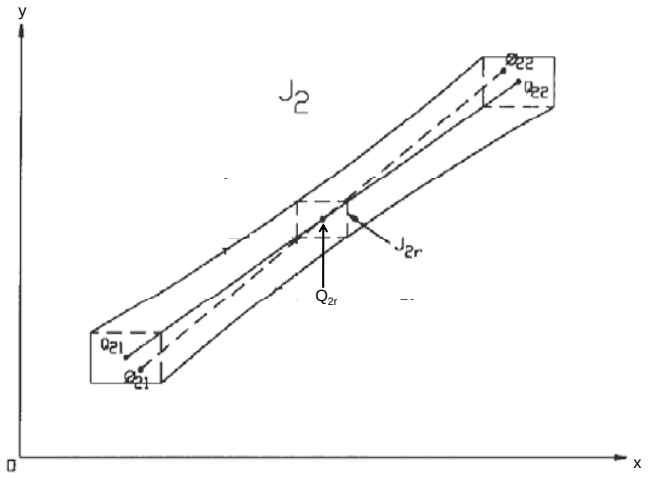
La modélisation de l'imperfection est un thème d'actualité au sein des Sciences de l'Information Géographique. En effet, que ce soit pour modéliser spatialement ou temporellement les objets, l'utilisation de modèles est fondamentale dans un objectif de simulation et d'uniformisation de traitement. Les enjeux et la complexité de la modélisation

sont exposés dans [Couclelis, 2005].

3.2.1 Modélisation spatiale

De nombreux modèles de représentation d'objets spatiaux aux frontières indéterminées sont proposés dans [Burrough et Frank, 1996]. Ces modèles peuvent être probabilistes, flous, possibilistes, approximatifs. Les fonctions de croyance sont plus utilisées dans le cadre de l'étude des relations topologiques entre objets spatiaux.

Figure 3.4 – Exemple de modélisation probabiliste de ligne : modèle de région de confiance pour le segment 2D (Q_{21}, Q_{22}) — source : [Shi, 2007]



[Shi, 2007] présente une revue des modèles probabilistes associés à la qualité de la donnée. La figure 3.4 illustre une des possibilités pour la modélisation des lignes. Dans cette modélisation, la région de confiance [47] (J_2) est l'union de toutes les régions de confiance (J_{2r}) des points (Q_{2r}) de la ligne.

Dans [Fisher, 1996], Fisher présente une étude comparative entre les modélisations utilisant la théorie des ensembles et les modélisations utilisant la théorie des ensembles flous pour la modélisation du territoire. Les unes simplifient la modélisation au risque d'erreurs, les autres complexifient la modélisation et le traitement. Dans [Dilo *et al.*, 2004], la notion d'objet vague est définie ; cette notion est une autre vision des modélisations floues. Une illustration pour la définition de zones floues est présentée dans la figure 3.5.

47. Une zone ou une région de confiance est un ensemble de localisations qui est supposé contenir, avec un certain degré de confiance, la localisation réelle de l'objet. Par exemple, une région de confiance à 95% (ou au seuil de risque de 5%) a 95% de chance de contenir la valeur du paramètre que l'on cherche à estimer mais cet intervalle de confiance est trompeur dans 5% des cas. Le seuil de confiance est à définir en fonction des applications.

Figure 3.5 – Exemple de modélisation floue de régions — source : [Dilo *et al.*, 2004]

Figure 3.6 – Exemple de modélisation approximative de région

Les approches privilégiant un modélisation par ensemble approximatif (figure 3.6) traitent l'imperfection mais donnent peu d'informations numériques durant le traitement. L'évaluation des objets se superposant spatialement est alors effectuée par l'évaluation de la qualité des zones.

Navratil propose d'utiliser la théorie des possibilités pour gérer la précision des mesures ; ils utilisent des nombres flous pour représenter les marges d'erreurs [Navratil, 2007]. La modélisation des territoires et de leurs frontières par des ensembles flous et possibilistes est aussi utilisée comme dans [Rolland-May, 2000] et [De Ruffray, 2007].

Les modélisations spatiales peuvent à la fois porter sur des objets ponctuels, des lignes, des polygones. Les modèles de diffusion peuvent être gaussiens (modélisation probabiliste), des ensembles flous convexes et normalisés (modélisation floue ou possibiliste) et des ensembles approximatifs. Ils auront tendance à élargir le champs de définition par rapport à une modélisation utilisant des ensembles classiques.

3.2.2 Modélisation temporelle

De nombreux modèles pour la gestion temporelle des objets géographiques existent. La plupart sont probabilistes à l'instar de [de Wit et de Bruin, 2006]. Ces modèles ont pour objectif la simulation de l'évolution territoriale entre deux dates précises.

Cependant, comme le remarquent Suzana Dragicevic et Danielle Marceau, le temps écoulé entre les dates nécessaires à la gestion de l'information temporelle (date de l'observation, date du stockage, date de l'analyse) implique que l'information temporelle peut aussi être considérée comme imprécise [Dragicevic et Marceau, 2000a]. Dans ce contexte, elles proposent d'utiliser une modélisation floue des espaces en fonction du temps. Cette approche n'est pas applicable aux données que j'ai manipulée ; ces objets n'ont pas pour objectif de représenter la période actuelle ou la période de fouille, mais une durée dans le passé.

Toutefois, Shokri *et al.* présentent les avantages et les inconvénients d'une modélisation du temps en périodes imparfaites par rapport à une représentation ponctuelle imparfaite du temps [Shokri *et al.*, 2006]. Dubois *et al.* proposèrent en 2003, dans [Dubois *et al.*, 2003], une étude de la modélisation du temps à l'aide de la théorie des ensembles flous et de ses dérivées. Les périodes sont dans ce cadre représentées par des nombres flous trapézoïdaux. Ce choix permet une représentation vague des périodes associées aux objets à modéliser.

3.2.3 Modélisation de l'information imparfaite

La modélisation de l'information imparfaite peut ne pas porter uniquement sur les composantes temporelles ou spatiales. Elle peut aussi porter sur la représentation des composantes descriptives et topologiques. Par exemple, pour les objets de *BDRues*, la composante orientationnelle est elle aussi imparfaite. Une modélisation des orientations avec leurs imperfections s'impose donc.

Il existe cependant des approches de modélisation de l'imparfait et de l'information spatiotemporelle dans les SIG n'utilisant aucune des théories présentées dans le chapitre précédent. Par exemple, les logiques modales [48] peuvent être utilisées à la fois pour la représentation de l'information [Papini, 2007b] et pour le raisonnement sur l'information [Papini, 2007a].

Les modélisations dépendent des objectifs et du choix de la théorie de représentation. À l'instar des théories associées, le choix de modélisation est déterminant dans le cadre de la pensée complexe.

Je présente dans la section suivante les choix faits pour la représentation et la modélisation des objets archéologiques contenus dans *BDRues* ; ces choix conduisent à la création d'une base de données imparfaites nommée *BDFRues*.

48. Les logiques modales sont toutes des extensions de la logique propositionnelle classique pour lesquelles on dispose d'opérateurs représentant une modalité à l'instar des opérateurs classiques représentant les modalités *possible* et *nécessaire*.

3.3 Application aux données archéologiques

Dans la base de données *BDRues*, les objets représentent les tronçons de rues datant de l'époque romaine trouvés à Reims lors de fouilles durant les 20 dernières années. Ce sont donc des éléments significatifs de Durocortorum. Dans ces objets, je distingue principalement trois caractéristiques :

- la localisation,
- la datation,
- l'orientation.

Ces trois caractéristiques représentent des composantes de l'information archéologique : la localisation et l'orientation correspondent à la composante spatiale, la datation, qui est la période d'activité, est associée à la composante temporelle.

Dans la suite de ce mémoire, un objet p de *BDRues* a une composante spatiale $p.Loc$, une composante temporelle $p.Date$ et une composante topologique ou orientationnelle $p.Orien$. Le choix de la représentation se fait donc pour chaque composante en fonction de l'imperfection possible de l'information contenue.

3.3.1 Choix de la théorie de représentation

Dans le contexte de l'information archéologique, l'estimation des périodes d'activité des objets archéologiques est soumise à la fois à des connaissances subjectives (estimation issues d'expertises) et sont aussi flexibles (sémantiques à signification variable). Par exemple dans *BDRues* certaines périodes d'activité sont définies comme « Haut Empire » dont les dates de début et de fin varient en fonction des experts. Les périodes d'activité sont donc sujettes à du vague. Pour leurs représentations dans le SIG, je propose, selon la taxonomie présentée dans la figure 3.3, d'utiliser la théorie des ensembles flous.

Les localisations sont à la fois soumises à des instrumentations, dont les GPS, ayant des marges d'erreurs mais aussi à des géoréférencements utilisant des techniques plus anciennes notamment des relations spatiales (en face de, à trois mètres de, etc.). De plus, dans le cas des données de *BDRues*, les objets n'ont été spatialement représentés que par un unique point auquel une orientation est associée, ce qui génère des problèmes de précisions lors des changements d'échelle par exemple. C'est pour cela que je considère que les composantes spatiales peuvent être considérées à la fois comme vagues et approximatives ; elles sont donc imprécises. En se reportant à la taxonomie de [Fisher *et al.*, 2005], la théorie des ensembles flous est dans ce cas aussi une bonne solution pour la représentation.

Ainsi, les trois principales composantes des tronçons de rues romaines seront représentées par des ensembles flous dans le SIG. Je propose donc de construire, à partir des objets de *BDRues*, une nouvelle base de données *BDFRues* comportant des objets ayant une composante temporelle floue ($fDate$), une composante spatiale floue ($fLoc$) et une composante orientationnelle floue ($fOrien$). À chaque objet p de *BDRues* ayant pour composantes $p.Date$, $p.Loc$ et $p.Orien$ est associé un objet pf de *BDFRues* ayant pour

composantes respectives $pf.fDate$, $pf.fLoc$ et $pf.fOrien$.

L'information archéologique présente aussi de l'absence et des lacunes. L'incomplétude de l'information n'est pas incompatible avec la théorie des ensembles flous. Le traitement des données devra en tenir compte.

Il s'agit maintenant de déterminer quelles seront les formes des fonctions d'appartenance $pf.fdate$, $pf.floc$ et $pf.forien$ de respectivement $pf.fDate$, $pf.fLoc$ et $pf.fOrien$ (voir 2.3).

3.3.2 Choix des modèles

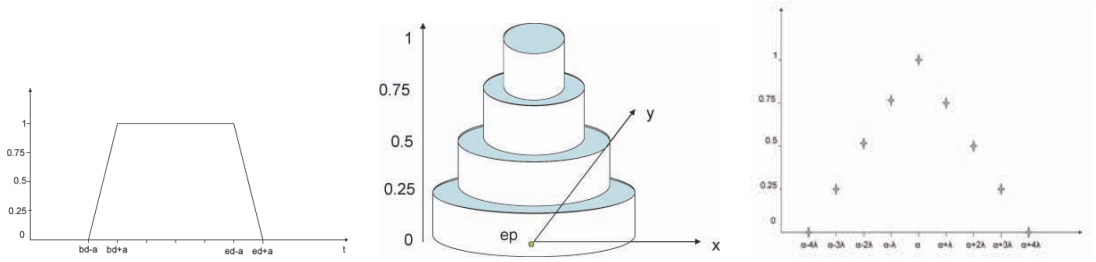
La théorie de représentation choisie pour les trois composantes des objets de *BDFRues* est la théorie des ensembles flous. Dans ce cadre, le choix de la modélisation temporelle des périodes d'activité se porte naturellement vers des nombres flous de forme trapézoïdale basée sur les composantes temporelles des objets de *BDRues*.

Ce choix permet, pour un objet pf de période d'activité $pf.fDate$, d'affecter à une date d un degré de confiance à 1 ($pf.fdate(d) = 1$) dès lors qu'il est certain que pf était en activité à la date d. De manière analogue, si à la date d, on est sûr que pf n'était pas en activité, alors la confiance associée à la présence de pf à la date d est nulle ($pf.fdate(d) = 0$). Le modèle utilisé pour la modélisation temporelle des objets de *BDFRues* est présenté Figure 3.7a. Les représentations des périodes ainsi construites seront appelées périodes floues.

Pour la modélisation de la représentation spatiale, les objets de *BDRues* ayant des points pour représentation spatiale, les modélisations floues associées à la géométrie des objets de *BDFRues* seront des points flous. Le modèle associé à la gestion des points flous est une diffusion circulaire centrée sur les points représentant spatialement les objets correspondants dans *BDRues*. Dans cette modélisation, pour un objet pf de *BDFRues* correspondant à un objet p de *BDRues*, la confiance associée à la présence de l'objet pf à la localisation $p.Loc$ est de 1 ($pf.floc(p.Loc) = 1$). Cette confiance diminue plus on s'éloigne de $p.Loc$. Le modèle utilisé pour la localisation de *BDFRues* est exposé Figure 3.7b.

Pour la modélisation de l'information orientationnelle (ou directionnelle) des objets de *BDFRues*, les modélisations floues associées sont des nombres flous triangulaires car les composantes orientationnelles des objets de *BDRues* sont de simples angles. Le modèle associé à la gestion des angles flous est un nombre flou triangulaire. Dans cette modélisation, pour un objet pf de *BDFRues* correspondant à un objet p de *BDRues*, la confiance associée à l'orientation $p.Orien$ dans la composante orientationnelle $pf.fOrien$ est de 1 ($pf.fOrien(p.Orien) = 1$). Cette confiance diminue plus l'angle est distant de $p.Orien$. Le modèle utilisé pour l'orientation de *BDFRues* est exposé Figure 3.7c.

Figure 3.7 – Modèles flous pour la localisation, l'orientation et les périodes d'activité des tronçons de rues romaines



a : Période d'activité b : Localisation c : Orientation

◇◇◇◇◇

En conclusion, ce chapitre permet de définir une taxonomie de l'imperfection au contexte de l'information archéologique. Dans cette taxonomie, une théorie de représentation de l'imperfection est associée à chaque taxon de plus bas niveau. Ce chapitre étudie ensuite la modélisation spatiotemporelle des données selon chaque théorie de représentation. À l'aide de la taxonomie et de l'étude des modèles, les données de *BDRues* sont modélisées en utilisant la théorie des ensembles flous et des modèles flous. Cette méthode implique la création d'une nouvelle base de données : *BDFRues*.

Dans cette base les objets archéologiques sont considérés selon leurs imperfections et donc selon leur qualité. J'ai donc préparé les données sur les tronçons de rues de Reims datant de l'époque romaine en vue de leur analyse.

Le travail présenté dans le chapitre 1 et celui-ci a été présenté lors de la conférence internationale "36th Annual Conference on Computer Applications and Quantitative Methods in Archaeology" en 2008 (voir [de Runz *et al.*, 2008e]).

Conclusion

L'objectif de cette partie fut d'exposer une démarche permettant une modélisation de l'information archéologique prenant en considération autant l'aspect spatial et temporel que son caractère imparfait. Cette démarche a pour but de préparer les données en vue de leur analyse (voire de la généralisation), et permet leur modélisation en prenant en considération leurs imperfections. Ainsi, la démarche proposée regarde les données dans leurs complexités.

La démarche exposée est structurée en trois étapes. Durant la première étape la nature de l'imperfection des objets archéologiques est étudiée ; les différentes formes de l'imperfection sont alors organisées dans une typologie. À partir de cette typologie et de l'étude des différents formalismes représentant l'imparfait, une taxonomie associant les types d'imperfections et les théories est construite. À l'aide de cette taxonomie, dans la seconde étape, la théorie de représentation de l'imperfection nécessaire à la modélisation des données est choisie. La troisième étape porte sur la modélisation des données. Celle-ci se fait via des modèles pour uniformiser le traitement, ces modèles variant selon la théorie de représentation choisie.

Le premier chapitre aborda la question de la nature de l'imperfection des données archéologiques. Pour cela, une description sémantique des termes utilisés dans ce mémoire est illustrée par des exemples courants mais aussi archéologiques. Une typologie de l'imperfection est ainsi proposée.

La méthodologie nécessitant le choix entre différentes théories de représentation formelle de l'imparfait, le second chapitre présenta différentes théories de représentations de l'imparfait. Les théories sont choisies parmi les plus courantes pour la modélisation des données dans les SIG.

Le troisième chapitre a lui porté sur les deux dernières étapes de la méthodologie. Dans ce chapitre, une taxonomie associant les types d'imperfection et les théories de représentation est proposée. Ensuite, en fonction de la théorie de représentation, une étude de la modélisation des données est exposée. Enfin, l'utilisation de la démarche pour la modélisation est illustrée sur les données de *BDRues* afin de les modéliser en tenant compte de leurs imperfections. De cette modélisation émerge une nouvelle base de données intitulé *BDFRues*.

La base de données archéologiques *BDFRues* est le support applicatif des prochaines parties. Celles-ci porteront sur l'analyse et la visualisation des données archéologiques dans

un SIG dans un objectif de permettre la généralisation de l'information archéologique. Dans ce contexte, la partie suivante étudiera l'analyse, à l'aide d'informations externes, des données spatiotemporelles imparfaites contenues dans une BDG associée au SIG.

Deuxième partie

Analyse de données spatiotemporelles imparfaites dans un SIG archéologique

« Life was a process of finding out how far you could go, and you could probably go too far in finding out how far you could go. »

Terry Pratchett, *Annals of the Discworld — Hogfather* (1996).

« “Pretty please with sprinkles on top’ is not a recognized method of interrogation!” »

Terry Pratchett, *Annals of the Discworld — Monstrous Regiment* (2003).

Introduction

Les SIG sont classiquement utilisés comme outils d'aide à la décision par le biais d'analyses de l'information à références spatiales. Ces analyses ont pour but de permettre notamment la généralisation des données afin d'insérer les informations dans leur espace. Elles ont besoin, pour cela, de techniques permettant de visualiser le positionnement de chaque objet contenu dans la base de données soit par rapport à de l'information externe (ou en entrée), soit par rapport aux autres objets de cette base (on parle alors d'exploration de données ou d'analyse exploratoire). Cette partie se situe dans le premier cas, le second étant l'objet de la partie III. Je m'intéresserai ici aux interactions entre les objets de la base de données associée au SIG et des éléments externes. Les méthodes d'analyse proposées permettent soit l'interrogation de la base de données, soit l'appariement de données externes avec celles de la base associée au SIG.

D'une manière générale, l'interrogation de la base consiste à traiter les données les unes après les autres. Une vision algorithmique de l'interrogation est proposée dans l'algorithme 3.1. L'interrogation de données consiste donc à visualiser dans un SIG le positionnement de chaque objet de la base par rapport à de l'information externe grâce à un traitement des données. Les méthodes d'interrogation présentées dans cette partie différeront donc d'une part par l'information qu'elles prennent en entrée et d'autre part par le traitement des données.

Algorithme 3.1 : Processus d'interrogation d'une base de données *bd* selon l'information externe *ie*

pour chaque ***objet*** *o* ***de*** *bd* **faire**
  | traitement(*o*,*ie*)
fin

L'appariement consiste à associer le mieux possible un objet issu d'une première source d'information à un objet issu d'une seconde source. Les techniques d'appariement d'objets utilisent des critères de similarité. Dans un SIG, chaque source est classiquement associée à une base de données géographiques. L'appariement se fait généralement entre des informations portant sur un même territoire et les critères correspondent à des distances entre objets. Plusieurs processus d'interrogation sont généralement utilisés par les techniques d'appariement dans un SIG. L'algorithme 3.2 présente la démarche générale de

l'appariement entre objets appartenant à deux bases de données différentes.

Algorithme 3.2 : Processus d'appariement entre une base de données $bd1$ et une base de données $bd2$ selon l'ensemble de critères cr

```
pour chaque objet o1 de bd1 faire
    pour chaque objet o2 de bd2 faire
        Calcul de cr(o1, o2)
    fin
    appariement de l'objet o1 avec l'objet o2 ayant le meilleur cr(o1, o2)
fin
```

Dans le contexte de l'interrogation de bases de données archéologiques, il est naturel de vouloir positionner chronologiquement l'ensemble des données par rapport à une période (ou une date) donnée en entrée. Malheureusement s'il est trivial d'obtenir l'antériorité entre deux dates parfaitement déterminées, cela devient beaucoup plus difficile de comparer des dates vagues ou des périodes représentées par des nombres flous. De plus, afin de pouvoir affiner leurs modèles, les experts ont besoin de pouvoir remettre en cause telle ou telle datation. Pour ce faire, je propose d'évaluer quantitativement l'antériorité entre deux nombres flous par un degré que j'appelle : indice d'antériorité. Pour calculer cet indice d'antériorité, je me place dans la démarche d'une des méthodes d'ordonnancement de deux nombres flous les plus intuitives — le Fuzzy Max Order (FMO) introduit par Ramik et Raminek en 1985 [49] — et je m'appuie aussi sur l'approche que Kerre proposa en 1982 [50]. Cet indice regarde la proximité relative de ces deux nombres flous à leur maximum ($\widetilde{max}$) et à leur minimum ($\widetilde{min}$) définis selon le principe d'extension de Zadeh [Zadeh, 1965]. Je détaillerai cette démarche dans le chapitre 4.

L'interrogation que je propose se base sur le calcul des valeurs de l'indice d'antériorité d'un objet par rapport à un autre. Ce calcul utilise l'indice proposé par Kerre pour l'ensemble formé des deux nombres flous représentant les périodes d'activité des deux objets en question. Pour la visualisation des résultats dans le SIG, le traitement de l'interrogation temporelle des données affecte à chaque objet une couleur attribuée en fonction de l'antériorité de l'objet à la période demandée. Ainsi l'expert peut interpréter visuellement à l'aide d'un *SIG* les résultats de son interrogation. Cette analyse temporelle est le cœur du chapitre 4.

De plus, il est naturel, quand on a des données spatiotemporelles imparfaites, de proposer des analyses tenant compte de l'imperfection des données. Dans le cadre du projet *SIGRem*, les objets contenus dans *BDFRues* correspondent à des tronçons de rues de Reims datant de l'époque romaine. Un des objectifs de l'utilisation d'un tel SIG est alors de mettre en avant la structure des rues de la ville en fonction de périodes fournies

49. Ramik et Raminek ont introduit le FMO dans [Ramik et Rimanek, 1985]
50. Kerre proposa le calcul d'un indice de position dans [Kerre, 1982].

en entrée. Pour cela, je propose une nouvelle méthode de traitement des objets pour l'interrogation de la base de données floues. Cette méthode se base sur la qualité des données et notamment sur le fait qu'elles sont lacunaires. Le traitement a pour objectif la complétion, au moins partielle, de ces informations.

Afin de compléter l'information, j'exploite les caractéristiques de forme des objets : les rues romaines ont la particularité d'être à priori linéaires. Le traitement utilise cette information pour la spécification d'une méthode de reconnaissance de formes particulières : la transformée floue de Hough [Han *et al.*, 1993]. Celle-ci se base notamment sur un accumulateur dans l'espace paramétrique des formes et sur une étape de sélection de l'information pertinente. Le traitement proposé, qui est l'objet du chapitre 5, a en entrée à la fois une période (information temporelle) et une forme (information géométrique) qu'il associe à la localisation des objets. L'analyse construite est donc spatiotemporelle.

Cependant, l'analyse de données dans un SIG ne porte pas seulement sur l'interrogation des données. En effet, analyser des données issues de sources différentes permet de dégager par exemple les relations entre des objets de bases différentes. Pour cela, il est souvent nécessaire de lier entre eux les objets de sources différentes. Cette recherche de lien se fait souvent à l'aide de techniques d'appariement de données basées sur des critères géométriques, topologiques, sémantiques ou encore toponymiques [Olteanu *et al.*, 2006]. Une approche pour l'appariement de données géographiques, qui s'applique quels que soient les critères choisis, est proposée dans [Olteanu, 2007a]. Cette approche, utilisée sur des données archéologiques, propose de passer par la théorie des fonctions de croyance pour orienter la décision lors de l'appariement des données. Ces fonctions de croyance sont définies grâce à des distances entre les différents éléments [Olteanu, 2007b].

En se basant sur cette approche, deux pistes sont proposées pour permettre l'analyse multi-source des données archéologiques. D'une part, on peut remarquer que, dans le cas général, les techniques d'appariement n'utilisent généralement pas de critère temporel. Il serait donc intéressant de proposer des critères propres au contexte archéologique comme cela est fait en géographie [Abadie *et al.*, 2007]. D'autre part, dans l'optique d'apporter plus de robustesse [51] vis-à-vis des données imparfaites, il serait judicieux d'ajouter une étape entre le calcul des critères et l'obtention des masses de croyance. Par analogie avec les statistiques d'ordre réputées pour leur robustesse, cette étape consistera à passer par les rangs définis selon les distances. Le chapitre 6 portera sur l'analyse multi-source de données archéologiques.

Cette partie présente donc un panel de méthodes d'analyse portant sur l'interaction entre les données déjà présentes dans le SIG dédié à l'archéologie avec de l'information donnée en entrée aux processus d'interrogation ou d'appariement. Ces informations peuvent être une période floue, une information de forme ou une base de données. Cette partie

51. En statistiques, la robustesse d'un procédé est sa capacité à ne pas être modifié par une petite variation dans les données ou dans les paramètres du modèle choisi pour l'estimation

a donc pour optique l'analyse des données spatiotemporelles imparfaites dans un SIG en fonction d'informations externes au système.

Chapitre 4

Analyse temporelle : interrogation selon l'antériorité

Sommaire

Les SIG permettent d'associer l'information contenue dans les BDG à une spatialisation et aussi de l'analyser. Dans cette optique, l'interrogation de la base de données du SIG est un outil essentiel à l'analyse géographique [Denègre et Salgé, 1996] et, pour le cas des données archéologiques, à l'expertise archéologique. Dans ce but, déterminer les éléments antérieurs à une période donnée s'avère pertinent et même indispensable. L'objet de ce chapitre est de mieux cerner le concept d'antériorité dans le contexte de données imparfaites (lacunaires et imprécises).

Dans un système de base de données spatiotemporelles, il est naturel d'extraire les objets selon leurs composantes temporelles et de les comparer. Une des comparaisons les plus usuelles est l'antériorité : « Cet objet est-il antérieur à telle date ou période ? »

Ces interrogations temporelles utilisent d'ordinaire des traitements se basant sur l'opérateur « inférieur ou égal » ($\leq$). En effet, on peut dire qu'un objet daté de l'année 1980 est antérieur à un objet daté de 2008 car $1980 \leq 2008$. Ces traitements qui sont donc naturels et aisés quand les objets sont des événements ponctuels, deviennent plus difficiles

quand leurs composantes temporelles sont des périodes, et encore plus si ces périodes sont vagues. Le sens de la question posée est alors important, car si on peut dire que la période $[1980, 1989]$ est antérieure à la période $[2000, 2008]$, la première n'est pas forcément inférieure ou égale à la seconde (le nombre d'années de la première étant supérieur à celui de la seconde). De plus, quand les périodes se chevauchent, comme par exemple les périodes $[1980, 2008]$ et $[2000, 2007]$, définir l'antériorité devient très subjectif. Dans le cas de *BDFRues*, les périodes d'activité étant représentées par des nombres flous (voir 3.3.2), considérer les objets selon leur antériorité est encore plus complexe. Il est donc nécessaire de résoudre le problème de l'interrogation pour des données imparfaites.

Pour résoudre ce problème, il serait naturel d'utiliser une méthode d'ordonnancement de nombres flous. Malheureusement, il n'existe pas de relation d'ordre totale sur les nombres flous. Cependant il existe de nombreuses techniques (voir [Wang et Kerre, 2001a] et [Wang et Kerre, 2001b]) définissant l'opérateur de pré-ordre $\preccurlyeq$. Mais ces différentes techniques peuvent pour la comparaison de deux nombres flous amener à des conclusions contradictoires (voir Annexe A).

Parmi ces méthodes, certaines considèrent les notions de maximum et/ou de minimum définies selon le principe d'extension de Zadeh [Zadeh, 1965]. Ces méthodes comparent les nombres flous en fonction de leurs similarités au maximum ou au minimum. Par exemple, dans le Fuzzy Max Order (FMO) introduit dans [Ramik et Rimanek, 1985], l'ordre entre deux nombres flous résulte de l'égalité au maximum ou au minimum de l'un des deux nombres flous. Cette méthode a l'inconvénient de ne pas être un ordre total.

Afin de pouvoir être utilisée sur des nombres flous quelconques, et d'éviter le problème de la transitivité des méthodes de comparaison deux à deux [Wang *et al.*, 1995], Kerre propose de positionner les nombres flous en fonction de leur distance au maximum de tous les nombres flous utilisés. Cependant, le calcul de ce maximum devient coûteux dans le cas de grandes bases de données et ne correspond pas à la démarche d'une interrogation de base de données, celle-ci se basant généralement sur un processus visant à étudier les données les unes après les autres.

De plus, dans le cadre de leurs travaux, les experts doivent pouvoir permettre de remettre en cause tel ou tel positionnement temporel. Une comparaison au résultat binaire ne donne alors que peu d'information.

Dans l'objectif analytique de positionnement temporel des objets archéologiques en fonction d'une date en entrée, je propose pour le traitement des données selon l'antériorité de calculer un indice quantifiant l'antériorité, appelé indice d'antériorité. Cet indice se base sur celui de Kerre défini sur la restriction de l'ensemble des nombres flous aux deux nombres flous à comparer. Les valeurs de l'indice donnent une indication de fiabilité de l'antériorité demandée.

Dans ce cadre, sur les données de *BDFRues* et pour une période donnée représentée par un nombre flou, je propose un traitement des données qui, pour chaque objet de la base, calcule la valeur de l'indice d'antériorité de la période en entrée à la période d'activité

de l'objet et affecte à cet objet une couleur en fonction de la valeur de l'indice afin de visualiser les résultats dans le SIG. Ce traitement permet d'évaluer la position temporelle des objets de la base par rapport à une période en entrée. C'est donc par un traitement temporel quantitatif que j'aborderai le problème de l'antériorité d'une période aux objets de la base.

Dans ce chapitre, je présenterai en premier lieu les méthodes de comparaison sur lesquelles se base l'indice d'antériorité que je propose. Ces méthodes sont le Fuzzy Max Order et la comparaison des valeurs de l'indice de Kerre. Ensuite, je définirai l'indice d'antériorité et j'en exposerai quelques caractéristiques. Enfin, j'illustrerai l'intérêt du calcul de la valeur de cet indice dans le contexte archéologique, ce qui me permettra de proposer une nouvelle forme d'interrogation temporelle tenant compte de l'imperfection des données.

4.1 Comparaison de nombres flous

Les méthodes de comparaison de nombres flous exposées ici se basent sur les notions de maximum et de minimum de nombres flous exprimées selon le principe d'extension de Zadeh [Zadeh, 1965]. Le FMO et l'indice de Kerre sont issus de la similarité entre les nombres flous à comparer et le maximum (ou le minimum).

Le Fuzzy Max Order (FMO) est l'une des méthodes les plus naturelles pour la comparaison de deux nombres flous. Il repose sur l'idée que si un nombre flou est égal au maximum (resp. minimum) des deux nombres à comparer alors il est le plus grand (resp. le plus petit). La méthode de Kerre considère que, dans un ensemble de nombres flous, plus un nombre est petit, plus sa distance au maximum de l'ensemble sera grande.

Dans cette section, je présenterai, dans un premier temps, les notions de maximum et minimum, ensuite le *FMO* pour finir par l'approche proposée par Kerre.

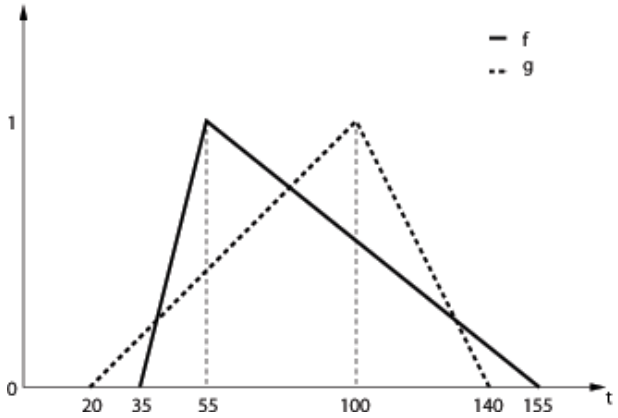
4.1.1 Notions préliminaires : le maximum et le minimum

Considérons deux nombres flous F et G quelconques de fonctions d'appartenance f et g — par exemple ceux dont les fonctions d'appartenance sont représentées dans la Figure 4.1 — en utilisant le principe d'extension de Zadeh présenté dans [Zadeh, 1965], le maximum de F et G est un ensemble flou $\widetilde{max}(F,G)$ dont la fonction d'appartenance $\widetilde{max}(f,g)$ est définie par :

$$\widetilde{max}(f,g)(z) = \sup_{\substack{x,\,y\,\in\,\mathbb{R},\\ z\,=\,max(x,\,y)}} \left(min(f(x),g(y))\right).$$

Le minimum de F et G est aussi un ensemble flou $\widetilde{min}(F,G)$ dont la fonction d'ap-

Figure 4.1 – Chevauchement de deux fonctions d'appartenance f et g des nombres flous F et G

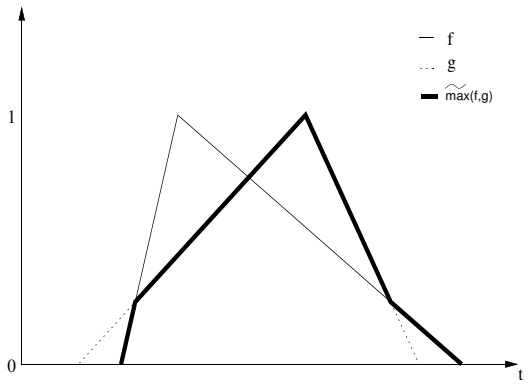


partenance $\widetilde{min}(f, g)$ est :

$$\widetilde{min}(f,g)(z) = \sup_{\substack{x,\, y \in \mathbb{R}, \\ z = min(x,\, y)}} (min(f(x), g(y)))$$

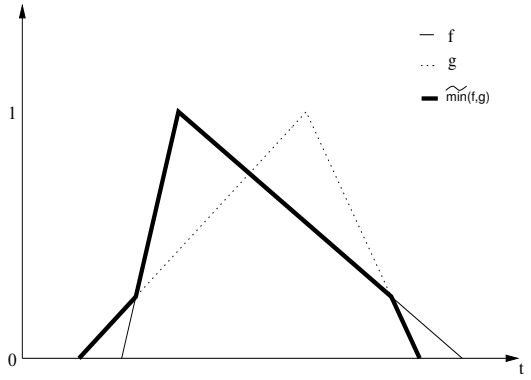
Notons que le maximum (resp. le minimum) de deux nombres flous est aussi un nombre flou (voir [Ramik et Rimanek, 1985]). Considérons maintenant les nombres flous F et G dont les fonctions d'appartenance sont représentées dans la Figure 4.1, la Figure 4.2 montre la fonction d'appartenance du $\widetilde{max}(F, G)$; la Figure 4.3 présente celle du $\widetilde{min}(F, G)$.

Figure 4.2 – Maximum de F et G



Le maximum ainsi défini forme la base du Fuzzy Max Order et de l'approche de Kerre. En effet, dans ces approches, les nombres flous sont rangés selon des critères fonctions du maximum. Le minimum est essentiel à la comparaison selon le FMO.

Figure 4.3 – Minimum de F et G



4.1.2 Fuzzy Max Order

Le Fuzzy Max Order (FMO), introduit dans [Ramik et Rimanek, 1985] par Ramik et Raminek, est un ordre partiel considéré comme l'un des plus naturels. Il est défini comme suit :

$$F \preccurlyeq G \Leftrightarrow \forall \alpha \in [0,1] \begin{cases} \sup(F_\alpha) \leqslant \sup(G_\alpha) \\ \inf(F_\alpha) \leqslant \inf(G_\alpha) \end{cases}$$

où F_α et G_α sont les α-coupes de F et G.

Selon cette comparaison, pour qu'un nombre flou (F) soit inférieur à un autre (G), il faut que, pour chaque couple d'α-coupes (F_α, G_α) où $\alpha \in [0,1]$, la borne inférieure de F_α soit inférieure à celle de G_α et que la borne supérieure de F_α soit aussi inférieure à la borne supérieure de G_α. Ce qui revient à assimiler F au $\widetilde{min}(F,G)$ et G au $\widetilde{max}(F,G)$.

D'ailleurs, d'après [Ramik et Rimanek, 1985], quand la relation $\preccurlyeq$ est définie par le FMO, alors les clauses suivantes sont équivalentes :

- $F \preccurlyeq G$
- $F = \widetilde{min}(F,G)$
- $G = \widetilde{max}(F,G)$

La relation $\preccurlyeq$ définie selon le FMO est un pré-ordre entre nombres flous basé sur l'égalité au maximum ou au minimum. Elle se rapproche donc de l'opérateur $\leq$ sur les réels, où si, pour deux réels a et b, $a \leq b$, alors $a = min(a,b)$ et $b = max(a,b)$. Le Fuzzy Max Order apparaît donc comme un ordre intuitif surtout si l'ordre souhaité a pour but de déterminer l'antériorité.

Malheureusement, le *FMO* est un ordre partiel : quand F et G se chevauchent (comme dans la Figure 4.1), les nombres ne peuvent alors pas être comparés. Cela est la conséquence du fait qu'aucun des nombres n'est égal au maximum ni donc au minimum. Il ne peut donc être utilisé pour interroger temporellement le système afin d'obtenir les positions temporelles de tous les objets datés présents dans la base de données du système

vis-à-vis d'une période donnée en entrée. Le traitement nécessaire pour permettre ces interrogations nécessite donc une démarche capable de positionner chaque objet vis-à-vis de la période fournie en entrée. De plus, comme les données sont imparfaites, ce positionnement temporel ne peut être net ou binaire, c'est à dire de type « avant-après » ou encore « antérieur-postérieur », l'aspect binaire ne rendant pas compte de l'ambiguïté de cette décision pour des données imparfaites.

Cependant, comme le *FMO* est un ordre partiel intuitif pour le problème de l'antériorité, le traitement des données devra dire qu'un objet o_1 est antérieur à une période D_{ref} si la période d'activité de o_1 est inférieure à D_{ref} pour le FMO. Il s'agit donc d'étendre le FMO à l'ensemble des représentations de toutes les périodes d'activité y compris celles qui ne seraient pas FMO comparables à la période donnée en entrée.

De plus, afin de pouvoir comparer deux nombres flous quelconques, les méthodes de comparaison deux à deux tendent à regarder les nombres flous par le prisme de valuations des nombres à comparer [Wang et Kerre, 2001a]. D'autres méthodes, à l'instar de celle proposée par Etienne E. Kerre, considèrent l'ensemble des nombres à classer pour calculer des indices pour chaque nombre par rapport à l'ensemble, et comparent les nombres via leurs indices.

4.1.3 Comparaison selon les indices de Kerre

La proposition de Kerre est la suivante : si on prend l'ensemble des nombres flous à classer, et que l'on compare les distances séparant chaque nombre flou du maximum de l'ensemble, alors on peut ranger les nombres flous par ordre de distance décroissante ; le plus petit nombre aura la distance au maximum la plus grande.

Pour un ensemble Ω de n nombres flous $\{A_1, A_2, \ldots, A_n\}$, Kerre propose de ranger ces nombres flous par la comparaison de leur distance de Hamming au maximum de l'ensemble Ω selon le principe d'extension [Kerre, 1982].

La distance de Hamming entre deux nombres flous F et G de fonctions d'appartenance f et g est définie comme suit dans [Wang et Kerre, 2001b] :

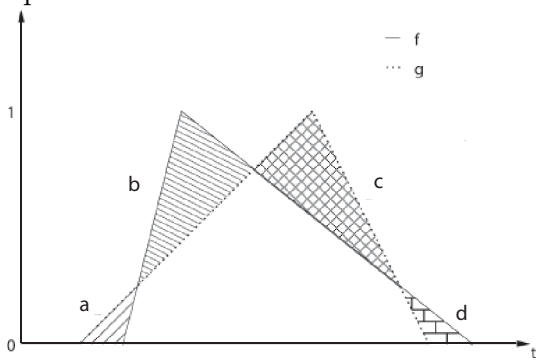
$$D_H(F, G) = \int |f(x) - g(x)| dx.$$

Ceci correspond pour les deux nombres flous présentés dans la Figure 4.1 à la somme des quatre aires (a, b, c et d) illustrées dans la Figure 4.4.

Cette distance est dite de Hamming car c'est une adaptation de la distance, proposée par Hamming dans [Hamming, 1950], entre deux chaînes de caractères A et B de même longueur ; cette distance est le cardinal de l'ensemble des valeurs de A qui diffèrent de celles de B. La distance dite de Hamming entre deux nombres flous peut aussi être vue comme la distance euclidienne L_1 [52] sur l'ensemble des fonctions d'appartenance des nombres flous.

52. La distance L_1 est appelée en anglais *Manhattan norm* ; pour le standard utilisé par le National

Figure 4.4 – Aires pour le calcul de la distance de Hamming entre F et G



La distance de Hamming entre A_i, avec $i \in [1, n]$, et $\widetilde{max}(A_1, A_2, \ldots, A_n)$ est appelée indice de Kerre $K_\Omega(A_i)$. Ainsi, l'indice de Kerre de A_i dans $\{A_1, A_2, \ldots, A_n\}$ est obtenu de la manière suivante :

$$K_\Omega(A_i) = D_H\left(A_i, \widetilde{max}(A_1, A_2, \ldots, A_n)\right) \tag{4.1}$$

Pour Kerre, un nombre flou est plus petit qu'un autre dans un ensemble donné si et seulement si son indice de Kerre est supérieur à celui de l'autre. C'est à dire $A_i \preccurlyeq A_j$ par rapport à $\Omega = \{A_1, A_2, \ldots, A_n\}$, avec $(i, j) \in [1, n]$, si et seulement si $K_\Omega(A_i) \geq K_\Omega(A_j)$.

Dans l'approche de Kerre, même si on ne souhaite comparer que deux nombres de Ω, on doit calculer le maximum de Ω défini par le principe d'extension, c'est-à-dire :

$$\widetilde{max}(\Omega)(z) = \sup_{\substack{x_1, \ldots, x_n \in \mathbb{R}, \\ z = max(x_1, \ldots, x_n)}} \left(min(A_1(x_1), \ldots, A_n(x_n))\right).$$

Ce calcul peut s'avérer coûteux dans le cas d'interrogation sur de volumineuses bases de données. De plus il devra être fait à chaque interrogation temporelle du système, car pour chaque traitement, l'ensemble Ω des éléments à ranger change. En effet la période en entrée est une variable introduite par l'utilisateur. L'algorithme d'une interrogation utilisant l'approche de Kerre nécessite un prétraitement des données avant le traitement réel (voir Algorithme 4.1) .

Une solution à ce problème est de ne considérer, durant le processus, que la réduction de l'ensemble des nombres flous à la paire à comparer :

- D_{ref} : la période pour laquelle on souhaite connaître la position temporelle des éléments de la base,
- $pf.fDate$: la période d'activité de l'objet pf en cours de traitement par le processus d'interrogation.

L'algorithme associé est alors l'algorithme 4.2.

Institut of Standards and Technology voir http ://www.nist.gov/dads/HTML/lmdistance.html

Algorithme 4.1 : Interrogation temporelle de la base de données *BDFRues* par rapport à la période D_{ref} et utilisant Kerre

$\Omega \leftarrow \{pf.fDate | pf \in BDFRues$ et $pf.fDate$: période d'activite de $pf\} \bigcup \{D_{ref}\}$;
Calcul du $\widetilde{max}(\Omega)$;
pour chaque ***objet*** pf ***de la BDFRues*** **faire**
  D_{ref} est antérieur à pf si $K_\Omega(pf.fDate) \leq K_\Omega(D_{ref})$;
fin

Algorithme 4.2 : Interrogation temporelle de la base de données *BDFRues* par rapport à la période D_{ref} et utilisant l'indice de Kerre, avec l'ensemble de référence réduit à la paire en cours de traitement

pour chaque ***objet*** pf ***de la BDFRues*** **faire**
  $\Omega \leftarrow \{pf.fDate, D_{ref}\}$;
  Calcul du $\widetilde{max}(\Omega)$;
  D_{ref} est antérieur à pf si $K_\Omega(pf.fDate) \leq K_\Omega(D_{ref})$;
fin

Bien que cette réduction du domaine de considération a pour conséquence la non transitivité des résultats, on considère la méthode de Kerre comme une extension du *FMO*. Les cas de non transitivité sont conséquences d'incohérence temporelle pour l'antériorité et mettent donc en évidence des inconsistances dans la base de données vis à vis de l'antériorité. Ces inconsistances peuvent attirer l'attention des experts et les aider à remettre en cause telle ou telle datation en fonction des informations de la base de données.

Comme les différentes méthodes de comparaison disponibles dans la littérature donnent dans certains cas des résultats qualitatifs contradictoires, je propose d'évaluer de manière quantitative l'antériorité entre deux nombres flous pour affiner l'aide fournie aux experts. De plus, vu que l'on ne considère les résultats que deux à deux, l'aspect non transitif n'est pas un réel problème dans ce cas. Ce travail propose de construire un indice d'antériorité permettant de relativiser l'aspect binaire et formel d'une décision issue d'une comparaison classique, et se basant sur l'approche de Kerre vue comme extension du FMO.

4.2 Indice d'antériorité

L'idée clé de l'approche de Kerre est : plus l'indice de Kerre d'un nombre flou est grand (par rapport à l'ensemble des entités à ranger), plus le nombre est petit. Pour le FMO, si un élément est identique au maximum, alors il est le plus grand. Ces deux approches sont donc basées sur des principes proches, voire se recoupent quand la cardinalité de l'ensemble de nombres flous à comparer est de 2.

L'indice d'antériorité, élément clé du traitement temporel des données flous établi dans ce travail, est défini à partir des valeurs de l'indice de Kerre pour une paire de nombres

flous (F,G) quelconques. Cet indice est relatif à la comparaison sur laquelle il porte.

4.2.1 Définition

Comme premier principe, si par *FMO* on a $F \preccurlyeq G$ alors la valeur de l'indice d'antériorité de F par rapport à G est égal à 1. Pour tous les autres cas, l'addition de la valeur de l'indice d'antériorité de F au regard de G et de celle de G par rapport à F est égale à 1. L'indice d'antériorité $Ant(F, G)$ de F vis-à-vis de G est défini comme suit :

$$Ant(F,G) = \begin{cases} \frac{K_\Omega(F)}{K_\Omega(F)+K_\Omega(G)} & \text{si } K_\Omega(F) + K_\Omega(G) > 0, \\ 1 & \text{si } K_\Omega(F) + K_\Omega(G) = 0; \end{cases}$$

avec $\Omega = \{F, G\}$.

Comme les indices de Kerre ne peuvent pas être négatifs par définition, le cas $K_\Omega(F) + K_\Omega(G) < 0$ ne peut pas exister. Pour F et G, et $\Omega = \{F, G\}$, dont les fonctions d'appartenance sont présentées dans la Figure 4.1 et dont les aires les séparant du maximum sont illustrées dans la Figure 4.4 par les zones (a, b, c et d), $K_\Omega(F)$ est égal à $b + c$ tandis que $K_\Omega(G)$ est égal à $a + d$.

Ainsi, comme $K_\Omega(F) + K_\Omega(G) > 0$ alors sur l'exemple des figures 4.1 et 4.4 :

$$Ant(F,G) = \frac{b+c}{a+b+c+d} = 0.86 \text{ et } Ant(G,F) = \frac{a+d}{a+b+c+d} = 0.14.$$

$Ant(F, G)$ est un indice d'éloignement entre F et $\widetilde{max}(F, G)$ utilisant les distances de Hamming de F et G au $\widetilde{max}(F, G)$. Plus la valeur de l'indice est grande, plus F est éloigné du maximum, et donc, plus F est petit par rapport à G. Ainsi, plus la valeur de $Ant(F, G)$ est grande, plus G est grand par rapport à F. Donc, $Ant(F, G)$ peut être considéré comme un indice de proximité entre G et $\widetilde{max}(F, G)$ utilisant les distances de Hamming de F et G au $\widetilde{max}(F, G)$.

Soient deux périodes P_F et P_G représentées par les nombres flous F et G, en utilisant l'indice d'antériorité pour connaitre l'antériorité de P_F vis à vis de P_G, on obtient :

- $Ant(F, G) = 0 \Rightarrow$ « P_F n'est pas antérieur à P_G »,
- $0 < Ant(F, G) \leq 0.5 \Rightarrow$ « P_F n'est pas suffisamment antérieur à P_G »,
- $0.5 \leq Ant(F, G) < 1 \Rightarrow$ « P_F est suffisamment antérieur à P_G »,
- $Ant(F, G) = 1 \Rightarrow$ « P_F est antérieur à P_G ».

Ainsi, si les périodes P_F et P_G sont représentées par les nombres flous F et G définis dans la figure 4.1, alors P_F est suffisamment antérieur à P_G car $Ant(F, G) = 0.86$ et P_G n'est pas suffisamment antérieur à P_F car $Ant(G, F) = 0.14$.

L'interrogation temporelle de *BDFRues* utilisant les valeurs de l'indice d'antériorité est exposée dans l'algorithme 4.3.

Un indice d'antériorité est donc indispensable aux processus de traitement des données imparfaites pour l'interrogation temporelle et donc à l'analyse temporelle de l'information.

Algorithme 4.3 : Interrogation selon l'antériorité de D_{ref} aux objets de la base de données *BDFRues*

pour chaque ***objet*** pf ***de la BDFRues où*** $pf.fDate$ ***est défini*** **faire**
- $\Omega \leftarrow \{pf.fDate, D_{ref}\}$;
- Calcul du $\widetilde{max}(\Omega)$;
- Calcul de $K_\Omega(pf.fDate)$ et de $K_\Omega(D_{ref})$;
- Calcul de $Ant(D_{ref}, pf.fDate)$;
- **si** $Ant(D_{ref}, pf.fDate) = 1$ **alors**
  - D_{ref} est antérieur à pf ;
- **fin**
- **si** $0.5 \leq Ant(D_{ref}, pf.fDate) < 1$ **alors**
  - D_{ref} est suffisamment antérieur à pf ;
- **fin**
- **si** $0 < Ant(D_{ref}, pf.fDate) < 0.5$ **alors**
  - D_{ref} n'est pas suffisamment antérieur à pf ;
- **fin**
- **si** $Ant(D_{ref}, pf.fDate) = 0$ **alors**
  - D_{ref} n'est pas antérieur à pf ;
- **fin**

fin

L'idée de l'utilisation de cet indice est d'obtenir une information quantitative descriptive des positions temporelles des objets de la base relativement à une période en entrée.

4.2.2 Propriétés et remarques

Si deux nombres flous F et G sont *FMO* comparables, alors on a :
- $Ant(G, F) = 1$ si $G \preccurlyeq F$,
- $Ant(F, G) = 1$ si $F \preccurlyeq G$.

De même, deux nombres flous sont égaux (voir 2.3.2) si et seulement si $Ant(F, G) = Ant(G, F) = 1$ (car on a selon le FMO : $G \preccurlyeq F$ et $F \preccurlyeq G$).

De plus, $Ant(F, G) = 0.5$ si et seulement si $Ant(G, F) = 0.5$. Dans ce cas, on peut remarquer que les nombres flous sont aussi proches de leur minimum que de leur maximum, et donc la décision « F est antérieur à G » est aussi légitime que la décision « G est antérieur à F ».

Une autre remarque importante porte sur le choix entre le minimum et le maximum pour la distance. De part sa construction, $Ant(F, G)$ est, en utilisant la distance de Hamming, un indice d'éloignement entre G et $\widetilde{max}(F, G)$ mais aussi entre F et $\widetilde{min}(F, G)$. En effet, calculer la distance de Hamming de F au minimum équivaut à calculer celle séparant G du maximum car on a :

$$\forall x \in Support(F) \cup Support(G), (f(x) = \widetilde{min}(f, g)(x)) \Leftrightarrow (g(x) = \widetilde{max}(f, g)(x)))$$

et réciproquement

$$\forall x \in Support(F) \cup Support(G), (g(x) = \widetilde{min}(f,g)(x)) \Leftrightarrow f(x) = \widetilde{max}(f,g)(x)).$$

On peut donc en déduire que $\int |f(x) - \widetilde{min}(f,g)|dx$ est égale à $\int |g(x) - \widetilde{max}(f,g)|dx$. Ainsi, les distances de Hamming des deux nombres flous à comparer à leur maximum suffit pour évaluer l'antériorité entre deux éléments. Cela implique aussi que l'indice d'antériorité de F à G peut être considéré comme un indice de postériorité de G à F.

Dans la partie suivante, l'indice d'antériorité est utilisé dans deux applications sur des données archéologiques. La première illustre l'intérêt de l'utilisation des valeurs de l'indice d'antériorité pour la résolution de conflit, la deuxième propose une méthode d'interrogation temporelle avec visualisation des résultats qui prend une période D_{ref} en entrée et qui associe à chaque objet de *BDFRues* une couleur en fonction de la valeur de l'antériorité de D_{ref} à la période d'activité de l'objet en cours de traitement.

4.3 Application aux données archéologiques

Dans les deux exemples suivants, l'indice d'antériorité est utilisé dans un contexte archéologique. Grâce à cet indice, on obtient une information graduée de l'antériorité des objets archéologiques. Le premier exemple prend en compte la valeur de l'indice pour la résolution de conflits architecturaux dans le cadre de la production de cartes simulées en fonction du temps.

Le second exemple présente, dans un SIG, une méthode d'interrogation visuelle et graduelle de l'antériorité entre une période en entrée et les données de *BDFRues*. Le traitement associe alors à chaque objet une couleur en fonction de la valeur de l'indice d'antériorité de la période en entrée vis-à-vis des tronçons de rues romaines.

4.3.1 Cas d'un conflit datation/architecture

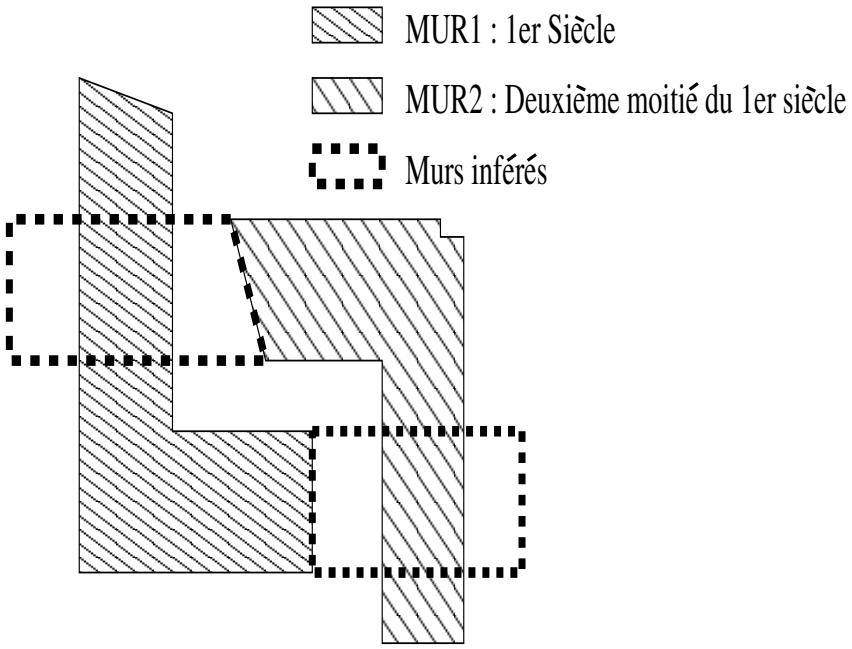
Les périodes d'activité des objets issus de fouilles archéologiques peuvent être représentées par des nombres flous (voir partie I). Afin de délivrer des cartes prédictives aux archéologues, il est nécessaire de faire de l'inférence spatiale en fonction du temps. Comme celle-ci a besoin de positionner temporellement les données, je propose d'utiliser l'indice d'antériorité pour les problèmes dus à la datation. Dans ce cadre, l'indice d'antériorité aide à la résolution de contradictions architecturales et spatiales dans les cartes issues des fouilles.

Prenons le cas de la découverte de deux murs (MUR1 et MUR2) datant respectivement du premier siècle après J.C. (*i.e. de 0 à l'an 100*) et de la seconde moitié du premier siècle après J.C. (*i.e. de l'an 50 à l'an 100*). Si on essaye de les restituer dans leur totalité sous des hypothèses d'extension architecturale présentées dans la figure 4.5, cette démarche serait source d'un conflit.

En représentation ensembliste classique, la période du mur MUR1 serait représentée par l'intervalle de temps $[0, 100]$, la représentation temporelle de MUR2 serait l'intervalle de temps $[50, 100]$.

Le but est de construire tout de même une carte pour la simulation temporelle basée sur les hypothèses des experts sur la datation. Ainsi, l'indice d'antériorité fournit une réponse pondérée à la question « Est-ce que le mur MUR1 fut construit avant le mur MUR2 ? ». Mais il faut tenir compte du caractère *vague* des périodes d'activité. Les deux

Figure 4.5 – Présentation d'un conflit spatiotemporel



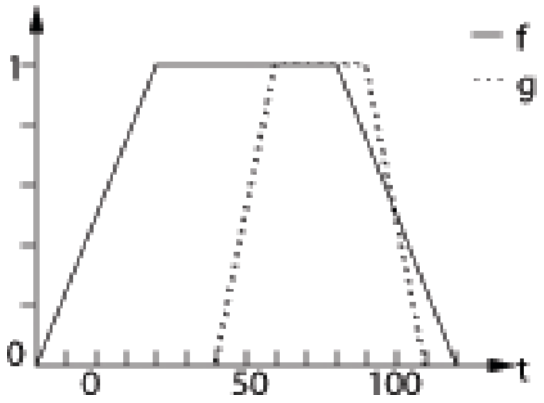
périodes d'activité sont modélisées par les nombres flous de forme trapézoïdale F (période du MUR1) et G (période du MUR2). Ces nombres flous sont déterminés en suivant la procédure qui suit. Soit $t1$ et $t2$ les valeurs de début et de fin de la période d'activité d'un objet o et $o.fDate$ la représentation temporelle floue associée à o, alors la fonction d'appartenance $o.fdate$ de $o.fDate$ est définie comme suit :

- t1 et t2 sont respectivement le début et la fin des périodes d'activité,
- $Support(o.fDate) = [(t1 - 0.2(t2 - t1)), (t2 + 0.2(t2 - t1))]$,
- $Noyau(o.fDate) = [(t1 + 0.2(t2 - t1)), (t2 - 0.2(t2 - t1))]$,

Pour le MUR1, t1 est égal à 0, et t2 à 100, tandis que pour le MUR2, t1 est égal à 50, et t2 à 100. Le coefficient 0.2 indique que l'imprécision des bornes des périodes d'activité est relativement faible : elle correspond à 10% des périodes d'activité. Donc, en utilisant

une notation classique d'un nombre flou en trapèze (sous la forme d'un quadruplet [53]), on a : $F = (-20, 20, 80, 120)$ et $G = (40, 60, 90, 110)$. Les fonctions d'appartenance des ensembles flous associés aux MUR1 et MUR2 sont illustrées par la Figure 4.6.

Figure 4.6 – Nombres flous représentant les périodes d'activité des murs



Le calcul des valeurs de l'indice d'antériorité entre F et G donne $Ant(F, G) = 0,95$ et $Ant(G, F) = 0,05$. Ainsi, on peut suggérer que le MUR1 est antérieur au MUR2. Ce qui implique bien sur, qu'il a été construit avant. L'indice d'antériorité permet donc une interprétation révisable de l'antériorité entre éléments laissant supposer que MUR1 a été détruit avant la construction de MUR2. L'enchaînement de ce type d'interprétation basé sur l'indice d'antériorité donne à l'expert un nouvel outil pour l'analyse des données archéologiques. L'application qui suit montre une autre utilisation de l'indice d'antériorité.

4.3.2 Cas des rues romaines

L'interrogation temporelle présentée dans cette application prend en entrée une période représentée par un nombre flou D_{ref} et assigne une couleur à chaque objet de la base de données *BDFRues* en fonction de la valeur issue du calcul d'indice d'antériorité entre D_{ref} et la période d'activité de l'objet.

Ainsi, pour chaque objet pf de *BDFRues*, on calcule, si cela est possible (i.e. si l'objet a une composante temporelle non nulle dans la base), l'indice d'antériorité de D_{ref} à la période d'activité floue $pf.fDate$ de pf : $Ant(D_{ref}, pf.fDate)$. Afin de visualiser ces valeurs de l'indice, on assigne à chaque objet la couleur correspondant à la classe de valeurs (*cf.* légende de la Figure 4.7). L'algorithme 4.4 correspond à cette interrogation sur l'ensemble des objets de la base.

53. Cette notation associe un quadruplet (a, b, c, d) à la forme trapèzoïdale de la fonction d'appartenance d'un nombre flou F. Ce quadruplet est défini selon les bornes inférieures a et supérieures d du support de F et selon les bornes inférieures b et supérieures c du noyau de F ; une autre notation dite de type $L - R$ associe un quadruplet $(m, n, a, b)_{LR}$ tel que $Support(F) = [m - a, n + b]$ et $Ker(F) = [m, n]$.

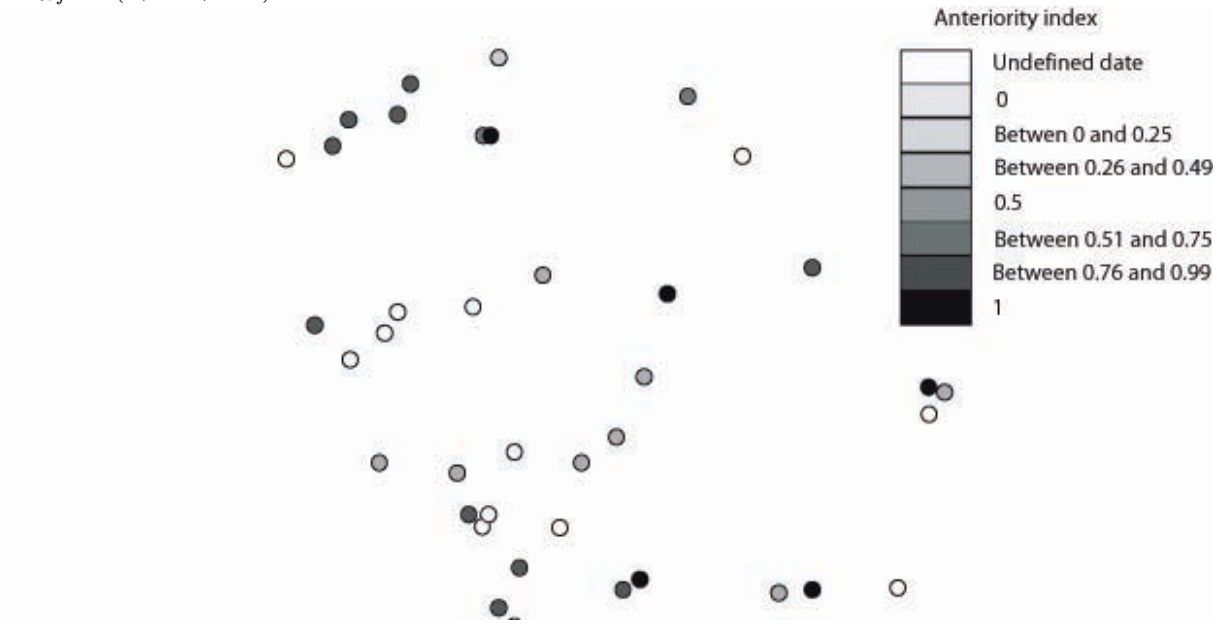
Algorithme 4.4 : Interrogation visuelle de la base de données *BDFRues* selon l'antériorité à D_{ref}

pour chaque ***objet*** pf ***de la BDFRues avec*** $pf.fDate$ ***défini*** **faire**
  Calcul de $Ant(D_{ref}, pf.fDate)$;
  Affectation d'une *couleur* à pf en fonction de la valeur de $Ant(D_{ref}, pf.fDate)$;
fin

Dans la Figure 4.7, les couleurs correspondent aux valeurs de l'indice d'antériorité pour D_{ref} qui est alors représentée par un nombre flou triangulaire noté par le triplet [54] $(0, 200, 400)$.

L'interrogation posée est donc "À quels objets la période D_{ref} est-elle antérieure ?" ; le système apporte une réponse quantitative à cette question grâce au calcul des valeurs de l'indice d'antériorité. Les objets en blanc sont ceux n'ayant pas de période d'activité définie dans *BDFRues*.

Figure 4.7 – Évaluation de l'antériorité de D_{ref} aux objets présents dans la base, avec $D_{Ref} = (0, 200, 400)$



L'interrogation temporelle de *BDFRues* par rapport à D_{ref} dans un SIG permet donc de visualiser par un code couleur le positionnement temporel de chaque objet relativement

54. Cette notation associe un triplet (a, b, c) à la forme triangulaire de la fonction d'appartenance d'un nombre flou F. Ce triplet est défini selon les bornes inférieures a et supérieures c du support de F et selon l'unique valeur b du noyau de F. Cette notation fait écho à celle des nombres flous trapézoïdaux dans laquelle un nombre triangulaire (a, b, c) serait (a, b, b, c). En notation $L-R$, on défini le triplet $(m, a, b)_{LR}$ tel que $Support(F) = [m - a, m + b]$ et $Ker(F) = m$.

à une date ou période de référence D_{ref}. Le positionnement de chaque objet se fait selon l'antériorité de D_{ref} à ce dernier.

Dans leurs prospections, les experts évaluent les objets qu'ils sont susceptibles de trouver dans un site tant d'un point de vue fonctionnel que temporel. Les résultats de ce type d'analyse peuvent donc apporter une information complémentaire en vue de leurs diagnostics. Ils peuvent, de même, utiliser ce traitement dans le but de déterminer l'évolution de la ville pendant la période romaine. C'est donc une information clé pour les archéologues qu'il faudra recouper avec d'autres pour une interprétation plus globale de la chronologie d'un site de fouille ou d'un ensemble d'objets archéologiques. Si ces informations sont connues des experts, elles deviennent plus accessibles et plus faciles à mettre en œuvre pour tester des hypothèses et bâtir des modèles.

◇◇◇◇◇

Dans ce chapitre, afin d'interroger un SIG archéologique temporellement, une évaluation de l'antériorité entre deux objets archéologiques est proposée. Pour établir cette évaluation, dans le cadre de la modélisation de l'information temporelle par des ensembles flous, on construit un indice, appelé indice d'antériorité, basé sur celui de Kerre et sur le FMO. Cet indice est ensuite utilisé pour le traitement des données archéologiques nécessaire à l'interrogation temporelle du SIG. L'indice d'antériorité apporte plus de souplesse qu'une comparaison classique (binaire ou qualitative) entre nombres flous.

Le traitement des données proposé dans ce chapitre a fait l'objet d'une communication internationale [de Runz *et al.*, 2006] et d'un article soumis à une revue internationale [de Runz *et al.*, 2008c]. De plus, la souplesse de la valuation me permettra de construire le graphe d'antériorité tel celui proposé dans le chapitre 7.

Ainsi, établir l'antériorité des objets archéologiques à une période de référence a son intérêt propre, l'analyse qui en découle est temporelle. Les données applicatives sont dans ce mémoire spatiotemporelles. Il apparaît donc intéressant de développer des méthodes d'analyse spatiotemporelle des données.

Dans le cadre des données archéologiques, l'un des buts de l'exploitation des SIG par les archéologues est d'arriver à reconstituer le passé. Cet objectif nécessite la généralisation de l'information associée à un territoire urbain. Pour cela une analyse spatiotemporelle est nécessaire. Cependant, comme l'information archéologique est lacunaire, *i.e.* les objets sont des fragments d'objets plus importants, celle-ci s'avère complexe. Pour cela, il est intéressant d'utiliser durant l'analyse les formes géométriques estimées des objets à reconstruire.

Dans le chapitre 5, je propose une méthode pour l'interrogation spatiotemporelle de

la base de données ayant pour but l'aide à la restitution du passé, qui pour une période et une forme données, estime les présences potentielles des objets pour la dite période.

Chapitre 5

Analyse spatiotemporelle : interrogation sous critère de forme

Sommaire

Comme il est indiqué dans le chapitre 1, les données archéologiques peuvent être lacunaires. La question qui se pose alors est de savoir si palier ce défaut d'information est possible, même partiellement, et si oui, comment. Il se trouve que dans un certain nombre de cas, les fouilles qui ont été réalisées ont permis de mettre à jour des objets qui sont des fragments d'objets plus importants. Or, si ces derniers ont la particularité d'avoir une structure géométrique particulière, on peut essayer de s'appuyer sur cette connaissance afin de construire des propositions d'hypothèses de présence.

Pour mettre en évidence cette fonctionnalité possible au sein d'un SIG, je propose de définir une méthode d'interrogation permettant de prendre en compte les structures spatiales et de les visualiser à partir de la recherche d'une forme particulière. Ces informations géométriques sont alors assimilées à une entrée du processus d'interrogation spatiotemporelle de base de données dans un SIG. Le traitement est donc paramétré par la forme

des objets qui ne sont que partiellement stockés dans la base de données du système.

Dans le cadre du projet *SIGRem*, l'origine du besoin et le cadre de validation de cette démarche a naturellement été le traitement de la présence des rues gallo-romaines au cours des différentes époques. Dans le contexte des rues de Reims à l'époque romaine, seuls des tronçons de rues sont stockés dans *BDFRues*. La structure que l'on souhaite dégager par la procédure d'interrogation spatiotemporelle est donc le plan des rues de Reims en fonction du temps. De plus, les rues de Reims à l'époque étant généralement linéaires, les formes à retrouver sont donc des droites, voire des segments.

La forme géométrique devant guider le traitement pour la complétion, il semble naturel d'utiliser une méthode de reconnaissance de formes. Comme la forme à retrouver est régulièrement simple géométriquement et que la transformée de Hough (TH), introduite dans [Hough, 1962], est un outil puissant pour la détection de formes simples (voir à ce propos [Duda et Hart, 1972]), je propose de l'appliquer au traitement des données archéologiques. Cependant, ces dernières sont imparfaites. Il existe des adaptations de la transformée de Hough dans le contexte d'informations imparfaites à l'instar de la transformée aléatoire de Hough [Xu et Oja, 1993] ou encore de la transformée floue de Hough [Han *et al.*, 1994]. La transformée de Hough et ses dérivées sont basées sur la projection des données dans un espace paramétrique spécifique aux formes recherchées. Celle-ci est réalisée sous la forme de votes des données dans une matrice discrétisant cet espace : l'accumulateur de Hough. La sélection des valeurs des paramètres associés aux cellules de l'accumulateur ayant recueilli le plus grand nombre de votes permet de proposer les formes recherchées et ainsi de compléter partiellement la lacune de l'information portée par les objets issus de fouilles archéologiques.

Comme les données de *BDFRues* sont modélisées à l'aide de la théorie des ensembles flous, je propose que le traitement s'appuie sur la transformée floue de Hough introduite dans [Han *et al.*, 1993] et [Han *et al.*, 1994]. Une seconde conséquence de la qualité des données est que, si le traitement proposé dans ce chapitre ne permet pas de visualiser le véritable plan, il donne cependant des estimations de la présence des rues romaines en fonction des époques. L'objectif de cette analyse n'est pas de fournir une interprétation mais d'aider les experts (archéologues) à construire leurs propres modèles d'interprétation car stocker des informations dans un SIG n'a que peu d'intérêt si ces informations ne sont pas exploitables et diffusées auprès de la communauté des archéologues et du public intéressé.

Ce traitement s'appuie donc sur la construction d'un accumulateur fonction de la forme pour chaque caractéristique à considérer. Pour les objets de *BDFRues*, les caractéristiques importantes sont les localisations, les orientations et les périodes d'activité. Le traitement portera ainsi sur la construction des accumulateurs correspondant à chacune des caractéristiques. Dans le but de pouvoir sélectionner les rues, les accumulateurs sont agrégés ; on fusionne donc les localisations, les orientations et les périodes d'activité par le prisme de ces accumulateurs. La sélection se fait sur le résultat de l'agrégation. On

visualise ensuite les résultats de la sélection dans le SIG sous formes de pré-cartes ou cartes incomplètes. Dans cette démarche, il est nécessaire d'effectuer une agrégation pour laquelle j'ai fait le choix d'utiliser une moyenne pondérée. Le choix des coefficients affecte nécessairement les résultats du traitement.

Dans ce chapitre, après avoir présenté la transformée de Hough, j'exposerai le traitement effectué pour chaque objet de la base en fonction du temps et de la forme en entrée. En appliquant le traitement aux objets archéologiques stockés dans *BDFRues*, j'étudierai l'influence des coefficients dans l'agrégation des accumulateurs, pour finir par la comparaison des résultats du traitement, utilisant une pondération déterminée empiriquement, avec des plans issus d'expertises.

5.1 Reconnaissance de formes simples dans les images : transformée de Hough

L'interrogation spatiotemporelle décrite dans ce chapitre se base sur une méthode particulière de reconnaissance de formes en traitement d'images : la transformée floue de Hough. Celle-ci est dérivée de la transformée de Hough.

5.1.1 Contexte général

Le champ disciplinaire de la reconnaissance de formes a connu des avancées très significatives dans les années 1960, comme l'indique [Webb, 2002]. Généralement, comme cela est précisé dans [Hjelmas et Low, 2001] et [Bow, 2002], les méthodes de reconnaissance de formes géométriques ou naturelles s'appuient sur l'information portée par les contours des objets. [Loncaric, 1998] présente une revue des méthodes d'analyse de formes dans les images.

Les images sont pré-traitées par des méthodes de détection de contours afin d'en dégager les pixels [55] appartenant aux contours des objets présents dans l'image. Le filtre du gradient selon les dérivées premières [Torre et Poggio, 1984], le filtre de Roberts (1965), le filtre de Canny [Canny, 1986], le filtre de Deriche [Deriche, 1987], le filtre de Prewitt (1970) et le filtre de Sobel (décrit dans [Duda et Hart, 1973]) sont des méthodes classiques de détection de contours [56].

Les méthodes de reconnaissance de formes, s'appuyant sur une extraction des contours, fonctionnent sur des images binaires (noir et blanc) ou en niveaux de gris et ne travaillent que sur les points d'intérêt (points qui appartiennent aux contours). Il existe de nombreuses méthodes de reconnaissance de formes. Pour les formes simples, une des méthodes les plus performantes est la transformée de Hough [Hough, 1962]. C'est une méthode d'estimation paramétrique de formes géométriques.

55. Pixel ou *picture element*.
56. Les filtres de Roberts, de Prewitt et Sobel sont illustrés en Annexe B.

5.1.2 Transformée de Hough

La transformée de Hough, introduite dans [Hough, 1962] fait partie des techniques classiques de détection de forme en traitement d'images. Bien qu'initialement introduite pour les droites, elle fut généralisée aux formes géométriques [Duda et Hart, 1972]. Elle fut étendue aux formes naturelles de manière non supervisée (voir [Bonnet, 2002]). Des études sur la transformée de Hough et ses dérivées ont été menées par Illingworth et Kittler [57] en 1988 ainsi que par Leavers [58] en 1993.

Le principe de la méthode est de projeter les points d'intérêt d'une image dans un espace paramétrique correspondant à l'équation des formes à reconnaître. On construit un accumulateur défini sur une discrétisation de l'espace paramétrique. La valeur d'une cellule de l'accumulateur correspond au crédit accordé à la présence, dans l'image, de la forme associée à cette cellule. Afin de définir ce crédit dans l'accumulateur, chaque point d'intérêt de l'image vote dans chaque cellule correspondant aux instances de la forme recherchée auxquelles le point peut appartenir. Par ce vote, la valeur des cellules est incrémentée. Plus le nombre de points votant dans une même cellule de l'espace des paramètres est important, plus la valeur de cette cellule est grande et plus le crédit accordé à la forme associée à la cellule est important.

Par exemple, prenons le cas où la recherche porte sur la reconnaissance de droites dans une image et considèrons celles-ci selon leurs équations polaires :

$$\mathcal{D} = \{(i,j) | \rho = i \times \cos(\theta) + j \times \sin(\theta)\},$$

où le couple (i,j) représente les coordonnées d'un point appartenant à la droite $\mathcal{D}$. L'espace des paramètres est alors à deux dimensions et chaque droite possible y est représentée par ses coordonnées (ρ,θ) (voir Figure 5.1).

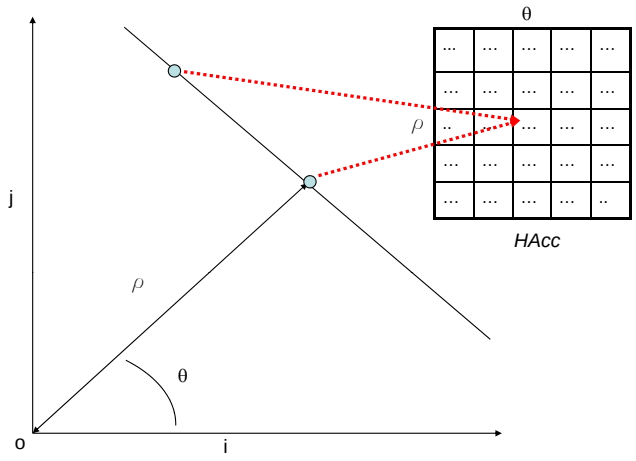
Un point d'intérêt appartient à un ensemble E de droites possiblement présentes dans l'image. Ce point vote dans l'accumulateur en (ρ,θ) dans toutes les cellules associées aux droites de E, c'est à dire que la valeur de chaque cellule associée à un (ρ,θ) correspondant à l'équation d'une droite de E est incrémentée (cf. Algorithme 5.1).

Algorithme 5.1 : Construction de l'accumulateur $HAcc$ dans la transformée de Hough pour la reconnaissance de lignes

```
pour chaque pixel (i,j) appartenant aux contours faire
    pour chaque angle θ faire
        ρ ← i × cos(θ) + j × sin(θ) ;
        HAcc(ρ,θ) ← HAcc(ρ,θ) + 1 ;
    fin
fin
```

57. Voir [Illingworth et Kittler, 1988]
58. Cf. [Leavers, 1993]

Figure 5.1 – Principe du vote dans l'accumulateur $HAcc$ selon la transformée de Hough



Ainsi, le crédit de toutes les droites de cet ensemble augmente du fait de la présence du point dans l'image. Donc, plus le nombre de points d'intérêt appartenant à une droite est important, plus la valeur de la cellule correspondant à cette droite sera grande. Dans l'espace des paramètres, les cellules de valeurs maximales correspondent à des droites potentiellement détectées car le crédit est maximal.

Cette méthode est cependant sensible au bruit dès qu'il s'agit de détecter plusieurs droites. Pour y remédier, différentes approches ont été proposées dont la transformée aléatoire de Hough et la transformée floue de Hough. Comme les données de *BDFRues* sont formalisées selon une représentation floue, j'ai fait le choix de la seconde approche.

5.1.3 Transformée floue de Hough

Pour faire face aux imperfections des images, Joon H. Han *et al* ont introduit, dans [Han *et al.*, 1993] et [Han *et al.*, 1994], la transformée floue de Hough (TFH). Cette méthode prend en considération à la fois le point d'intérêt et son voisinage. Un ensemble flou est utilisé pour définir une matrice W sur le voisinage du point étudié ; W est de dimension $m \times n$. La valeur du vote dans l'accumulateur est celle de la cellule de la matrice qui doit voter.

La figure 5.2 illustre l'intérêt d'une telle approche pour la reconnaissance de ligne. L'algorithme 5.2 présente alors le traitement fait pour la construction de l'accumulateur lors de la recherche de droites à l'aide de la transformée floue de Hough.

Le traitement proposé dans ce chapitre adapte la transformée floue de Hough pour

Figure 5.2 – Exemple de reconnaissance d'une droite à l'aide de la TH et de la TFH

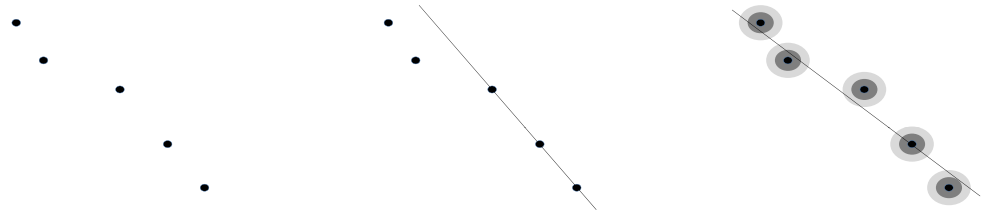


Image contenant les points d'intérêt | Droite reconnue par la TH | Droite reconnue par la TFH

Algorithme 5.2 : Construction de l'accumulateur $FHAcc$ dans la transformée floue de Hough (TFH) pour la reconnaissance de droites selon la matrice de voisinage W

pour chaque ***pixel(i,j) appartenant aux contours*** **faire**
  pour chaque $x \in [-m/2, m/2]$ **faire**
    pour chaque $y \in [-n/2, n/2]$ **faire**
      pour chaque ***angle*** θ **faire**
        $\rho \leftarrow (i + x) * \cos(\theta) + (j + y) * \sin(\theta)$;
        $FHAcc(\rho, \theta) \leftarrow FHAcc(\rho, \theta) + W(x, y)$;
      fin
    fin
  fin
fin

l'interrogation spatiotemporelle sous critères de formes.

5.2 Traitement des données

L'interrogation proposée se structure en trois étapes : construction des accumulateurs, agrégation des accumulateurs, sélection des cellules.

La première étape porte sur l'application de la transformée de Hough aux différentes caractéristiques étudiées des données. Cette application implique la construction d'un accumulateur par caractéristique prise en compte dans le traitement.

Les accumulateurs construits, il faut les agréger afin de pouvoir dégager les informations importantes. Cette fusion des informations accumulées constitue le corps de la seconde étape. Elle est obtenue par la normalisation des accumulateurs afin d'en faire des ensembles flous puis par l'utilisation d'une fonction d'agrégation d'ensembles flous.

L'agrégation réalisée, l'interrogation nécessite une phase de sélection des éléments les plus importants. Cette troisième étape constitue l'étape finale du traitement.

5.2.1 Construction des accumulateurs de Hough

La fonction des objets (ex : murs, rues) sur lesquels portent les traitements définit la forme géométrique à détecter. Si l'interrogation est effectuée sur des données représentant un amphithéâtre romain, la forme à reconnaître sera une ellipse ou un cercle. Pour les données de *BDFRues*, la forme est la droite car les objets représentent des tronçons de rues datant de l'époque romaine et les rues romaines ont à Reims la caractéristique d'être linéaires.

En traitement des images, un point d'intérêt (point de contour) d'une image peut être vu comme une donnée 2D. Dans le cadre des SIG, une base de données géographiques ne contient souvent que les objets importants avec leurs coordonnées (2D). On peut donc considérer que les points d'intérêts seront, dans ce cadre, les objets issus de la BDG. L'utilisation des méthodes de reconnaissance de formes devient ainsi pertinente dans les SIG.

Les objets issus de la BDG ont différentes caractéristiques imparfaites. Dans le contexte de l'information archéologique dont je dispose, ces caractéristiques sont représentées par des ensembles flous. L'idée principale est alors de construire pour chaque caractéristique un accumulateur dans l'espace paramétrique dépendant de la forme géométrique à détecter. Ces accumulateurs sont donc construits sur le même espace.

Dans le cadre des données de *BDFRues*, les objets pf ont trois principales caractéristiques — la localisation, l'orientation et la période d'activité — représentées par des ensembles flous. L'objectif est de construire, pour chaque caractéristique, un accumulateur dans l'espace paramétrique correspondant à la forme recherchée. Or ces objets représentent des tronçons de rues linéaires. Ainsi, dans le cadre de *BDFRues*, je propose la construction de trois accumulateurs — $TFHLocAcc$, $TFHOrienAcc$ et $TFHDateAcc$ — dans l'espace des paramètres (ρ, θ) car la forme géométrique étudiée est la droite.

Le premier ($TFHLocAcc$) est dédié à l'information sur la localisation, le second ($TFHOrienAcc$) porte sur l'orientation tandis que le dernier ($TFHDateAcc$) a pour objet la correspondance de l'objet avec la période de référence fournie en entrée de l'interrogation. Seule la valeur incrémentée change en fonction de l'accumulateur (Algorithme 5.3). La valeur du vote dans $TFHLocAcc$ est $pf.floc(x, y)$, celle dans $TFHOrienAcc$ est $pf.forien(\theta)$. Enfin, pour obtenir les rues pour une période D_{ref} donnée, la cellule de $TFHDateAcc$ correspondant à la droite en cours d'étude est incrémentée par une valeur déterminant la capacité de l'objet pf à être contemporain à la période D_{ref}. Cet indice peut correspondre à la hauteur de l'ensemble flou représentant le ET logique (ET [59]) entre D_{ref} et la période d'activité $pf.fDate$ de l'objet pf, c'est à dire : $Hauteur(D_{ref} \wedge pf.fDate)$).

Ainsi, en fonction des caractéristiques étudiées, plusieurs accumulateurs peuvent être construits. Afin de pouvoir exploiter les données des accumulateurs dans le SIG, une

59. La $t-norme$ de [Zadeh, 1965] fait correspondre au ET logique entre deux ensembles flous le minimum des fonctions d'appartenance des deux ensembles flous (voir 2.3.3).

Algorithme 5.3 : Construction des accumulateurs pour les périodes d'activité, les localisations et les orientations des objets de *BDFRues* selon la détection de droite

pour chaque ***objet*** pf ***de BDFRues*** **faire**
  pour chaque (x, y) ***tel que*** $(x, y) \in Support(pf.fLoc)$ **faire**
    pour chaque θ ***tel que*** $\theta \in Support(pf.fOrien)$ **faire**
      $\rho \leftarrow x \times \cos(\theta) + y \times \sin(\theta)$;
      $TFHLocAcc(\rho, \theta) \leftarrow TFHLocAcc(\rho, \theta) + pf.floc(x, y)$;
      $TFHOrienAcc(\rho, \theta) \leftarrow TFHOrienAcc(\rho, \theta) + pf.forien(\theta)$;
      $TFHDateAcc(\rho, \theta) \leftarrow TFHDateAcc(\rho, \theta) + Hauteur(pf.fDate \wedge D_{ref})$;
    fin
  fin
fin

fusion des informations spatiales et temporelles est nécessaire. Je propose donc d'agréger les accumulateurs. L'étape suivante est dédiée au choix de la fonction d'agrégation des accumulateurs.

5.2.2 Agrégation des accumulateurs

On peut constater que les opérateurs d'agrégation classiques portent sur la fusion de données homogènes. Or, normalisés par leur maximum, les accumulateurs peuvent être vus comme des ensembles flous ayant un domaine de définition commun (l'espace paramétrique) et donc des entités homogènes. Ainsi normalisés, les accumulateurs peuvent être agrégés par les opérateurs classiques (*t-norme*, moyenne, OWA, etc.) de la théorie des ensembles flous.

Didier Dubois et Henri Prade remarquent que le choix du mode de fusion dépend de la nature des données à fusionner et de leur cadre de représentations [Dubois et Prade, 2004]. Dans sa thèse, Marçin Detyniecki propose une étude des opérateurs traditionnels d'agrégation et de fusion [Detyniecki, 2000]. La *t-norme* et la *t-conorme* dans le contexte des ensembles flous (pour [Zadeh, 1965] le Min et le Max, voir 2.3.3) ne considèrent pas que l'ordre des données soit important (fonctions symétriques). De plus, si l'on regarde les fonctions de fusion, définies selon le Min et le Max des valeurs des fonctions d'appartenance, elles admettent un élément neutre [60] et un élément absorbant [61]. Ceci implique que certaines valeurs des fonctions d'appartenance des ensembles flous peuvent être trans-

60. Soient N ensembles flous $A_1, \ldots, A_N$ et fus la fonction de fusion des fonctions d'appartenance $a_1, \ldots, a_N$ de $A_1, \ldots, A_N$ définie en fonction de l'opérateur choisi pour l'agrégation des ensembles flous, $e(x)$ est un élément neutre pour fus si :

$$fus(a_1(x), \ldots, a_{i-1}(x), e(x), a_{i+1}(x), \ldots, a_n(x)) = fus(a_1(x), \ldots, a_{i-1}(x), a_{i+1}(x), \ldots, a_n(x)).$$

61. Soient N ensembles flous $A_1, \ldots, A_N$ et fus la fonction de fusion des fonctions d'appartenance $a_1, \ldots, a_N$ de $A_1, \ldots, A_N$ définie en fonction de l'opérateur choisi pour l'agrégation des ensembles flous,

parentes pour le processus de fusion. Avec les intégrales de Choquet [Choquet, 1954], la fonction de paramétrisation permet de définir la valeur relative de chaque ensemble flou. Ainsi, le choix de l'opérateur d'agrégation dépend des objectifs de l'application.

Dans l'application à *BDFRues* de ce traitement, chaque caractéristique a sa propre importance. Chaque valeur des fonctions d'appartenance doit donc transparaître dans celle de la fonction d'appartenance de l'ensemble flou résultant de l'agrégation. La fusion des valeurs des fonctions d'appartenance ne doit donc pas avoir d'élément neutre ou absorbant. Cependant, l'importance de chaque caractéristique peut varier. Par exemple, l'information temporelle peut avoir plus d'importance que l'information de localisation. Dans ce cas, les valeurs des cellules de $TFHDateAcc$ doivent avoir une plus grande influence que celles de $TFHLocAcc$ dans le calcul des valeurs de la fonction d'appartenance de l'ensemble flou résultant de l'agrégation. Fusionner les valeurs des fonctions d'appartenance par une moyenne arithmétique pondérée appliquée à ces valeurs permet cela. La moyenne pondérée correspond pour les ensembles flous à une intégrale de Choquet spécifique [Detyniecki, 2000]. En utilisant cette fonction, l'agrégation des accumulateurs, dont le résultat sera l'ensemble flou Ag de fonction d'appartenance ag, peut être définie de la manière suivante :

$$ag(\rho,\theta) = \lambda * TFHLocAcc(\rho,\theta) + \mu * TFHOrienAcc(\rho,\theta) + \nu * TFHDateAcc(\rho,\theta),$$

avec les coefficients λ, ν et μ positifs et $\lambda + \nu + \mu = 1$.

Les coefficients sont définis lors de la paramétrisation de l'interrogation par les utilisateurs. Cette agrégation des accumulateurs correspond à une fusion des informations temporelle, spatiale et directionnelle des objets archéologiques par l'intermédiaire d'une recherche de forme géométrique dans les données.

5.2.3 Sélection des cellules

Dans la transformée de Hough, la méthode classique de sélection des objets détectés dans l'accumulateur consiste à prendre les maxima. Utiliser cette méthode dans le processus d'interrogation reviendrait à ne sélectionner que les éléments les plus pertinents. Si le but est d'obtenir des zones [62] de fort crédit, utiliser une α-coupe (voir 2.3.2) représente une bonne solution. La méthode d'interrogation affiche donc dans le SIG l'ensemble des instances de forme correspondant aux cellules sélectionnées dans l'accumulateur. La procédure d'interrogation est alors telle que proposée dans l'algorithme 5.4.

Le schéma de construction de l'interrogation spatiotemporelle sous critères de formes

$e(x)$ est un élément absorbant pour fus si :

$$fus(a_1(x),\ldots,e(x),\ldots,a_n(x)) = e(x).$$

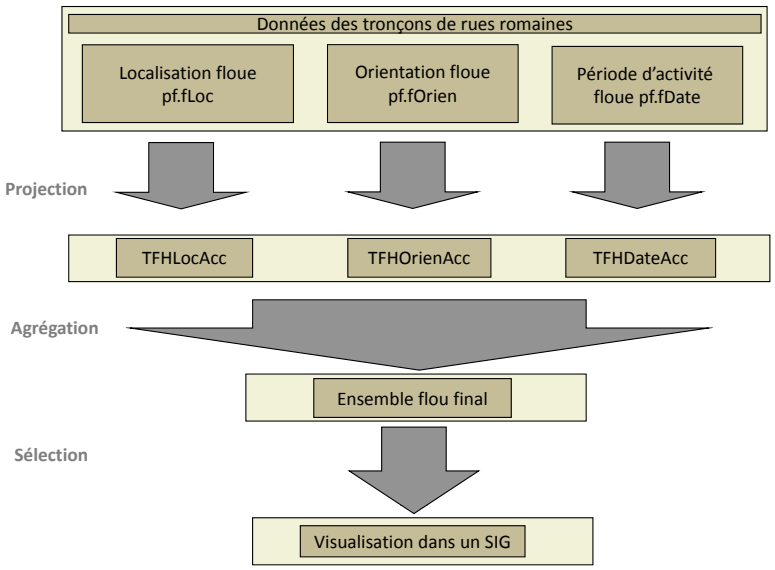
62. Le mot zone doit être compris comme le regroupement d'instances de formes spatialement proches.

Algorithme 5.4 : Interrogation spatiotemporelle de la base bd pour la période D_{ref} reconnaissant la forme $forme$

pour chaque ***objet*** o ***de*** bd **faire**
  vote de o dans les accumulateurs pour l'espace paramétrique correspondant à $forme$ en fonction de D_{ref} ;
fin
Ag ← Agrégation des accumulateurs ;
α ← valeur pour l'α-coupe pour la sélection ;
pour chaque $X \in A_\alpha$ **faire**
  Dessiner dans le SIG l'instance de la forme $forme$ associée à X ;
fin

est présenté dans la Figure 5.3.

Figure 5.3 – Schéma méthodologique pour l'interrogation sur critère de formes



La prochaine section porte sur l'utilisation de la méthode d'interrogation spatiotemporelle sur les données de *BDFRues* afin de déterminer l'influence des coefficients dans la fusion et de connaître la pertinence d'un tel outil d'analyse.

5.3 Application aux données archéologiques

Dans l'analyse du cadre applicatif *BDFRues*, le SIG doit afficher les segments de droites correspondant aux rues potentielles. Il doit de même afficher l'ensemble des droites de forts

crédits et non uniquement les droites ayant un crédit maximal. Pour cela, les résultats de l'interrogation sont insérés dans une nouvelle couche du SIG.

5.3.1 Influence des coefficients

Comme la fusion des informations se fait à l'aide d'une moyenne pondérée, les coefficients attribués aux accumulateurs amènent à différents résultats. L'impact de cette paramétrisation est étudié avec une sélection des valeurs dans le résultat de l'agrégation des accumulateurs par une α-coupe avec $\alpha = 0.95$.

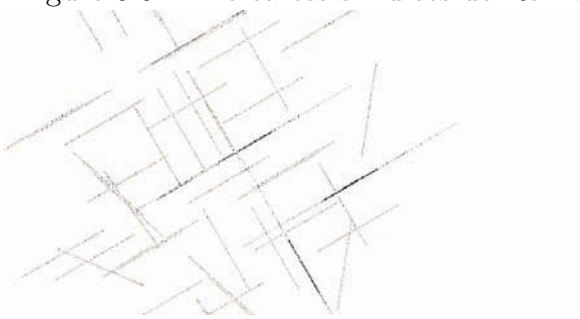
Figure 5.4 – Pré-cartes simulées de Reims obtenues avec $(\lambda, \mu, \nu) = (5/6, 1/12, 1/12)$

a : 3ème siècle

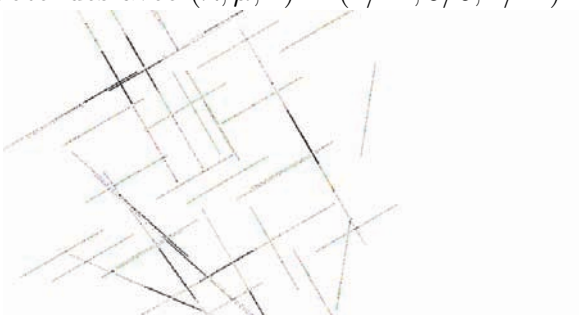
b : 4ème et 5ème siècles

Par exemple, les pré-cartes des figures 5.4a et 5.4b sont identiques et ne présentent qu'une seule droite car l'influence de la valeur de l'indice temporel et de celle de l'orientation est minimisée par rapport à l'influence de la valeur de la localisation c'est-à-dire que, dans la moyenne pondérée, les coefficients μ et ν affectés à $TFHDateAcc$ et à $TFHOrienAcc$ sont inférieurs à celui (λ) associé à $TFHLocAcc$ (dans l'exemple le coefficient associé à TFHLocAcc est dix fois supérieur aux autres). Ainsi, seule la droite de plus grand crédit spatial est affichée, c'est-à-dire la rue pour laquelle on a le plus de tronçons de rues présents dans la base.

Figure 5.5 – Pré-cartes simulées de Reims obtenues avec $(\lambda, \mu, \nu) = (1/12, 5/6, 1/12)$



a : 3ème siècle



b : 4ème et 5ème siècles

Quand le coefficient (μ) associé à $TFHOrienAcc$ est supérieur (dix fois supérieur, dans l'exemple des figures 5.5a et 5.5b) à ceux affectés à $TFHLocAcc$ (λ) et $TFHDateAcc$ (ν),

alors l'importance des localisations et celle des périodes d'activité sont minimisées par rapport à celle de l'orientation. Les pré-cartes résultantes de ces interrogations sont proches quelles que soient les périodes d'activité. Dans ce cas, les interrogations ne permettent l'affichage que de la principale orientation des rues.

Figure 5.6 – Pré-cartes simulées de Reims obtenues avec $(\lambda, \mu, \nu) = (1/12, 1/12, 5/6)$

a : 3$^{\text{ème}}$ siècle b : 4$^{\text{ème}}$ et 5$^{\text{ème}}$ siècles

Dans l'hypothèse où c'est le coefficient des périodes d'activité que l'on maximise, c'est à dire que le coefficient associé à $TFHDateAcc$ est supérieur aux autres dans la moyenne pondérée, alors les cartes différent totalement en fonction de la période souhaitée, mais toutes les évaluations des rues sont similaires comme le montre l'exemple des figures 5.6a et 5.6b où $\nu = 10 \times \lambda = 10 \times \mu$. Donc, l'objet de ce traitement est de visualiser l'influence des droites.

Figure 5.7 – Pré-cartes simulées de Reims obtenues avec $(\lambda, \mu, \nu) = (1/3, 1/3, 1/3)$

a : 3$^{\text{ème}}$ siècle b : 4$^{\text{ème}}$ et 5$^{\text{ème}}$ siècles

Dans les interrogations dont les résultats sont présentés dans la figure 5.7a et la figure 5.7b, une moyenne où tous les coefficients sont égaux est appliquée. Les résultats permettent bien de visualiser les rues, mais ne permettent pas d'identifier les éléments ayant les plus fortes valeurs. Dans ce cas, l'α-coupe réduit l'affichage aux rues dont l'indice temporel, la valeur de l'orientation et celle de la localisation sont maximaux.

Dans la section suivante, les résultats des interrogations sur les données pour l'estimation des présences potentielles des rues pour le 3$^{\text{ème}}$ siècle et pour les 4$^{\text{ème}}$ et 5$^{\text{ème}}$ siècles sont présentés.

5.3.2 Estimation des rues potentielles

J'ai utilisé la démarche précédente pour analyser les données spatiotemporelles imparfaites de *BDFRues* selon un critère de forme géométrique. Le but de cette analyse est de fournir des pré-cartes permettant d'estimer les présences des rues romaines de Reims en fonction des périodes définies par les utilisateurs. J'ai donc premièrement projeté les données de *BDFRues* dans des accumulateurs en adaptant la transformée floue de Hough, puis fusionné ces accumulateurs pour enfin sélectionner les droites pertinentes dans l'ensemble flou issu de la fusion.

Dans l'interrogation, les coefficients λ, μ et ν de la moyenne pondérée permettant d'obtenir l'ensemble flou Ag sont évalués empiriquement, et leurs valeurs sont respectivement 3/13, 1/13, et 9/13. Cette affectation donne trois fois plus d'importance au temps qu'à la localisation et aussi trois fois plus d'importance à la localisation qu'à l'orientation. Cela permet, à l'aide de la même méthode de sélection que dans le 5.3.1 (*i.e.* une α-coupe avec $\alpha = 0.95$), de définir des zones de présence potentielle correspondant assez bien à la période choisie, car l'influence des orientations et des localisations est minimisée.

Ces zones sont plus réduites que pour les traitements effectués dans la figure 5.6a et la figure 5.6b car l'importance de la localisation est plus grande que celle de l'orientation. Les résultats des interrogations sur *BDFRues* pour le 3$^{\text{ème}}$ siècle ou pour le 4$^{\text{ème}}$ et 5$^{\text{ème}}$ siècles permettent d'estimer les zones de présence potentielle des rues à ces époques.

La comparaison des pré-cartes issues de ce traitement (figure 5.8a et figure 5.8c) et des plans issus d'expertises (figure 5.8b et figure 5.8d) induit que l'interrogation spatiotemporelle sous critère de forme est une méthode d'analyse pratique pour l'aide à l'expertise archéologique.

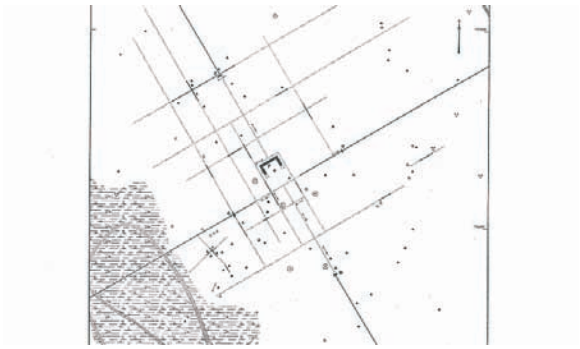
On peut remarquer que les rues issues du traitement et celles indiquées dans les plans issus d'expertises sont le plus souvent similaires, bien que quelques différences soient visibles. En effet, certaines rues présentes dans les pré-cartes n'ont pas d'équivalent dans celles présentes dans les plans d'experts. Ces différences peuvent s'expliquer par le fait que les plans d'expertises sont basés sur des informations anciennes. Comme il est coûteux en temps de refaire de nouvelles expertises à chaque nouvel objet découvert, ces plans ne sont pas régulièrement mis à jour. La démarche d'analyse proposée dans ce chapitre offre aux experts la possibilité de visualiser instantanément les scénarios possibles en fonction de leurs paramétrages.

La comparaison des pré-cartes (figure 5.8a et figure 5.8c) entre elles permet une étude de la dynamique temporelle des tronçons de rues. Ces pré-cartes diffèrent légérement, des rues présentent aux 3$^{\text{ème}}$ siècle ont disparu aux 4$^{\text{ème}}$ et 5$^{\text{ème}}$ siècles.

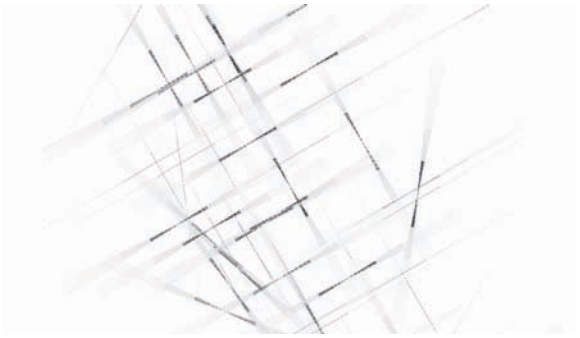
De plus, les experts évaluent l'intérêt des différents éléments et ne prennent pas en compte les moins intéressants. Dans l'interrogation, l'intérêt des éléments est défini par les valeurs de la fonction d'appartenance ag de l'ensemble flou Ag. Le processus de paramétrisation de l'interrogation, notamment la flexibilité laissée à l'utilisateur dans la définition des coefficients de la moyenne pondérée utilisée pour l'agrégation des accumu-

Figure 5.8 – Pré-cartes issues d'interrogations spatiotemporelles sous critère de forme — Plans d'experts

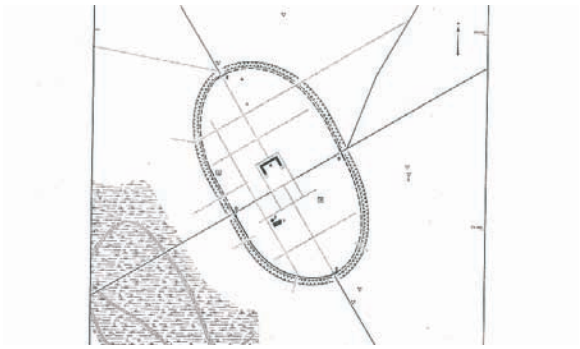
a : Pré-carte simulée pour le 3ème siècle



b : Plan d'experts pour le 3ème siècle



c : Pré-carte simulée pour les 4ème et 5ème siècles



d : Plan d'experts pour les 4ème et 5ème siècles

lateurs, permet à l'expert d'étudier plusieurs scénarios afin d'affiner ses travaux.

◇◇◇◇◇

L'intérêt principal de cet outil d'analyse est de définir les zones de présence possible des rues romaines pour une période donnée. Ces zones peuvent alors être utilisées, par les archéologues, pour l'aide au diagnostic. Elles peuvent de même contribuer à la mise en relation des objets afin de procéder à la généralisation de l'information. L'outil d'analyse proposé contribue fortement à l'intérêt de l'utilisation d'un SIG pour la gestion des données géographiques en permettant l'exploitation (l'analyse selon la forme) et la diffusion (pré-cartes simulées) de l'information et non uniquement son stockage.

Dans le contexte de ce travail, on dispose maintenant d'un outil qui a fait l'objet de deux communications internationales [de Runz *et al.*, 2007b] et [de Runz *et al.*, 2007c], d'un article dans une revue française [de Runz *et al.*, 2007a] et d'un article soumis à une

revue internationale [de Runz *et al.*, 2008d].

Les informations stockées dans un SIG archéologiques sont susceptibles d'être mises en correspondance avec d'autres informations issues d'une nouvelle interprétation ou d'une nouvelle source. Les données de la BDG du SIG doivent donc être appariées avec des données issues d'une autre BDG.

Dans ce cadre, la procédure classique est l'appariement de données, méthode aussi utilisée dans la démarche d'intégration de ces nouvelles bases dans le systèmes. C'est le sujet du prochain chapitre.

Chapitre 6

Analyse multi-source : appariement

Sommaire

La recherche de relations entre données géographiques est l'un des objectifs de l'analyse à l'aide d'un SIG. Les relations recherchées peuvent être spatiales, sémantiques, topologiques et/ou temporelles dans le contexte d'informations spatiotemporelles. Les données archéologiques à l'instar des données géographiques peuvent être issues de sources d'information différentes et peuvent représenter le même territoire. Dans le cadre de l'étude des évolutions structurelles d'un territoire au cours des temps, certaines données peuvent être issues d'une étude de ce territoire pendant la période romaine (à l'image des données de *BDFRues*) et les autres d'une étude de ce territoire mais pendant le Moyen-Âge. L'analyse multi-source associée à l'étude des évolutions structurelles peut, dans ce cas, être la recherche de relations spatiales, sémantiques, toponymiques et topologiques entre les objets stockés dans les différentes bases.

Cette recherche de relations peut se baser sur des techniques de mise en correspondance des objets comme l'intégration des données [Sheeren, 2005]. Classiquement, ces techniques

portent souvent sur l'appariement d'objets, c'est-à-dire assortir par paires les objets issus de sources différentes.

De nombreuses méthodes ont été proposées dans la littérature parmi lesquelles on peut, par exemple, citer [Devogele, 1997] et [Mantel et Lipeck, 2004]. Classiquement, considérons deux sources S_1 et S_2 dont les bases de données associées sont respectivement bd_1 et bd_2, pour chaque objet o_1 de bd_1, le processus interroge la base de données bd_2 afin de sélectionner l'objet o_2 correspondant le mieux à o_1. Pour cela, les techniques procèdent à une extraction de candidats possibles dans la base bd_2, et évaluent la correspondance des candidats à o_1.

Selon [Volz, 2006], les techniques d'appariement peuvent porter sur le couplage de points, de lignes et/ou de zones. Par exemple, les algorithmes de [Beeri *et al.*, 2004] portent sur l'appariement de points, celui proposé dans [Mantel et Lipeck, 2004] couple les lignes, celui de [Bel Hadj Ali, 2001] met en correspondance des zones, et la démarche exposée dans [Xiong et Sperling, 2004] combine à la fois les points, les lignes et les zones. Toutes ces différentes méthodes utilisent des fonctions d'évaluation de la correspondance.

L'étude de la correspondance entre objets est capitale pour l'évaluation des possibles correspondances. Cette étude nécessite la définition de critères basés sur des distances ou des indices, ces critères sont généralement des critères de similarité. Or, comme le remarque Ana-Maria Olteanu dans [Olteanu *et al.*, 2006], dans le cadre de l'information géographique, quelle que soit la méthode d'appariement utilisée, les critères sont de nature semblable. À partir de ce constat, elle propose, dans [Olteanu, 2007a] et [Olteanu, 2007b], une approche, basée sur les fonctions de croyance, adaptable aux données à mettre en correspondance. En utilisant les fonctions de croyance, l'approche d'Olteanu considère aussi le fait que les données géographiques sont imparfaites.

Comme les données archéologiques sont elles aussi imparfaites, l'utilisation de cette méthode dans le contexte de l'information archéologique semble possible. Cependant, pour cela, les critères utilisés devront être adaptés à l'information archéologique.

En effet, dans le contexte de l'information géographique, des critères géométriques, topologiques, descriptifs sont utilisés [Olteanu *et al.*, 2006]. En extension de cette démarche, afin de prendre en compte la dimension « temps » de mes données archéologiques, il est nécessaire de définir de nouveaux critères.

La démarche d'Olteanu se base sur la théorie des fonctions de croyance (voir 2.2) et sur le maximum de probabilité pignistique (voir 2.2.2) pour le choix des candidats. Les fonctions de croyance sont définies sur les valeurs possibles des critères d'évaluation de la correspondance entre objets. Les masses de croyance associées aux candidats sont donc directement liées aux valeurs des critères. Afin de donner plus de robustesse à la méthode, je propose de définir les fonctions de croyance sur les rangs attribués aux candidats par l'objet à apparier selon les critères choisis.

Dans ce chapitre, je présenterai dans un premier temps le contexte, le principe et les approches classiques de l'analyse multi-source en portant une attention particulière

à l'appariement. Dans un second temps, je proposerai des critères de correspondance pour des données archéologiques. Enfin, j'exposerai la méthode choisie puis j'expliquerai l'apport que peut être le passage par les rangs.

6.1 Analyse multi-source dans un SIG

De nos jours, une multitude d'informations géographiques issues de sources différentes et décrivant le même territoire est disponible. Ces informations sont généralement exploitables par l'intermédiaire de bases de données géographiques associées à des SIG. Avoir dans un SIG la capacité d'analyser les données issues des différentes sources est important afin de permettre, par exemple, le recoupement d'informations ou d'étudier les dynamiques spatiotemporelles.

Cependant, les données peuvent être exprimées à des échelles différentes et peuvent provenir de sources diverses ainsi que d'applications variées. Par conséquent, bien que les informations portent sur la même réalité terrain, une certaine indépendance entre les bases de données est observable [Devogele, 1997]. Il en va de même pour l'information archéologique. Cette indépendance complique l'analyse multi-source. Afin de permettre celles-ci, la mise en correspondance dans un même système de l'ensemble des données portant sur le même territoire est une solution riche de promesses.

6.1.1 Contexte général

La mise en correspondance de données à références spatiales est un champ d'actualité fécond de la recherche dans les Sciences de l'Information Géographique, particulièrement avec l'impulsion des nouvelles technologies pour la récolte et l'utilisation des informations. Elle nécessite une analyse détaillée des spécifications, des données et des schémas associés.

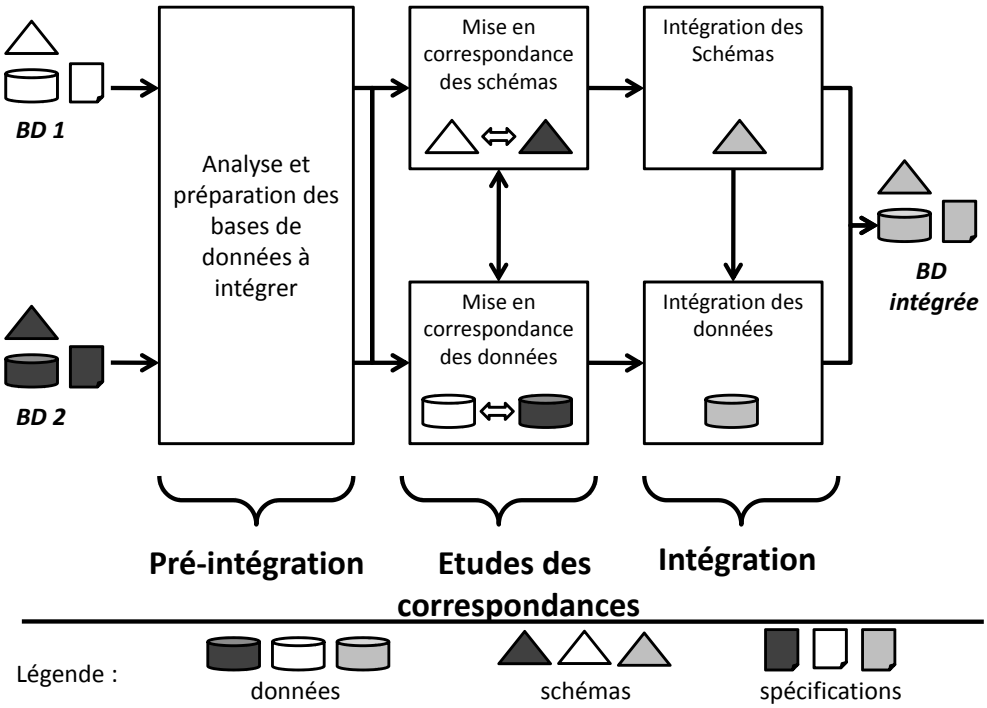
L'analyse de données multi-sources dans un SIG peut se faire par l'association à chaque source de données d'une couche d'information dans le SIG. L'analyse multi-source revient alors à mettre en correspondances les données issues des couches différentes. Elle peut aussi se faire via la communication réseau entre différents SIG et correspond alors à l'extraction des correspondances entre les bases de données ainsi reliées. Elle peut enfin se faire par l'intégration dans le SIG des différentes bases de données par la création d'une unique base des données résultant de la mise en relation des bases issues des différentes sources d'information.

6.1.2 Intégration des bases de données

L'intégration de bases de données dans un système consiste à trouver un modèle commun à toutes les bases afin de créer une nouvelle base, et à insérer les données issues des différentes bases dans cette nouvelle base en évitant les doublons dans l'information. Selon [Sheeren *et al.*, 2008], les méthodes d'intégration de bases de données comportent trois

phases calculatoires (voir la figure 6.1) : la pré-intégration, l'étude des correspondances et l'intégration proprement dite des données dans la base résultat.

Figure 6.1 – Cadre général de l'intégration de bases de données selon [Sheeren, 2005]



Durant la première phase, chaque base de données est étudiée pour obtenir une bonne vision de son contenu, afin d'arriver à un schéma commun nécessaire à la définition de la base résultant du processus.

L'étude des correspondances — la seconde phase — permet d'identifier les correspondances à la fois dans les schémas et dans les objets contenus dans les bases de données. Bien que les correspondances entre objets et schémas soient liées, pour les bases de données géographiques, du fait de l'absence d'identifieur universel, il faut mener, parallèlement mais séparément, la recherche de correspondances entre objets et entre schémas.

L'étape d'intégration des données termine les processus d'intégration des bases de données. Durant cette étape un nouveau schéma est construit et les objets sont, dans le système, fusionnés ou liés en fonction de la stratégie d'intégration choisie.

La base de données issue de l'intégration reste compatible avec les bases de données d'origine et inclut les mêmes entités géographiques mais avec des niveaux de détails différents. Le cadre général — figure 6.1 — de l'intégration permet par une vision globale et synthétique de bien appréhender le domaine de recherche.

Ainsi que ce soit par l'intégration des données, par la superposition de couches d'information dans le SIG, ou par la communication entre différents SIG, la mise en correspondance des objets est une démarche capitale pour l'analyse de données multi-sources. Cette dernière se fait généralement à l'aide d'une méthode d'appariement de données, c'est à dire qu'à chaque objet d'une base de données correspond au plus un objet de la seconde base.

6.1.3 Appariement des objets

L'appariement est une vision particulière de la mise en correspondance qui consiste à mettre en relation les éléments deux à deux. Elle est très utilisée dans les processus d'intégration de données géographiques. L'appariement est une procédure d'association dépendante de l'approche de mise en correspondance choisie. Une étude synthétique de ces approches est proposée dans la suite.

A l'instar du recalage en traitement d'images dont les principales méthodes sont présentées dans [Brown, 1992] et dont les techniques numériques [Modersitzki, 2004] sont régulièrement utilisées en imagerie médicale [Scherzer, 2006], quatre approches d'appariement de données géographiques sont prépondérantes [Volz, 2006] : par points, par lignes, par régions ou zones et les approches mixtes. On peut aussi classer, comme le proposent [Olteanu *et al.*, 2006], les méthodes selon qu'elles considèrent chaque objet géographique de manière isolée ou dans des réseaux (par exemple [Devogele, 1997]).

Les méthodes d'appariement de points regardent la similarité entre les points à lier. Par exemple l'algorithme proposé dans cadre du projet EVIDENCE [Pandazis *et al.*, 1999] se base sur l'idée de définir les intersections de rues comme des objets ponctuels caractérisés par leurs coordonnées, les abréviations et les noms des rues adjacentes de même que le nombre de rues passant par les points. L'appariement se fait à l'aide d'une étude comparative de ces objets. Dans [Beeri *et al.*, 2005], d'autres algorithmes pour des ensembles de points sont développés.

Les méthodes d'appariement de lignes sont particulièrement intéressantes pour l'appariement de réseaux de rues. Elles ont été étendues aux objets cartographiques dans [Mantel et Lipeck, 2004]. La première étape de ces méthodes liste toutes les correspondances potentielles des lignes connectées topologiquement. La liste obtenue contient un nombre important de paires. Puis, une sélection des paires les plus pertinentes s'effectue selon des paramètres relationnels (p.ex. topologiques) et des critères basés sur les caractéristiques (p.ex. les angles). Enfin, dans le but de réduire la liste des correspondances possibles à une unique combinaison de paires, chaque couple de lignes est alors évalué selon une fonction déterminant son mérite.

L'appariement de régions ou zones considère celles-ci comme des polygones. L'approche de Kraft, datant de 1995 et présentée succinctement dans [Volz, 2006], minimise les différences géométriques globales entre deux jeux de données par le prisme des aires de correspondance. Toutes les paires d'objets dont la distance entre les aires est en dessous

d'un seuil sont considérées comme paires potentielles. Toutes ces paires potentielles sont ensuite évaluées par une fonction de coût calculée par des critères pondérés comme la surface, le centre de gravité, etc. La liste des paires potentielles est ensuite ordonnée par les coûts afin d'obtenir la moins coûteuse combinaison de paires.

Les approches mixtes combinent les procédures à base de points, de lignes et/ou de régions. Par exemple, dans [Xiong et Sperling, 2004], les appariements de nœuds, de segments et de contours sont conjugués dans la procédure globale. Leur projet est de faire correspondre des caractéristiques extraites d'images satellitaires avec des bases de données existantes portant sur les réseaux routiers. Les nœuds sont regroupés en classes. La méthode évalue ensuite la similarité entre classes selon des critères géométriques et topologiques. C'est une approche semi-automatique, c'est-à-dire qu'une intervention extérieure est demandée.

Cependant,que les méthodes apparient des points, des lignes et/ou des zones, l'étude de la similarité entre données nécessite la définition de critères géométriques, topologiques, sémantiques et/ou toponymiques [Olteanu *et al.*, 2006]. Ces critères, bien que de même nature, sont, dans le cadre de l'information géographique, la base de l'étude présentée dans [Abadie *et al.*, 2007]. Cependant, de par les spécificités de l'information archéologique (voir Partie I), des critères adaptés aux données archéologiques doivent être définis.

Par exemple dans le contexte du projet *SIGRem*, il peut être envisagé, pour Reims, d'étudier l'évolution entre les tronçons de rues de la période antique et :

- ceux issus de fouilles archéologiques mais portant sur le Moyen-Âge,
- les rues que l'on aura préalablement extraites du cadastre napoléonien,
- les rues de la période actuelle extraites du parcellaire urbain de Reims contenus dans BD PARCELLAIRE[63].

Pour cela, il sera nécessaire de définir des critères pour évaluer les correspondances possibles entre les objets de *BDFRues* et les autres selon les orientations, les localisations, les descriptions et le temps associés aux tronçons.

6.2 Critères pour la correspondance d'objets archéologiques

L'information archéologique et l'information géographique contiennent en commun la composante topologique et la composante géométrique. La composante descriptive et fonctionnelle de l'information archéologique est très proche de la composante descriptive et sémantique en géographie. La composante temporelle présente dans l'information archéologique n'est généralement pas considérée comme telle dans l'information géographique (le caractère temporel est le plus souvent associé à un attribut de l'information).

63. BD PARCELLAIRE est une BDG élaborée par l'IGN à partir de l'assemblage du plan cadastral dématérialisé.

Ainsi, les critères de correspondance topologiques et géométriques utilisés dans le cadre de l'appariement de données géographiques doivent pouvoir aussi être utilisés pour la mise en correspondance de données archéologiques. Cependant, l'aspect lacunaire des ces données nuance ce jugement. En effet, si on considère deux tronçons de rues issus de sources différentes représentant la même rue mais spatialement éloignés, les critères géométriques classiques ne pourront les associer sauf si l'on considère les objets englobants (les rues et non les tronçons) issus par exemple de l'analyse proposée dans le chapitre II.

De plus, il faut adapter les critères sémantiques utilisés pour l'appariement de données géographiques au contexte des données archéologiques. Il faut par exemple considérer les échelles fonctionnelles de l'information archéologiques dans les critères sémantiques. Par contre, les critères toponymiques sont utilisables directement.

Enfin, coupler des objets de la même époque amène à utiliser des critères de similarité temporelle. temporel.

L'appariement de données archéologiques amène à l'utilisation soit de critères classiques en géomatique, soit de critères spécifiques à la nature de l'information archéologique. Le contexte d'utilisation est la clé pour le choix des critères. Je propose dans la suite quelques critères géométriques, descriptifs (sémantiques et toponymiques) et temporels utilisables pour l'appariement de données archéologiques comme par exemple les données de *BDFRues*.

6.2.1 Critères géométriques

Comparer les objets archéologiques selon la composante géométrique nécessite l'utilisation de critères s'adaptant notamment aux formes des objets à apparier et à la modélisation de ces objets tenant compte de leurs imperfections. De plus, en fonction de l'objet de l'analyse, la lacune associée aux objets archéologiques peut amener à considérer plusieurs critères géométriques.

Par exemple, pour l'appariement de bases de données avec les données de *BDFRues*, le critère géométrique devra être choisi par rapport à la théorie des ensembles flous.

Pour la similarité de localisation, un indice de dissimilarité sur des ensembles flous de domaine de définition de dimension 2 devra être utilisé afin d'obtenir son dual (indice de similarité). On peut soit définir cet indice en projetant les ensembles flous spatiaux sur des ensembles à 1 dimension (voir 8.2.2) soit choisir une mesure *ad-hoc*.

De même, pour les données de *BDFRues*, il faut aussi regarder la similarité dans les orientations des objets archéologiques. Pour cela, on peut utiliser la distance de Hamming (voir 4.1.3) ou la métrique proposée dans [Grzegorzewski, 1998] par la relation duale. D'autres indices de dissimilarité entre ensembles flous sont présentés dans [Lowen et Peeters, 1998].

La combinaison de critères géométriques pour la similarité spatiale (localisation et orientation) peut limiter l'impact de la lacune de l'information archéologique. En effet, pour les données de *BDFRues* par exemple, elle peut coupler des tronçons de rues pas

très éloignés mais définissant la même rue et éviter d'apparier des tronçons de rues spatialement proches mais d'orientations éloignées ou de rues d'orientations proches mais spatialement éloignées. La combinaison permet donc de limiter le nombre de mauvaises associations d'objets et facilite la résolution des conflits portant sur la définition d'objets se recoupant dans les différentes sources.

Cependant, comme en géographie, comparer les composantes descriptives des données est nécessaire dans le cadre d'une analyse de l'information dans sa complexité.

6.2.2 Critères descriptifs

Comparer les informations contenues dans la composante descriptive nécessite de considérer deux aspects principaux de cette composante : l'aspect sémantique et l'aspect toponymique. Le premier donne du sens à l'information, le second la spécifie. Prenons par exemple l'*avenue des Gobelins* à Paris, son aspect sémantique portera sur son rôle (sa fonction) et portera donc sur le sens du mot *avenue* tandis que les mots *avenue des Gobelins* définiront l'aspect toponymique de l'objet étudié. Regarder l'information descriptive selon ces deux aspects est donc nécessaire.

Dans le contexte de données archéologiques, le choix d'un critère sémantique met en évidence le problème de la non-spécificité des éléments et définir un critère de similarité entre les objets est complexe. En effet, l'appariement de données à l'échelle de la structure avec des données à l'échelle du fait nécessitera notamment de comparer les informations selon la qualification thématique de l'objet et de la position de celle-ci dans la structuration sémantique (lattice, ontologie, etc.) associée aux descriptions des objets archéologiques.

De plus, si l'on considère deux bases de données portant sur des époques différentes mais à la même échelle, l'aspect toponymique est alors capital. Il permettra de lier par exemple deux objets aux sémantiques différentes mais à la toponymie proche. Cela permet de lier par exemple, à Paris, la *prison de la bastille* du 18^{e} siècle à la *place de la Bastille* actuelle.

Dans le cadre de *BDFRues*, le problème de la structuration sémantique ne se pose pas, toutes les données se trouvant forcément à la même échelle. Par contre, la similarité toponymique pourra et devra être utilisée. Les distances de Hamming [Hamming, 1950] et de Levenshtein [Levenshtein, 1966] sont de bons candidats à l'évaluation de la similarité toponymique. Cependant, en fonction du temps séparant les périodes d'activité des objets à apparier, il faudra, si nécessaire, enlever la partie sémantique de l'information descriptive.

La composante temporelle des données archéologiques à apparier est donc un aspect capital pour le bon couplage des objets associés aux données.

6.2.3 Critère temporel

Le choix de l'indice de similarité entre périodes d'activité des objets archéologiques dépend de l'objectif de l'analyse. Si son but est l'étude de la dynamique des objets, la comparaison du positionnement temporel des objets est un bon critère. Par contre, si le but de l'analyse est le recoupement de l'information, apparier les objets selon qu'ils sont ou non contemporains entre eux est une solution.

Pour les données de *BDFRues*, si l'analyse a pour but le positionnement temporel entre objets, l'indice d'antériorité (*cf.* 4.2) peut être un critère intéressant. Si l'objectif de l'analyse est de faire correspondre des objets selon leur capacité à être contemporain entre eux, l'indice utilisé dans l'interrogation spatiotemporelle sous critère de forme (voir 5.2.1) pour la projection temporelle des objets dans l'espace des formes ou bien la distance dite de Hamming (voir 4.1.3) sont des critères adéquats.

Dans le contexte de l'appariement d'objets archéologiques, les critères doivent être combinés et, si possible, traités simultanément. De plus, les données archéologiques étant imparfaites, la technique d'appariement doit pouvoir en tenir compte. La démarche proposée par Ana-Maria Olteanu dans notamment [Olteanu, 2007a] regarde les données avec leurs imperfections et considère les critères simultanément.

6.3 Appariement de données spatiotemporelles utilisant les fonctions de croyance

Les critères d'évaluation de la correspondance entre objets ne sont pas utilisés simultanément dans les approches classiques, mais les uns après les autres. De plus, ces approches ne modélisent pas explicitement l'imperfection des données. L'objectif de la démarche présentée par Ana-Maria Olteanu dans [Olteanu, 2007a] et [Olteanu, 2007b] est de proposer une stratégie qui considère l'imperfection des données et utilise les critères simultanément.

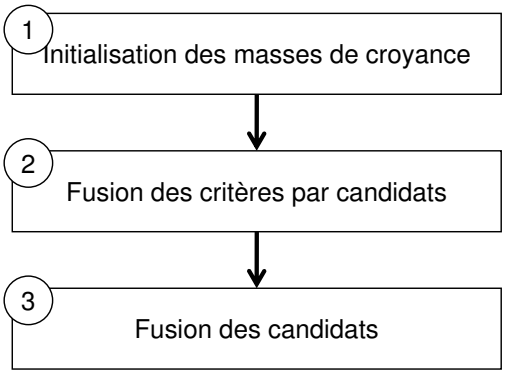
Pour cela, elle se place dans le cadre de la théorie de l'évidence. Elle fait ainsi le choix d'une théorie faisant le pont entre l'incertitude et l'imprécision. Le choix de ce formalisme est d'autant plus judicieux que l'on va vouloir fusionner de l'information venant de plusieurs sources ne donnant pas forcément la même définition à un objet commun. On fait donc face à de l'ambiguïté et plus spécifiquement à du conflit. Ce choix de représentation se justifie pleinement selon la taxonomie de Peter Fisher (voir 3.1.1 et [Fisher *et al.*, 2005]) pour l'information géographique et selon la taxonomie proposée dans le chapitre 3 (voir figure 3.3).

De plus, grâce à la théorie des fonctions de croyance, les critères d'évaluation de la possible correspondance peuvent être traités simultanément en associant à chaque critère de correspondance une fonction de croyance et en les fusionnant par la suite.

6.3.1 Appariement selon Olteanu

La méthode d'appariement est structurée en trois étapes (voir figure 6.2) : l'initialisation des masses de croyance, la fusion des critères par candidats et la fusion des candidats. Elle fusionne donc les informations localement puis globalement.

Figure 6.2 – Étapes de la méthode d'Olteanu [Olteanu, 2007a, Olteanu, 2007b]



L'utilisation de la théorie des fonctions de croyances nécessite la définition d'un cadre de discernement. Considérons deux bases de données bd_1 contenant M objets et bd_2 en contenant N. Soit P le nombre de critères d'évaluation de la correspondance. La démarche vise à coupler des objets de bd_1 avec des objets de bd_2. Les objets de bd_2 seront considérés comme les candidats. Comme tout objet de la base ne doit pas nécessairement être apparié, l'hypothèse NA (non apparié) est définie.

Pour chaque objet o_j, $j \in [1, M]$, de bd_1, on extrait les objets de bd_2 dont la distance spatiale est inférieure à un seuil afin d'éliminer un maximum de candidats impossibles et de limiter le temps de calcul. Le cardinal de l'ensemble des objets extraits de bd_2 est égal à P. On a donc P candidats à l'appariement avec o_j.

Le candidat numéro i et le critère numéro k seront respectivement notés c_i et CS_k.

Le cadre de discernement pour l'appariement de l'objet o_j est alors défini de la manière suivante :

$$o_j : \Omega = \{c_1, c_2, \ldots, c_P, NA\}, j \in [1, M].$$

Ω est défini comme exhaustif c'est-à-dire que la masse de l'ensemble vide est nulle.

La première étape de la méthode consiste à initialiser les masses de croyance pour chaque candidat c_i et pour chaque critère CS_k. De manière indépendante à l'objet en cours d'analyse, des fonctions de croyance sont définies sur les valeurs possibles de critères d'appariement.

Pour chaque objet o_j, pour chaque critère CS_k, pour chaque candidat c_i, on calcule trois masses de croyance à partir de la valeur de l'indice CS_k pour o_j et c_i :

- $m_{j,k}(c_i)$ la masse de confiance sur l'appariement de o_j avec c_i pour le critère CS_k,
- $m_{j,k}(\neg c_i)$ la masse de confiance sur le non appariement de o_j avec c_i pour le critère CS_k,
- $m_{j,k}(\Omega)$ la masse de confiance pour l'ignorance sur le fait que l'objet o_j doive ou non être apparié avec c_i pour le critère CS_k .

La contrainte associée à ces masses est :

$$m_{j,k}(c_i) + m_{j,k}(\neg c_i) + m_{j,k}(\Omega) = 1.$$

Pour un objet o_j, pour un candidat c_i et un critère de similarité CS_k, l'hypothèse de base pour l'initialisation des masses de croyance est que plus la similarité est forte entre o_j et c_i (plus la dissimilarité est faible) plus les chances que c_i soit l'objet homologue à o_j sont fortes (i.e. $m_{j,k}(c_i)$ sera grand et $m_{j,k}(\neg c_i)$ sera petit). A contrario, plus la dissimilarité entre o_j et c_i est forte (plus la similarité est faible) plus les chances que c_i soit l'objet homologue à o_j sont faibles (i.e. $m_{j,k}(c_i)$ sera petit et $m_{j,k}(\neg c_i)$ sera grand). Entre les deux, on se situe dans l'ignorance.

La seconde étape de la méthode d'appariement est une fusion locale des critères pour chaque candidat. Les candidats sont analysés séparément.

Pour chaque candidat, les critères sont fusionnés à l'aide de la règle de combinaison de Dempster (voir 2.2.2). Soit $m_{j,12}$ la masse résultant de la combinaison de deux critères CS_1 et CS_2 selon la règle de Dempster pour le candidat c_i est :

$$m_{j,12}(c_i) = (m_{j,1} \otimes m_{j,2})(c_i) = \frac{\sum_{B,C \subseteq \Omega \& B \cap C = c_i} m_{j,1}(B) \times m_{j,2}(C)}{1 - \sum_{B,C \subseteq \Omega \& B \cap C = \emptyset} m_{j,1}(B) \times m_{j,2}(C)}$$

Ce qui permet d'obtenir pour le candidat c_i et les deux critères CS_1 et CS_2, les masses de croyance suivantes :

$$m_{j,12}(c_i), m_{j,12}(\neg c_i), m_{j,12}(\Omega).$$

Dans cette étape, tous les critères sont fusionnés pour chaque candidat en suivant la même démarche que celle illustrée pour deux critères.

Après avoir fusionné localement (fusion des critères pour chaque candidat) les masses de croyance, une fusion globale (fusion des candidats) est faite.

Dans cette dernière étape, les résultats de l'étape précédente sont combinés. La démarche est la suivante : soient trois candidats c_1, c_2 et c_3 et deux critères CS_1 et CS_2, les résultats des candidats c_1 et c_2 sont combinés, puis la combinaison de c_1 et c_2 est elle-même combinée avec c_3. C'est-à-dire que l'on combine $m_{j,12}(c_1), m_{j,12}(\neg c_1)$ avec $m_{j,12}(c_2), m_{j,12}(\neg c_2)$. Puis les résultats de la précédente combinaison sont combinés avec $m_{j,12}(c_3), m_{j,12}(\neg c_3)$.

Cette démarche est alors généralisée à tous les objets de bd_1 et à tous les candidats associés à ces objets. Le choix du candidat (ou le non choix) se fait ensuite dans le cadre

des modèles de croyances transférables (voir 2.2.3) en utilisant le maximum de probabilité pignistique après la fusion globale.

6.3.2 Passage par les rangs

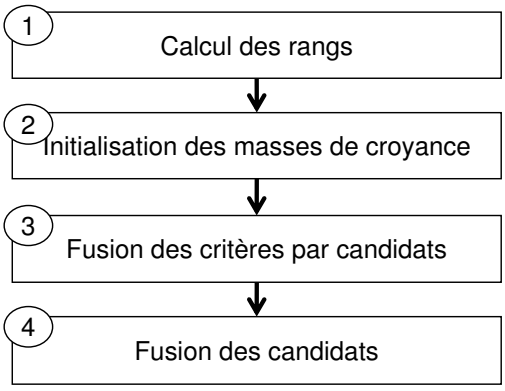
Dans la méthode d'appariement exposée précédemment, il y a une forte dépendance aux similarités dans l'obtention des masses de croyance. Une variation dans les informations associées aux objets affecte les valeurs des masses de croyance et peut donc provoquer de mauvaises associations. Afin de gagner en robustesse, je propose d'ajouter une étape intermédiaire entre le calcul des dissimilarités et la détermination des masses de croyance : le calcul des rangs des candidats.

Les rangs sont obtenus ainsi : on classe les candidats selon la distance les séparant de l'entité à apparier. On passe donc de la notion de distance à la notion de rang ajoutant ainsi un peu plus d'abstraction au modèle.

Cette distanciation entre les critères et les masses de croyances a pour but de réduire les mauvaises associations d'objets dues à des problèmes de déformation pour le cas de données rasters, de mises à l'échelle, etc.

Pour les données archéologiques qui sont généralement imparfaites, la robustesse vis-à-vis de l'imperfection est un élément important pour l'efficacité de l'appariement. L'utilisation des rangs a donc pour objectif d'augmenter la robustesse de la méthode proposée par Ana-Maria Olteanu. Cela impose donc de structurer l'appariement selon les étapes présentées dans la figure 6.3.

Figure 6.3 – Étapes de l'approche utilisant les rangs



Le principe pour l'initialisation des masses de croyance est le suivant : meilleur est le classement d'un candidat pour l'objet-cible, plus la croyance dans son appariement avec l'objet-cible sera élevée.

L'initialisation des masses de croyance se passe donc de la manière suivante : pour chaque objet o_j, pour chaque candidat c_i pour chaque critère CS_k, les masses de croyance

ne seront pas initialisées directement à partir de la valeur du critère CS_k pour o_j et c_i mais du rang $Rg_{(k,o_j)}(c_i)$ que cette valeur procure à c_i dans le classement fait par o_j des différents candidats selon les valeurs du critère pour chaque candidat.

Cette étape est classiquement utilisée en statistique pour ajouter de la robustesse à l'évaluation de la position d'une donnée ou d'un échantillon ; le médian est un estimateur de position de l'échantillon. Les statistiques d'ordres (voir [David et Nagaraja, 2003]) sont en informatique de plus en plus utilisées notamment en traitement numérique des images à l'instar de la méthode proposée dans [Vautrot *et al.*, 2006]. En passant par cette étape, les filtres proposés sont moins dépendants aux variations dans les données. C'est aussi l'objectif de l'ajout de cette étape à la démarche d'Olteanu.

◇◇◇◇◇

L'analyse multi-source se fonde classiquement sur la mise en correspondance des données issues des différentes sources. Dans cette démarche, l'appariement des données est une démarche souvent utilisée.

L'appariement de données nécessite la définition de critères évaluant les correspondances entre les objets. J'ai présenté, dans ce chapitre, une réflexion sur le choix de ces critères dans le cadre des données archéologiques. Celles-ci présentent plusieurs composantes et différentes formes d'imperfection. Les critères d'évaluation des correspondances doivent les prendre en considération, c'est pourquoi j'ai suggéré l'utilisation de plusieurs critères simultanément.

La démarche d'Olteanu permet à la fois de considérer l'imperfection de l'information mais aussi d'évaluer les possibles appariements en traitant les critères de manière simultanée. Pour cela, elle se base sur la théorie de l'évidence. Sa démarche est structurée en 3 étapes : initialisation des masses de croyance, fusion des critères par candidats, fusion des candidats.

Dans la première étape, elle associe à chaque critère des fonctions de croyance définies sur les valeurs possibles des critères d'évaluation de la correspondance. Afin d'augmenter la robustesse de la méthode, j'ai proposé l'ajout d'une étape calculatoire — le passage par les rangs — en préalable à l'initialisation des masses de croyance. Les fonctions de croyance sont alors définies sur les rangs possibles des candidats et non sur les valeurs possibles des critères d'évaluation de la correspondance. Le calcul des rangs permet la distanciation entre valeurs des critères et masses de croyance.

Ainsi, dans ce chapitre, j'ai présenté une approche de l'analyse multi-source de données spatiotemporelles imparfaites. Cette approche considère les différentes caractéristiques des données dans le choix des critères et dans le choix de la méthode d'appariement et son

adaptation au contexte archéologique qui nécessite plus de robustesse pur tenir compte de l'imperfection des données.

Conclusion

Dans la problématique de la généralisation de l'information archéologique, l'analyse des données avec leurs imperfections est nécessaire. Dans cette optique, l'objet de cette partie fut l'analyse de données spatiotemporelles imparfaites à partir d'informations extérieures au SIG ; les données ont été préalablement modélisées selon la démarche proposée dans la partie I.

Dans ce contexte, j'ai présenté trois méthodes d'analyse de données. La première est une analyse temporelle de l'information basée sur l'étude de l'antériorité aux données. Une telle analyse uniquement temporelle de données imparfaites est rarement mise en œuvre dans un SIG et elle s'avère néamoins indispensable en archéologie. La seconde est une analyse spatiotemporelle suggérant des complétions possibles aux données lacunaires en fonction du temps et d'une forme géométrique. Elle nécessite la fusion d'information de nature différentes sous critère de forme, ce qui constitue une approche originale généralisable à d'autres domaines applicatifs. La troisième méthode présente une démarche pour l'analyse multi-source se basant sur l'appariement de données de façon robuste pour tenir compte de l'imperfection. Les deux premières méthodes proposent des outils d'interrogation, la troisième exploite une technique d'appariement.

L'interrogation temporelle selon l'antériorité présentée dans le chapitre 4 utilise une quantification de l'antériorité. Cette quantification, appelée indice d'antériorité, permet de nuancer la validité du positionnement temporel des objets. Cet indice est basé sur celui de Kerre [Kerre, 1982] et se place dans la démarche d'une extension du FMO [Ramik et Rimanek, 1985]. Les valeurs de l'antériorité d'une période fournie en entrée vis-à-vis des objets sont visualisées durant l'analyse par l'intermédiaire d'une couleur affectée à chaque objet du SIG.

La seconde interrogation, exposée dans le chapitre 5, a pour objectif d'aider à la reconstruction du passé malgré la lacune de l'information archéologique. Pour cela, elle utilise les localisations des objets archéologiques présents dans la BDG associée au SIG, la forme géométrique supposée des objets englobant ceux de la BDG et la période sujet de la recherche. Par l'utilisation des principes d'une méthode de reconnaissance de forme (la transformée de Hough), l'analyse présentée estime les présences potentielles des objets englobants en abordant ainsi la notion de facteur d'échelle.

Dans le chapitre 6, j'ai étudié l'appariement des données géographiques afin de proposer une démarche applicable aux données archéologiques. Cette approche analytique

s'appuie sur l'étude de critères utilisables pour l'information archéologique au regard de l'imperfection des données. Cette approche permet de traiter les critères simultanément. Elle combine des fonctions de croyance définies sur les valeurs possibles des différents critères. Pour gagner en robustesse, j'ai proposé de calculer les rangs des objets candidats pour les objets cibles et de définir les fonctions de croyance sur leurs rangs. Ce besoin de robustesse devient plus imperatif dans un contexte d'utilisation de données largement imparfaites.

Les analyses utilisant des processus d'interrogation présentées dans cette partie ont été exposées aux communautés scientifiques en informatique et en géomatique à travers trois communications internationales ([de Runz *et al.*, 2006], [de Runz *et al.*, 2007b] et [de Runz *et al.*, 2007c]). De plus, les résultats de l'interrogation spatiotemporelle sous critère de forme ont été présentés aux experts à travers le magazine « Culture et Recherche » publié par le Ministère de la Culture. Par ailleurs, deux articles ont été soumis à des revues internationales ([de Runz *et al.*, 2008d] et [de Runz *et al.*, 2008c]).

Cependant, l'analyse de l'interaction des données avec de l'information extérieure n'est pas la seule démarche possible. La recherche des relations entre objets déjà présents dans la BDG associée au SIG représente une autre démarche analytique : l'analyse exploratoire. De plus, une fonction essentielle des SIG est la visualisation. L'analyse exploratoire et la visualisation sont au cœur de la partie III de ce mémoire.

Troisième partie

Analyse exploratoire et visualisation de données spatiotemporelles imparfaites dans un SIG archéologique

« The drinkers were the usual bunch of heroes, cut throats, mercenaries, desperados and villians, and only microscopic analysis could have told which was which. »

Terry Pratchett, *Annals of the Discworld — Hogfather* (1996).

-

« Granny Weatherwax was not lost. She wasn't the kind of person who ever became lost. It was just that, at the moment, while she knew exactly where she was, she didn't know the position of anywhere else. »

Terry Pratchett, *Annals of the Discworld — Wyrd Systers* (1988).

Introduction

L'utilisation des SIG en archéologie se justifie grandement par la disponibilité des méthodes d'analyse de l'information spatiotemporelle mais aussi des processus de visualisation. Les méthodes d'analyse permettent de révéler notamment le positionnement de chaque objet contenu dans la base de données par rapport aux autres objets de celle-ci. Les processus de visualisation permettent à l'utilisateur d'appréhender graphiquement la plus grande part possible de l'information associée à chaque objet ou des relations entre objets. Dans cette partie, je m'intéresse à deux aspects : d'une part, l'analyse exploratoire des données pour étudier le positionnement temporel des objets et pour extraire les objets les plus représentatifs de la base, d'autre part la visualisation des objets tant par leur composantes temporelles que par leurs dissimilarités.

D'un côté, l'analyse exploratoire de données, aussi appelée fouille ou exploration de données, utilise un ensemble d'algorithmes issus de disciplines scientifiques diverses (statistiques, intelligence artificielle, base de données) pour extraire un maximum de connaissances utiles. Dans ce mémoire, afin d'aider à l'expertise archéologique et à l'utilisation des SIG, les algorithmes exposés ont pour but de dégager les positions temporelles des objets de *BDFRues* (chapitre 7) ainsi que d'extraire les tronçons de rues les plus représentatifs (chapitre 8).

De l'autre, la visualisation des données rassemble toutes les techniques pour créer des images, des diagrammes ou des animations afin de représenter de l'information. Elle donne à l'utilisateur la possibilité d'explorer de larges bases de données par l'intermédiaire d'outils informatiques. Les techniques de visualisation proposent de faciliter la compréhension humaine des données par l'intermédiaire de sélections, de transformations et de représentations d'un résumé des celles-ci. Ce sont des techniques devant permettre à l'utilisateur de découvrir intuitivement des formes ou/et des structures liant les données. La technique proposée dans le chapitre 9 fournit une image couleur représentant des informations quantitatives relatives aux objets de *BDFRues*. Le but est d'explorer des masses de données pour une meilleure compréhension de celles-ci.

Dans le contexte de l'analyse exploratoire, les processus présentés ont pour but de déterminer le positionnement temporel et de regarder la représentativité de chaque objet au regard de tous ceux de la base.

Dans le chapitre 7, j'examine les liens temporels entre les objets afin de dégager si possible une structure temporelle de l'ensemble des données. L'information disponible étant imparfaite, les périodes d'activité des objets de *BDFRues* sont représentées par des nombres flous. L'étude du positionnement temporel d'un objet nécessite donc le positionnement du nombre flou représentant sa période d'activité dans l'ensemble des nombres flous associés aux autres objets de la base. De nombreuses méthodes, à l'instar de celle de [Kerre, 1982] ou de [Jain, 1977], rangent les éléments d'un ensemble de nombres flous. Cependant, ces méthodes considèrent le nombre flou à ranger par rapport à l'ensemble et non pas par rapport à chacune des données séparément. En effet, des comparaisons deux à deux génèrant des incohérences, les approches classiques élaborent un classement en considérant l'ensemble des données.

Dans l'approche exposée dans ce mémoire, j'utilise l'indice d'antériorité proposé dans 4.2.1. Cet indice, qui définit la capacité d'un objet à être antérieur à un autre, apporte plus d'information qu'une décision binaire (antérieur/non antérieur). En calculant cet indice pour chaque couple d'objets de *BDFRues*, je construis un graphe orienté pondéré dont les sommets seront les tronçons de rues et dont les arcs auront pour coût la valeur de l'indice d'antériorité de l'origine de l'arc par rapport à la destination. À l'aide des coûts des arcs entrants et sortants d'un sommet, je détermine le potentiel d'antériorité, de postériorité ainsi que la position temporelle de l'objet associé au sommet dans l'ensemble des objets de la base. Ce graphe permet une représentation synthétique et formelle des structures temporelles de l'ensemble des objets de la base de données. L'information dégagée par ce graphe met en lumière les rapports temporels entre objets et permet notamment d'extraire l'objet le plus antérieur ou l'objet le plus postérieur à l'ensemble des autres. L'étude des positions temporelles dégage une connaissance nouvelle sur les données ; ce n'est cependant pas la seule connaissance que l'on peut obtenir par un processus d'exploration des données.

Le chapitre 8 présente un processus exploratoire ayant pour objectif la recherche d'éléments représentatifs. En effet, déterminer l'élément qui représente le mieux un ensemble de fouilles archéologiques en terme de localisation, de période et de forme est important pour la classification des données car les éléments représentatifs sont des objets réels reflétant l'ensemble de la base.

Cette recherche exploite la notion de représentativité introduite par Frédéric Blanchard dans [Blanchard, 2005]. Cette notion m'a permis de définir le Vecteur de Meilleur Rang Moyen (*VMRM*). Ce vecteur est une statistique de rang qui sélectionne l'élément le plus représentatif d'un échantillon. Les statistiques de rangs augmente la robustesse nécessaire à la définition d'un bon représentant, particuliérement de données imparfaites. La notion —quantitative— de représentativité est basée sur la transformation par rangs des dissimilarités entre éléments. J'appliquerai cette méthode statistique à la recherche de l'élément le plus représentatif des données de *BDFRues* selon leurs localisations, leurs orientations, leurs périodes d'activité et des trois conjuguées. Pour cela, afin de tenir compte de l'imperfection de l'information, le processus utilise des critères de dissimilarité

temporelle, d'orientation et de localisation. Les objets extraits estiment la position de *BDFRues* et sont donc caractéristiques des objets contenus dans *BDFRues* et donc des rues gallo-romaines trouvées à Reims. Par ce processus exploratoire, après avoir regardé la représentativité de chaque objet archéologique relativement à la base de données, j'extrais des représentants de la base au regard de leurs représentativités selon les critères choisis.

L'analyse et l'exploration des données spatiotemporelles dans un SIG ne sont pas les seules fonctionnalités nécessaires à la compréhension des données, il faut aussi les visualiser. Généralement, la visualisation se fait dans un SIG par la production de cartes thématiques. Dans le chapitre 9, la volonté est d'utiliser les informations quantitatives associées aux objets d'une base de données afin de dégager visuellement des groupes d'objets. Les données sont vues ici comme des collections d'information [Guptill, 2005]. Les données quantitatives associées aux objets sont assimilées à des données multidimensionnelles : chaque attribut est une dimension. La démarche proposée exploite une technique de visualisation orientée pixel [64] représentant les données par une image couleur [Blanchard *et al.*, 2005a]. En attribuant un pixel couleur à chaque objet et en organisant spatialement les pixels dans une image, l'approche exposée veut regrouper par similitude couleur et proximité spatiale, les objets dont les informations descriptives ou temporelles sont proches.

Pour les données de *BDFRues*, l'ambition est de permettre, par ce type d'image, non seulement une analyse temporelle de la base, mais aussi une analyse de la similarité des objets à un objet particulier. Dans le premier cas, on considère les ensembles flous représentant les périodes d'activité des objets de *BDFRues*. Pour chaque objet, on quantifie sa représentation temporelle par de multiples techniques [65] afin d'avoir des valeurs numériques le décrivant, et on applique la technique de visualisation sur ces valeurs. Dans le second cas, on calcule, pour un objet sélectionné dans la base, les dissimilarités des autres objets à l'objet sélectionné en fonction des représentations en ensembles flous des localisations, des orientations et des périodes d'activité. J'applique ensuite la méthode de visualisation sur ces dissimilarités. Les images résultantes de ces visualisations fournissent des cartes synthétiques permettant une analyse intuitive de l'information.

Cette partie propose donc un ensemble de techniques de fouilles de données et d'exploration visuelle dédié à l'analyse de données spatiotemporelles et adapté plus spécifiquement aux données imparfaites. Ces techniques exploratoires ou pré-exploratoires traitent aussi bien du positionnement temporel, de la représentativité et de la similarité des données tant numériquement que visuellement.

64. Pixel ou picture element : unité de base d'une image 2D

65. Les techniques de défuzzification d'ensembles flous utilisées dans le chapitre 9 sont extraites de [VanLeekwijck et Kerre, 1999] et permettent d'obtenir des informations quantitatives sur ces ensembles

Chapitre 7

Exploration temporelle : graphe d'antériorité

Sommaire

Dans un SIG dédié à l'archéologie, l'analyse exploratoire de données cherche à dégager des relations et des corrélations afin d'extraire de nouvelles connaissances entre les objets d'une base de données spatiotemporelles. Dans ce but, le sujet de ce chapitre est l'étude du positionnement temporel de chaque objet dans la base de données.

Le positionnement temporel des objets fournit de nouvelles connaissances sur l'information portée par ces objets. Ces nouvelles connaissances facilitent la compréhension des données archéologiques et donc l'expertise.

Dans ce chapitre, je positionnerai temporellement, dans une BDG dédiée à l'archéologie, chaque objet archéologique par rapport aux autres objets, dont les périodes d'activité sont représentées par des nombre flous.

Comme les comparaisons de deux nombres flous sont souvent non transitives (voir [Wang *et al.*, 1995]), celles-ci ne sont pas directement utilisées pour le classement de

nombres flous dans l'ensemble les comportant. Cependant, une alternative consiste à positionner chaque nombre flou individuellement dans l'ensemble. Celle-ci a conduit à la définition de nombreuses techniques permettant de positionner un nombre flou relativement à un ensemble de nombres flous et non vis-à-vis de chaque nombre.

Parmi ces méthodes, on retrouve l'approche proposée par Kerre [66] en 1982 et exposée dans 4.1.3. Pour Kerre, après avoir déterminé le maximum de l'ensemble selon le principe d'extension de Zadeh, le positionnement d'un élément correspond à sa distance au maximum des éléments. Une autre approche est celle de Jain qui considère les nombres selon leur intersection à un ensemble flou appelé « maximisant » construit sur l'union des supports des nombres flous à positionner [Jain, 1977]. Dans la démarche proposée dans ce chapitre, je propose de regarder le positionnement de chaque objet de la base de données via sa capacité à être antérieur ou/et postérieur à chaque objet de la base.

Pour cela, j'utilise l'indice d'antériorité proposé dans le chapitre 4. En utilisant cet indice sur un ensemble d'objets archéologiques dont les périodes d'activité sont représentées par des nombres flous, je propose de définir un graphe orienté pondéré dont les sommets représentent les tronçons de rues et dont les arcs ont pour coût l'indice d'antériorité de l'origine de l'arc par rapport à la destination.

La capacité d'antériorité d'un objet est déterminée par la somme des coûts des arcs sortants du sommet associé à l'objet. La capacité de postériorité d'un objet est défini par la somme des coûts des arcs entrants dans le sommet. L'indice temporel d'un objet est le différentiel entre sa capacité de postériorité et sa capacité d'antériorité.

Le positionnement temporel de chaque objet dans un ensemble correspond au rang obtenu par la valeur de son indice temporel dans l'ensemble des valeurs de l'indice temporel de tous les objets. Ce graphe permet donc une représentation synthétique et formelle des structures temporelles de l'échantillon.

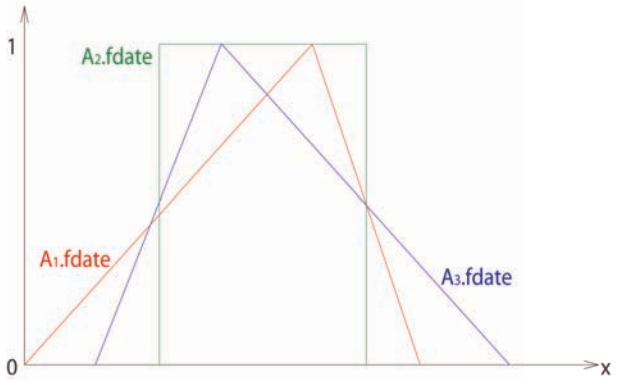
Par la construction de ce graphe sur les objets de *BDFRues*, je propose donc de déterminer le positionnement temporel de chaque objet. À partir de ces positions temporelles, j'extrais de *BDFRues* l'objet le plus antérieur ou l'objet le plus postérieur.

Afin d'illustrer ce chapitre, le cas de trois objets spatiotemporels A_1, A_2 et A_3, dont les composantes temporelles sont respectivement représentées par les nombres flous $A_1.fDate$, $A_2.fDate$ et $A_3.fDate$, dont les fonctions d'appartenance sont présentées figure 7.1, sera considéré.

Ainsi, après avoir présenté la démarche classique pour le positionnement d'un nombre flou dans son ensemble, j'étudierai via la construction du graphe, la position temporelle de chaque objet par l'intermédiaire de sa capacité d'antériorité et de sa capacité de postériorité dans l'ensemble des objets. Je terminerai ce chapitre par l'application de ma démarche exploratoire sur les données archéologiques contenues dans *BDFRues*.

66. Voir [Kerre, 1982]

Figure 7.1 – Fonctions d'appartenance $A_1.fdate$, $A_2.fdate$ et $A_3.fdate$ de respectivement $A_1.fDate$, $A_2.fDate$ et $A_3.fDate$



7.1 Rangement de nombres flous dans un ensemble

Considérons un ensemble Ω de n nombres flous $\{F_1, \ldots, F_n\}$. Les méthodes usuelles de classement les rangent par comparaison des valeurs associées à chacun d'entre eux et calculées selon un indice. Celles-ci sont obtenues à l'aide d'une ou plusieurs valeurs de référence (val_{ref}) définies sur l'ensemble des nombres flous [Wang et Kerre, 2001a]. Dans Ω, la position d'un nombre flou F_i correspond au rang que confère la valeur de l'indice pour F_i au regard des valeurs de l'indice pour tous les autres nombres flous. L'algorithme 7.1 illustre la structure classique d'une méthode de rangement de nombres flous dans un ensemble.

Algorithme 7.1 : Approche classique du rangement d'un ensemble Ω de n nombres flous $\{F_1, \ldots, F_n\}$ selon la méthode m

calcul des valeurs de références $\{val_{ref}\}$ sur Ω ;
pour chaque i ***allant de*** 1 ***à*** n **faire**
  $Indice(F_i) = m(F_i, \{val_{ref}\})$;
fin
tri des nombres flous selon les valeurs de l'indice ;

Classiquement on retrouve deux grands types de méthodes de rangement. Dans le premier type, les techniques (*c.f.* [Jain, 1977] et ses dérivées) se basent sur un ensemble flou de référence, appelé ensemble flou maximisant Ω. Dans le second type (*c.f.* [Kerre, 1982] et ses dérivées), les méthodes utilisent l'ensemble flou maximum sur Ω, au sens de l'extension de Zadeh.

7.1.1 Approche de Kerre et ses dérivées

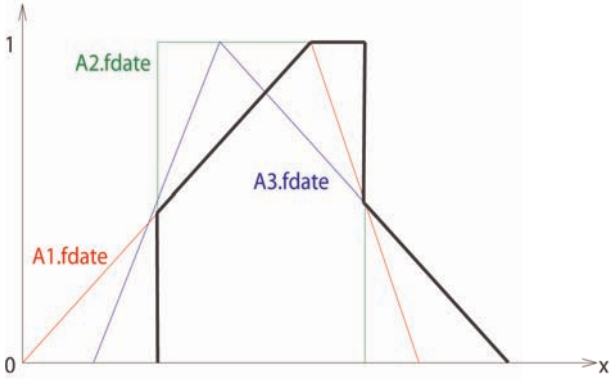
L'indice de Kerre d'un nombre flou correspond à sa distance de Hamming au maximum de l'ensemble des nombres flous selon le principe d'extension de Zadeh [Zadeh, 1965].

L'indice de Kerre d'un nombre F_i dans Ω est noté $K_\Omega(F_i)$ (voir chapitre 5). Ensuite, les nombres sont rangés de la manière suivante : pour deux nombres F_i et F_j de Ω, F_i a un plus grand rang que F_j si et seulement si l'indice de Kerre de F_i est plus petit que celui de F_j. Contrairement au contexte du chapitre 5, l'objet n'est plus l'interrogation des données par la comparaison deux à deux des nombres flous mais leur rangement dans l'ensemble de référence.

Depuis, différentes approches se sont basées sur l'évaluation de la proximité de chaque nombre flou F_i au maximum ou au minimum de Ω. Ainsi, par exemple, en 1987, Wang (voir dans [Wang et Kerre, 2001a]) proposa d'utiliser d'autres méthodes de quantification de la proximité au maximum.

Pour les objets A_1, A_2 et A_3 dont les représentations des périodes d'activité sont illustrées dans la Figure 7.1, Ω est composé de $A_1.fDate$, $A_2.fDate$ et $A_3.fDate$. La fonction d'appartenance du maximum sur Ω selon le principe d'extension de Zadeh est représentée Figure 7.2.

Figure 7.2 – $\widetilde{max}(A_1.fDate, A_2.fDate, A_3.fDate)$



Les valeurs de l'indice de Kerre pour les représentations des périodes d'activité des objets A_1, A_2 et A_3 et pour $\Omega = \{A_1.fDate, A_2.fDate, A_3.fDate\}$ sont :

$$K_\Omega(A_1.fDate) = 11.1, \quad K_\Omega(A_2.fDate) = 13.1, \quad K_\Omega(A_3.fDate) = 10.7.$$

Ainsi $K_\Omega(A_1.fDate) < K_\Omega(A_2.fDate)$ et $K_\Omega(A_1.fDate) > K_\Omega(A_3.fDate)$. À partir de ces comparaisons, le rangement, selon Kerre, des objets par l'intermédiaire de leurs périodes d'activité est, dans l'ordre croissant des rangs $A2.fDate$, $A1.fDate$, $A3.fDate$.

7.1.2 Approche de Jain et ses dérivées

Dans l'approche de Jain [Jain, 1977], considérant une valeur $k > 0$ donnée, on définit un ensemble flou F^k_{max} maximisant Ω dont la fonction d'appartenance f^k_{max} est :

$$f^k_{max}(x) = \left(\frac{x}{x_{max}}\right)^k,$$

avec $k \in \mathbb{R}^{+*}$, $x \in \bigcup_{i=1}^{n} Support(F_i)$, $x \geq 0$, $x_{max} = \sup(\bigcup_{i=1}^{n} Support(F_i))$.

Cela est valable si : $\min(\bigcup_{i=1}^{n} Support(F_i)) \geq 0$. Si ce n'est pas le cas on procède à un changement de variable.

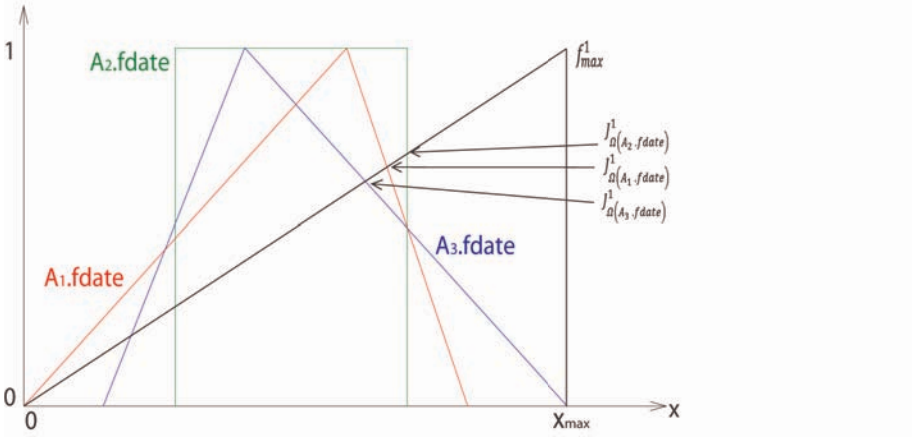
L'indice de Jain d'un nombre flou F_i ($J^k_\Omega(F_i)$) est alors la hauteur de la *t-norme* proposée par Zadeh entre F_i et F^k_{max} (voir 2.3.3) :

$$J^k_\Omega(F_i) = Hauteur(F_i \wedge F_{max})$$

Soient F_i et F_j appartenant à Ω, si $J^k_\Omega(F_i)$ est plus grand que $J^k_\Omega(F_j)$ alors F_i aura un plus haut rang que F_j.

Pour les objets A_1, A_2 et A_3 dont les périodes d'activité sont illustrées dans la Figure 7.1, Ω est composé de $A_1.fDate$, $A_2.fDate$ et $A_3.fDate$. La fonction d'appartenance f^1_{max} de l'ensemble flou F^1_{max} maximisant Ω pour $k = 1$ est illustrée dans la figure 7.3.

Figure 7.3 – f^1_{max} pour $\{A_1.fdate, A_2.fdate, A_3.fdate\}$ avec $k = 1$



Les valeurs de l'indice de Jain pour les représentations des périodes d'activité des objets A_1, A_2 et A_3, pour $\Omega = \{A_1, A_2, A_3\}$ et pour $k = 1$ sont :

$$J^1_\Omega(A_1.fDate) = 0.67, \quad J^1_\Omega(A_2.fDate) = 0.71, \quad J^1_\Omega(A_3.fDate) = 0.63.$$

Donc $J^1_\Omega(A_3.fDate) < J^1_\Omega(A_1.fDate) < J^1_\Omega(A_2.fDate)$. Le rangement selon Jain pour

$k = 1$ sera donc dans l'ordre croissant des rangs $A_3.fDate$, $A_1.fDate$, et $A_2.fDate$. Les objets sont donc positionnés dans l'ordre croissant de la manière suivante : A_3, A_1 et A_2.

L'approche de [Chen, 1985] complète celle de Jain par l'ajout d'une valeur de référence : l'ensemble flou minimisant Ω. A l'aide des ensembles flous maximisant et minimisant Ω, il détermine une utilité gauche et une utilité droite d'un nombre flou sur lesquelles il base le calcul de l'indice pour le nombre flou considéré. Liou et Wang se sont basés sur la démarche de Chen pour proposer une autre méthode de rangement de nombres flous [Liou et Wang, 1992].

Comme on peut le voir, dans ces méthodes, le rangement dans l'ensemble ne se fait pas par comparaison deux à deux des nombres flous mais par le calcul d'un indice reposant sur la définition de valeurs de référence sur l'ensemble des nombres à comparer. Une des causes à cela est que les méthodes de comparaison deux à deux de nombres flous sont le plus souvent non transitives [Wang *et al.*, 1995].

Dans l'approche proposée ci-après, l'idée est d'utiliser l'indice d'antériorité *Ant* entre deux nombres flous (voir chapitre 4) plutôt qu'une méthode de comparaison deux à deux donnant une décision binaire. Ainsi, au regard des objets archéologiques, dont les périodes d'activité sont représentées par un nombre flou, la quantification de l'antériorité permet d'étudier le positionnement temporel de chaque objet à l'aide de la construction d'un graphe orienté pondéré (le graphe d'antériorité). Ce graphe permet l'exploration schématique des données selon l'information temporelle.

7.2 Graphe d'antériorité

L'analyse exploratoire proposée dans ce chapitre se base sur la construction d'un graphe orienté pondéré, appelé graphe d'antériorité, à partir de l'indice d'antériorité *Ant* (*cf.* 4.2). Ce dernier quantifie une relation binaire, définie sur les nombres flous et interprétée comme antériorité, entre deux nombres flous quelconques. L'ensemble des nombres flous à comparer, la relation d'antériorité et les valeurs de l'indice d'antériorité fournissent les sommets, les arcs et les coûts du graphe d'antériorité.

La construction du graphe d'antériorité ayant pour objectif l'observation du positionnement temporel de chaque objet d'une base de données archéologiques par rapport aux autres, les nombres flous mis en relation sont les représentations des périodes d'activité des objets archéologiques. Ainsi on apparie temporellement les objets archéologiques les uns aux autres et on quantifie les paires obtenues à l'aide d'une représentation schématique.

7.2.1 Construction du graphe

Considérons un ensemble d'éléments E, une relation binaire $\mathcal{R}$ sur E (*i.e.* un sous-ensemble de $E \times E$) et une application *App* sur cette relation binaire prenant valeur dans $\mathbb{R}$.

Un graphe orienté pondéré $G_{\mathcal{R}}(L_S, L_A, L_C)$ est une représentation schématique constituée d'un ensemble L_S de sommets, d'un ensemble L_A d'arcs reliant les sommets deux à deux, et d'un ensemble L_C de coûts associés aux arcs.

Soit deux éléments A et B de E reliés par la relation binaire $\mathcal{R}$, on associe, dans $G_{\mathcal{R}}(L_S, L_A, L_C)$, à A le sommet S_A et à B le sommet S_B :

$$A \in E \Leftrightarrow S_A \in L_S.$$

L'arc (S_A, S_B) représente alors le fait que $A\mathcal{R}B$:

$$A\mathcal{R}B \Leftrightarrow (S_A, S_B) \in L_A.$$

Le coût $C(S_A, S_B)$ d'un arc (S_A, S_B) est égal à $App(A, B)$:

$$(S_a, S_b) \in L_A \Leftrightarrow (C(S_A, S_B) = App(A, B) \text{ et } C(S_A, S_B) \in L_C)\,.$$

L'analyse exploratoire de cette partie se base sur la construction d'un graphe orienté pondéré à partir de l'indice d'antériorité et de l'ensemble Ω d'objets d'une base de données archéologiques dont les périodes d'activité sont définies et représentées par des nombres flous. Ce graphe sera noté $G_{Ant}(L_S, L_A, L_C)$ et sera appelé graphe d'antériorité.

La construction de ce graphe suit la démarche suivante. La relation $\mathcal{R}$ est une relation qui relie deux à deux tous les objets de Ω :

$$\forall A_i, A_j \in \Omega,\ A_i\mathcal{R}A_j \text{ et } A_j\mathcal{R}A_i.$$

À chaque objet A_i de Ω de période d'activité $A_i.fDate$, j'associe un sommet S_{A_i} du graphe $G_{Ant}(L_S, L_A, L_C)$. Ainsi, le cardinal de L_S est égal à celui de Ω. Soit deux objets A_i et A_j de Ω représentés respectivement par les sommets S_{A_i} et S_{A_j} du $G_{Ant}(L_S, L_A, L_C)$, l'arc (S_{A_i}, S_{A_i}) représente l'antériorité possible de A_i à A_j au regard de leurs périodes d'activité. Le coût $C(S_{A_i}, S_{A_j})$ de l'arc (S_{A_i}, S_{A_j}) est égal à $Ant(A_i.fDate, A_j.fDate)$ et représente la quantification de l'antériorité de A_i à A_j.

Ainsi, soit Ω un ensemble de n objets $\{A_1, \ldots, A_n\}$, le graphe d'antériorité $G_{Ant}(L_S, L_A, L_C)$ est donc construit comme le suggère l'algorithme 7.2.

Pour l'exemple des objets dont les périodes d'activité sont représentées dans la figure 7.1, $\Omega = \{A_1, A_2, A_3\}$ et les valeurs de l'indice d'antériorité pour l'ensemble des couples de Ω sont :

$$\begin{array}{cc} Ant(A_1.fDate, A_2.fDate) = 0.44, & Ant(A_2.fDate, A_1.fDate) = 0.56, \\ Ant(A_1.fDate, A_3.fDate) = 0.51, & Ant(A_3.fDate, A_1.fDate) = 0.49, \\ Ant(A_2.fDate, A_3.fDate) = 0.5, & Ant(A_3.fDate, A_2.fDate) = 0.5. \end{array}$$

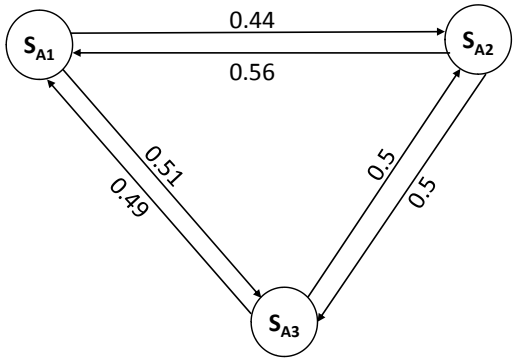
Le graphe d'antériorité $G_{Ant}(L_S, L_A, L_C)$ est alors représenté dans la figure 7.4, dans

Algorithme 7.2 : Construction du graphe d'antériorité pour un ensemble Ω de n objets archéologiques $\{A_1, \ldots, A_n\}$

$L_S \leftarrow \{\}$;
$L_A \leftarrow \{\}$;
$L_C \leftarrow \{\}$;
pour chaque i ***allant de*** 1 ***à*** n **faire**
  Association de S_{A_i} à A_i ;
  $L_S \leftarrow L_S \cup \{S_{A_i}\}$;
  pour chaque j ***allant de*** 1 ***à*** $i-1$ **faire**
    $L_A \leftarrow L_A \cup \{(S_{A_i}, S_{A_j}), (S_{A_j}, S_{A_i})\}$;
    $C(S_{A_i}, S_{A_j}) \leftarrow Ant(A_i.fDate, A_j.fDate)$;
    $C(S_{A_j}, S_{A_i}) \leftarrow Ant(A_j.fDate, A_i.fDate)$;
    $L_C \leftarrow L_C \cup \{C(S_{A_i}, S_{A_j}), C(S_{A_j}, S_{A_i})\}$;
  fin
fin

lequel un sommet S_{A_i} correspond à l'objet A_i ayant pour période d'activité $A_i.fDate$.

Figure 7.4 – Graphe d'antériorité pour $\Omega = \{A_1, A_2, A_3\}$



Pour chaque sommet du graphe $G_{Ant}(L_S, L_A, L_C)$ ainsi construit, la somme des coûts des arcs sortants, celle des coûts des arcs entrants et leur différence sont des valeurs particulières et ont une signification importante.

7.2.2 Capacité d'antériorité, capacité de postériorité et positionnement d'un objet

Soient S_{A_i} et S_{A_j} deux sommets de $G_{Ant}(L_S, L_A, L_C)$ correspondant respectivement aux objets A_i et A_j, le coût $C(S_{A_i}, S_{A_j})$ correspond à la valeur de l'indice d'antériorité $Ant(A_i.fDate, A_j.fDate)$. Ainsi, la somme des coûts des arcs sortants de S_{A_i} correspond

à la capacité d'antériorité $CapAnt_\Omega(A_i)$ de A_i relativement à l'ensemble des nombres flous Ω :

$$CapAnt(A_i) = \sum_{\substack{S_{A_j} \in L_S \\ (S_{A_i}, S_{A_j}) \in L_A}} C(S_{A_i}, S_{A_j}).$$

Pour l'exemple où $\Omega = \{A_1, A_2, A_3\}$, d'après le graphe d'antériorité présenté Figure 7.4, les capacités d'antériorité des objets de Ω sont les suivants :

$$CapAnt_\Omega(A_1) = 0.95, \quad CapAnt_\Omega(A_2) = 1.06, \quad CapAnt_\Omega(A_3) = 0.99.$$

L'indice $Ant(A_j.fDate, A_i.fDate)$ entre deux nombres flous $A_j.fDate$ et $A_i.fDate$ a été présenté comme l'indice d'antériorité de A_j à A_i. On peut aussi le voir comme l'indice de postériorité de A_i à A_j. Ainsi, la somme des coûts des arcs entrants sur le sommet S_{A_i} de $G_{Ant}(L_S, L_A, L_C)$ associé à l'objet A_i de Ω peut être assimilée à la capacité de postériorité $CapPost_\Omega$ de A_j relativement à l'ensemble des nombres flous Ω :

$$CapPost(A_i) = \sum_{\substack{S_{A_j} \in L_S \\ (S_{A_j}, S_{A_i}) \in L_A}} C(S_{A_j}, S_{A_i}).$$

Pour l'exemple où $\Omega = \{A_1, A_2, A_3\}$, d'après le graphe présenté Figure 7.4, les capacités de postériorité des objets de Ω sont les suivants :

$$CapPost_\Omega(A_1) = 1.05, \quad CapPost_\Omega(A_2) = 0.94, \quad CapPost_\Omega(A_3) = 1.01.$$

L'indice temporel d'un objet A_i doit permettre de définir sa position temporelle. Son calcul doit donc prendre en considération à la fois sa capacité de postériorité et sa capacité d'antériorité. C'est pourquoi, l'indice temporel $IndTemp_\Omega(A_i)$ de A_i est déterminé comme suit :

$$IndTemp_\Omega(A_i) = \sum_{\substack{S_{A_j} \in L_S \\ (S_{A_j}, S_{A_i}) \in L_A}} C(S_{A_j}, S_{A_i}) - \sum_{\substack{S_{A_j} \in L_S \\ (S_{A_i}, S_{A_j}) \in L_A}} C(S_{A_i}, S_{A_j}).$$

Les valeurs de l'indice temporel des objets A_1, A_2 et A_3, dont les fonctions d'appartenance des périodes d'activité sont représentées dans la figure 7.1 et pour le graphe d'antériorité présenté dans la figure 7.4 (avec $\Omega = A_1, A_2, A_3$), sont :

$$IndTemp_\Omega(A_1) = 0.1, \quad IndTemp_\Omega(A_2) = -0.12, \quad IndTemp_\Omega(A_3) = 0.02.$$

À l'aide de cet indice, je détermine la position temporelle $PosTemp(A_i)$ de A_i et A_j dans l'ensemble Ω en suivant le principe suivant :

Si $IndTemp_\Omega(A_i) > IndTemp_\Omega(A_j)$, alors $PosTemp(A_i) > PosTemp(A_j)$.

La position temporelle de A_i correspond au rang de A_i dans la liste des objets de Ω ordonnée selon les valeurs de l'indice temporel.

Pour l'exemple de la figure 7.1, avec $\Omega = \{A_1, A_2, A_3\}$, d'après le graphe présenté figure 7.4, les positions temporelles des objets de Ω sont dans l'ordre croissant : A_2, A_3, A_1. Cet ordre n'est donc pas le même que celui obtenu par l'approche de Kerre ni le même que celui issu de la démarche de Jain. Toutefois, grâce aux valeurs de l'indice d'antériorité, la méthode proposée définit un indice de position temporelle d'un individu vis-à-vis des autres dans un ensemble. On obtient une vision pondérée de la structure temporelle de l'ensemble.

Les valeurs de l'indice temporel et les positions temporelles des objets donnent des informations importantes sur les relations temporelles entre objets.

7.2.3 Analyse des objets selon le graphe d'antériorité

À l'aide du graphe d'antériorité, on peut extraire trois objets particuliers : l'objet le plus antérieur, l'objet le plus postérieur et l'objet temporellement médian.

L'objet le plus vieux dans le cadre applicatif, c'est-à dire le plus antérieur, noté PA, est celui dont la valeur de l'indice temporel est la plus petite dans l'ensemble des valeurs de l'indice temporel des objets de Ω. La position temporelle de l'objet le plus antérieur est la position temporelle minimale des objets de Ω :

$$PosTemp(PA) = \min_{A_i \in \Omega} (PosTemp(A_i))$$

L'objet le plus récent dans le cadre applicatif, c'est-à dire le plus postérieur, noté PP, est celui dont la valeur de l'indice temporel est la plus grande dans l'ensemble des valeurs de l'indice temporel des objets de Ω. La position temporelle de l'objet le plus antérieur est la position temporelle maximale des objets de Ω :

$$PosTemp(PP) = \max_{A_i \in \Omega} (PosTemp(A_i))$$

Pour les objets de la figure 7.1, l'objet A_2 est le plus antérieur.

Grâce à l'approche par rang (position temporelle), il est trivial de définir l'objet temporellement médian, noté TM. Il est celui dont la valeur de l'indice temporel est médiane à l'ensemble des valeurs de l'indice temporel pour les objets de Ω.

Pour les objets de l'exemple, on a :

$$PA = A_2, \quad PP = A_1, \quad TM = A_3.$$

De plus, on peut considérer qu'un objet ayant un indice temporel négatif peut être considéré, par rapport à l'ensemble des objets, comme un objet « plutôt antérieur » tandis qu'un objet ayant un indice temporel positif correspond à un objet « plutôt postérieur ».

Ceux ayant un indice d'antériorité nul ont une capacité d'antériorité égal à celle de postériorité. Ils occupent une position intermédiaire spécifique dans l'ensemble : je les nomme « anté-postérieurs ». Je propose ainsi une classification des objets selon qu'ils soient « plutôt antérieurs », « plutôt postérieurs » ou « anté-postérieurs ».

Dans l'exemple de la figure 7.1, $\Omega = \{A_1, A_2, A_3\}$, il n'y a pas d'élément anté-postérieur, A_2 est « plutôt antérieur » dans Ω et $\{A_1, A_3\}$ forme l'ensemble des objets « plutôt postérieurs » dans Ω.

La construction du graphe d'antériorité est une approche originale pour le rangement d'objets archéologiques dont les périodes d'activité sont imparfaites et sont représentées par des nombres flous. Ce graphe donne une vision globale des relations chronologiques entre les objets archéologiques. Il donne de nombreuses indications en vue de la classification des objets archéologiques, de l'analyse à l'échelle locale (le chantier de fouille) et globale (la ville) et donc de la généralisation. Les périodes d'activité des objets de *BDFRues* étant représentées par des nombres flous, ces objets font dans la suite l'objet d'une analyse exploratoire dégageant les positions temporelles des objets.

7.3 Application aux données archéologiques

Le processus exploratoire proposé dans ce chapitre est appliqué sur les données de *BDFRues*. Afin d'obtenir une grille de lecture des résultats plus globale, je présenterai dans un premier temps les rangs des objets obtenus par le classement des représentations de leurs périodes d'activité selon les valeurs des indices de Jain et de Kerre. Ensuite, je présenterai les résultats issus de la construction du graphe d'antériorité.

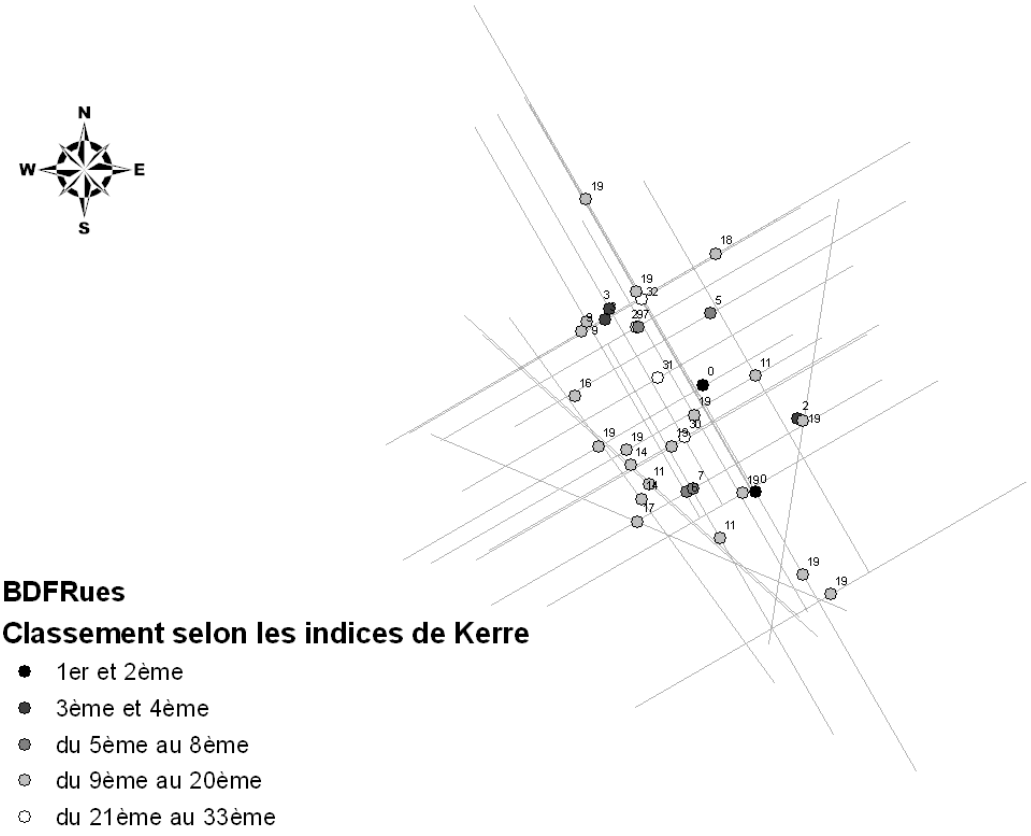
Remarquons en préliminaire que comme les données l'ensemble de la période Gallo-Romaine, les fonds de carte aux différents siècles ne sont pas représentés dans les figures de cette partie.

7.3.1 Rangs des objets selon leurs périodes d'activités pour Kerre et Jain

En utilisant Kerre pour classer les représentations des périodes d'activité des objets de *BDFRues* on obtient la figure 7.5. Dans cette figure, relativement à l'ensemble des nombres flous associés aux périodes d'activité des objets de *BDFRues*, plus le rang d'un objet est grand plus le nombre flou associé à la période d'activité de l'objet est grand selon [Kerre, 1982].

En utilisant l'indice de Jain avec $k = 1$ pour classer les nombres flous associés aux périodes d'activité des objets de *BDFRues* on obtient la figure 7.6. Dans cette figure plus le rang d'un objet est grand plus le nombre flou associé à sa période d'activité est grand pour [Jain, 1977].

Figure 7.5 – Rangs, déterminés à l'aide de Kerre, des 33 objets de *BDFRues* selon leurs périodes d'activité



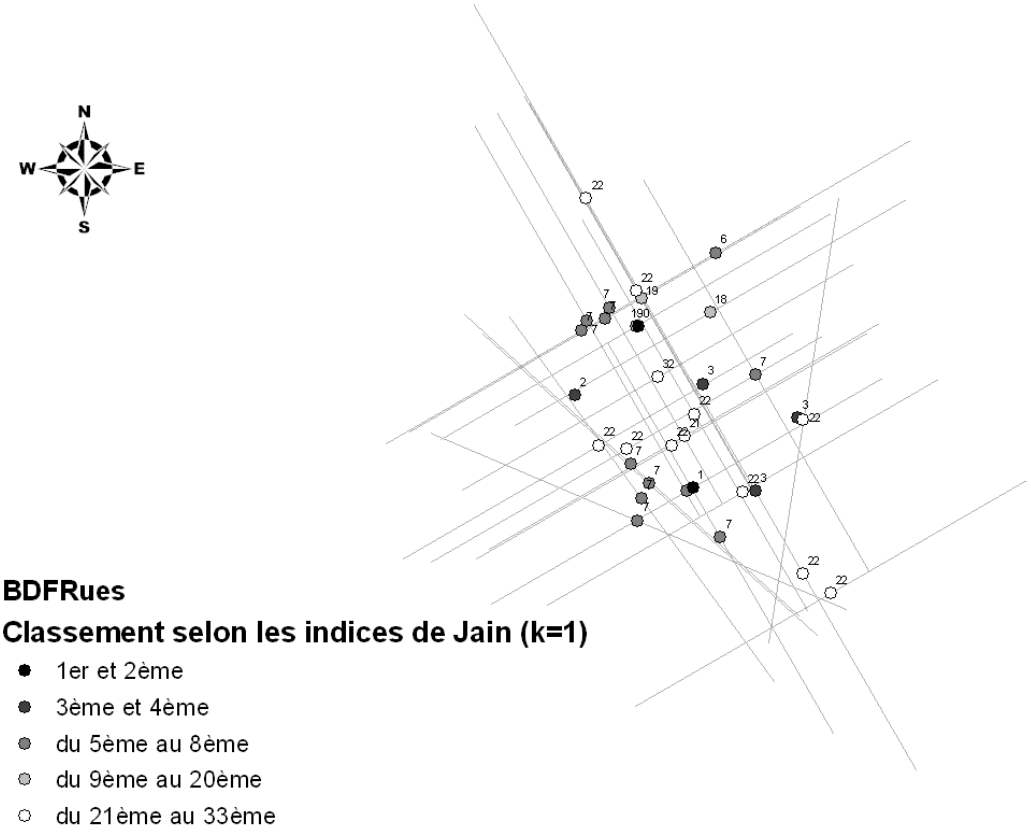
On peut remarquer que certains rangs attribués aux objets selon l'indice de Jain diffèrent de ceux attribués selon l'indice de Kerre. L'interprétation de ces rangs est difficile, car il n'y a pas de sémantique associée à ces indices. A contrario, les positions temporelles obtenues à l'aide de la construction du graphe d'antériorité ont une forte interprétabilité qui sera illustrée dans la suite.

7.3.2 Positionnement temporel des objets selon le graphe d'antériorité

En construisant le graphe d'antériorité sur les données de *BDFRues* on obtient la figure 7.7. Dans cette figure, plus le rang d'un objet est grand, plus la position temporelle est élevée, c'est-à-dire plus sa capacité de postériorité à tous les autres est grande.

On peut remarquer que les positions temporelles obtenues à l'aide du graphe d'antériorité, et les rangs obtenus à l'aide des approches de Kerre ou de Jain présentent dans certains cas des divergences. Cependant, les positions temporelles ont une interprétabilité plus forte que les rangs issus de Kerre ou de Jain, ce qui permet de définir une classification

Figure 7.6 – Rangs, déterminés à l'aide de Jain ($k = 1$), des 33 objets de *BDFRues* selon leurs périodes d'activité



temporelle des objets selon leur positionnement temporel.

La position temporelle d'un objet archéologique est obtenue à l'aide de l'indice temporel de l'objet lui-même issu des capacités d'antériorité et de postériorité. Ces capacités de l'objet sont calculées à l'aide de l'indice d'antériorité de l'objet à chacun des autres objets archéologiques. Cet indice quantifie l'antériorité d'un objet à un autre. Ainsi, la position temporelle d'un objet reflète la relation temporelle de l'objet à l'ensemble des autres.

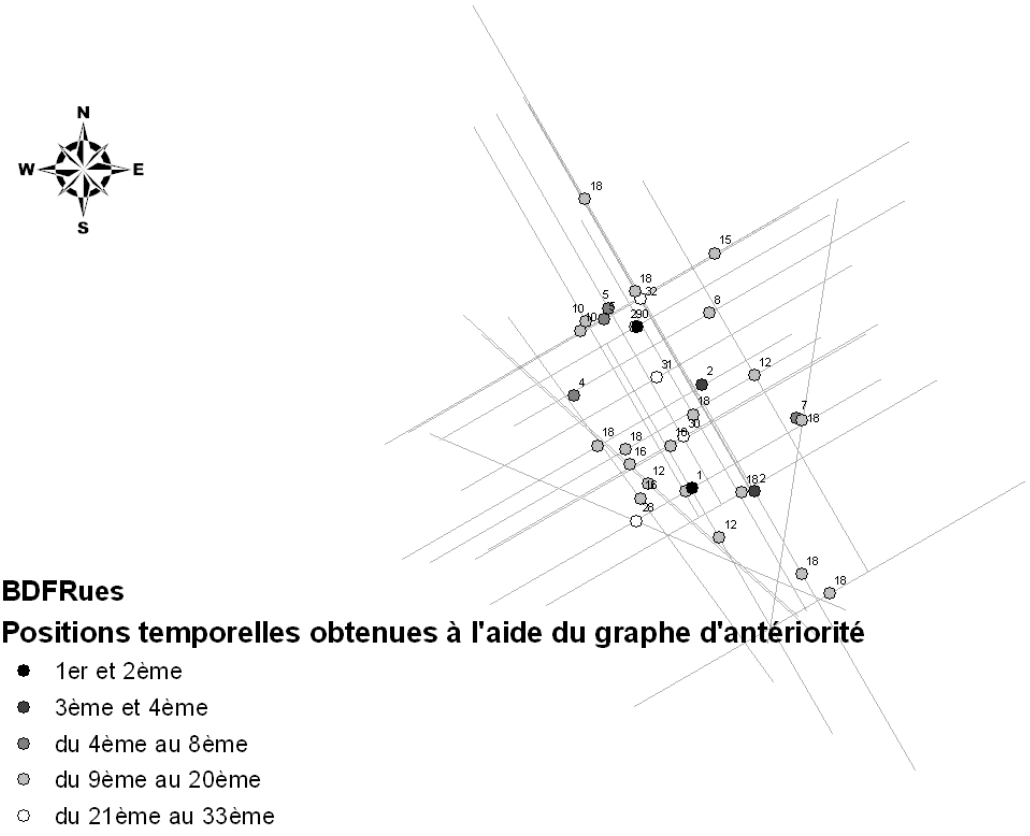
Par exemple, de ces positions temporelles, on peut extraire les trois objets particuliers suivants : le plus antérieur, le plus postérieur et le médian. Ces objets particuliers sont identifiables dans la figure 7.8.

La figure 7.9 présente les objets de *BDFRues* selon les trois classes : « plutôt postérieurs », « plutôt antérieurs » et « anté-postérieurs ».

Pour les données de *BDFRues*, la classe des objets « plutôt postérieurs » regroupe deux fois plus d'objets que celle des objets « plutôt antérieurs ». La classe des objets « anté-postérieurs » est vide.

On peut supposer que les objets « plutôt antérieurs » ont des valeurs de l'indice d'an-

Figure 7.7 – Positions temporelles des 33 objets de *BDFRues* selon leurs périodes d'activité



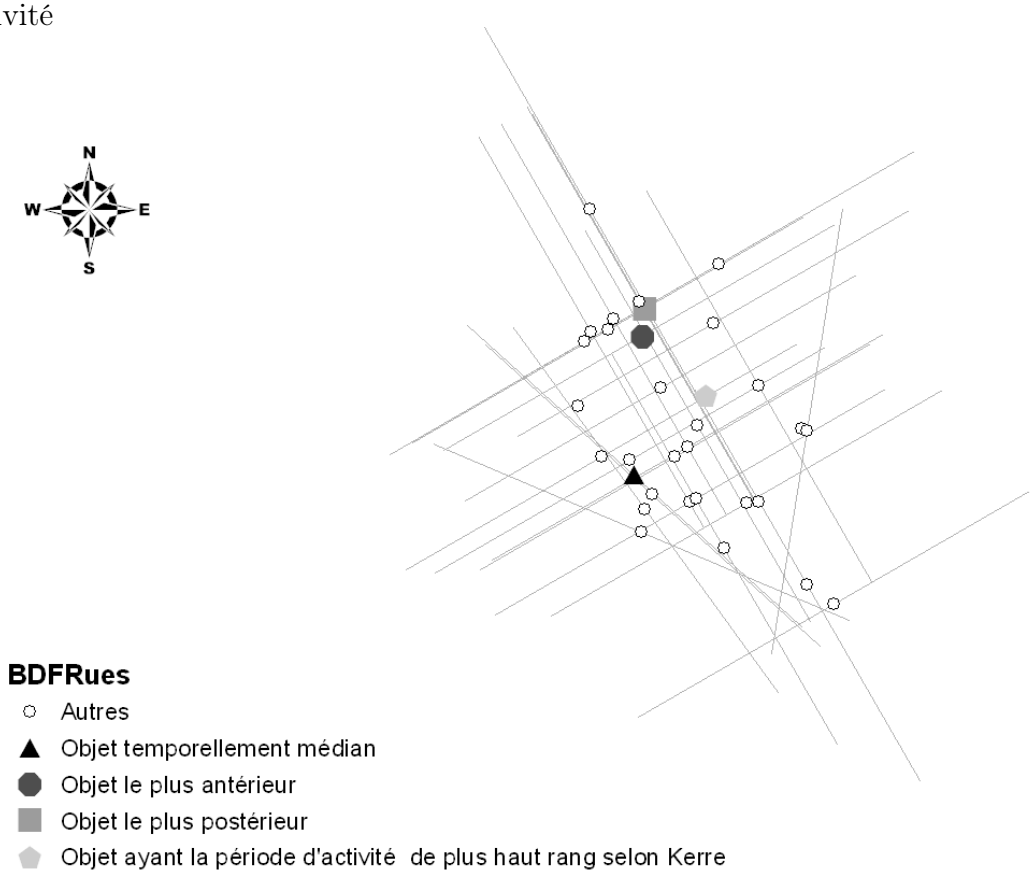
tériorité avec les autres objets proches de 1. Cela expliquerait le fort différentiel entre le cardinal de la classe des objets « plutôt postérieurs » avec celui de la classe des objets « plutôt antérieurs ». Cette hypothèse est d'ailleurs vérifiée puisque, sachant qu'il y a trente trois objets :

- la valeur de l'indice temporel de l'élément le plus antérieur est de -29.8 ;
- la plus grande valeur d'indice temporel des éléments « plutôt antérieurs » est -5.4 ;
- plus de 72% des objets « plutôt postérieurs » ont une valeur d'indice temporel supérieure à 6.

La bipolarisation des valeurs est donc pertinente. Cependant, on peut remarquer que trois objets ont une valeur de l'indice temporel supérieur à 0 mais proche de 0 . Une définition plus vague de la notion *anté-postérieure* aurait sûrement permis de détecter ces éléments. Mais cette nouvelle définition demande un paramétrage fortement dépendant des données. Ce paramétrage n'entre pas dans la démarche exploratoire proposée dans cette partie et donc dans ce chapitre.

De plus, par le recoupement de la figure 7.9 avec la figure 7.8, on remarque que l'objet temporellement médian appartient aux objets « plutôt postérieurs » à l'instar de l'objet

Figure 7.8 – Positions temporelles particulières des objets de *BDFRues* selon leurs périodes d'activité



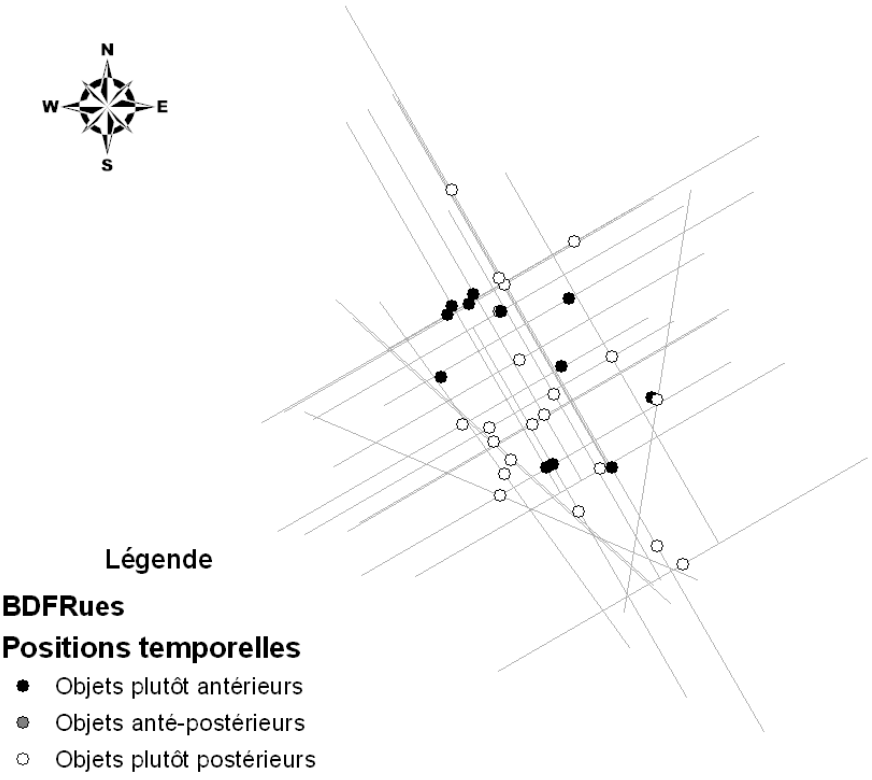
le plus postérieur. L'objet le plus antérieur appartient à la classe des « plutôt antérieurs ».

Ainsi, les valeurs de l'indice temporel des objets de *BDFRues* obtenues à partir du graphe d'antériorité permettent de classer les objets archéologiques selon l'antériorité aux objets de la base. Ces positions permettent de définir des classes d'objets et d'extraire des objets particuliers. Ce graphe permet donc, par une organisation schématique structurée, de dégager de nouvelles connaissances temporelles sur les informations contenues dans *BDFRues*.

◇◇◇◇◇

Dans ce chapitre, j'ai proposé un processus exploratoire ayant pour but l'analyse des relations entre objets selon la temporalité de l'information archéologique disponible. Pour cela, j'ai construit un graphe orienté pondéré à partir des objets de la base et des valeurs

Figure 7.9 – Objets de *BDFRues* classés en « plutôt postérieurs », « plutôt antérieurs » et « anté-postérieurs »



de l'indice d'antériorité les reliant. De ce graphe, j'ai dégagé l'indice temporel de chaque objet afin d'en déterminer sa position temporelle dans l'ensemble et l'analyse associée (« plutôt antérieur »,« plutôt postérieur », « anté-postérieur »). Les positions, relativement au temps, des objets archéologiques dans une base de données facilitent la compréhension des relations liant les objets de la base dans le SIG.

Dans leurs expertises, les archéologues évaluent les objets qu'ils sont susceptibles de trouver dans un site tant d'un point de vue fonctionnel que temporel. L'étude des relations à l'échelle de la ville entre objets stockés dans les bases de données permet de vérifier si la logique temporelle est respectée et d'estimer une évolution temporelle de la cité. Dans cet objectif, le graphe d'antériorité peut être utilisé.

Toutefois, l'exploration temporelle ne révèle que le positionnement temporel des objets. Or les objets archéologiques ont de par leurs spatialisations et leurs formes d'autres relations possibles et envisageables. Dans ce cadre, l'un des buts de l'exploitation des SIG est la recherche d'un ou plusieurs représentants des objets.

Dans le chapitre 8, je propose une méthode qui se base sur la représentativité [67] de chaque objet dans leur ensemble. Elle permet à partir d'un indice de dissimilarité, d'extraire de la base les éléments les plus représentatifs en fonction de l'indice de dissimilarité choisi.

67. Représentativité : notion introduite dans [Blanchard, 2005]

Chapitre 8

Exploration spatiotemporelle : éléments représentatifs

Sommaire

Par l'utilisation d'un SIG archéologique, l'expert aimerait regarder le positionnement et la représentativité de chaque objet présent dans la base de données associée au SIG tant spatialement que temporellement. Ainsi, est-il important de regarder les relations entre des objets archéologiques, et de positionner chaque objet dans l'ensemble associé. Pour cela je propose d'étudier la dissimilarité entre objets non pas dans le but de découvrir les objets dissimilaires mais dans celui d'extraire de l'ensemble les objets les plus représentatifs.

Frédéric Blanchard a défini durant sa thèse ([Blanchard, 2005]) une notion de la représentativité d'une donnée dans un échantillon. La définition de la notion de représentativité d'un élément au sein d'un échantillon de données utilise les statistiques de rangs. Les statistiques non paramétriques (par exemple les statistiques de rangs) connaissent depuis quelques années un regain d'intérêt [David et Nagaraja, 2003]. Leur utilisation en analyse de données permet notamment de s'affranchir de l'hypothèse de normalité et apporte une robustesse vis à vis des données aberrantes (voir [Galambos, 1975] et [Barnett, 1976]).

La définition quantitative de la représentativité nécessite de disposer d'un indice de dissimilarité entre éléments afin de calculer les rangs des données. Cette dissimilarité est liée à la description des données. Or les données de *BDFRues* sont représentées par des ensembles flous convexes et normalisés. C'est pourquoi, je propose d'utiliser une métrique classique entre nombres flous afin de déterminer les dissimilarités temporelles et orientationnelles. Pour les localisations, j'utilise la métrique précédente sur la projection des représentations des localisations (ensembles flous convexes et normalisés définis sur $\mathbb{R} \times \mathbb{R}$) vers des nombres flous (ensembles flous convexes et normalisés définis sur $\mathbb{R}$). La dissimilarité globale est la moyenne de la normalisation des trois précédentes.

Au sein d'un SIG, il est intéressant de distinguer les objets les plus représentatifs (relativement à des critères fixés) de l'ensemble des objets stockés. Par exemple, si l'on recherche quel est le boulevard le plus représentatif des boulevards parisiens sur le plan architectural, le résultat est vraisemblablement un boulevard de type haussmanien. Dans le cadre d'une analyse spatiotemporelle, il s'avère intéressant de déterminer celui qui représente le mieux les éléments découverts en terme de localisation, de période d'activité ou de forme ; cet élément sert de représentant à l'ensemble et le caractérise. Je cherche dans *BDFRues* le tronçon de rue le plus représentatif en termes de localisation, de temps et/ou d'orientation.

À partir de la notion de représentativité, je propose dans ce chapitre de construire une statistique pour déterminer l'élément le plus représentatif d'un échantillon de données : le Vecteur de Meilleur Rang Moyen (*VMRM*). Le *VMRM* extrait l'élément de représentativité maximale d'un ensemble de données. Bien que les statistiques de rangs soient très utilisées notamment en filtrage d'images [Pitas et Tsakalides, 1991], elles ne le sont que peu dans l'exploitation des *SIG* archéologiques pour la classification. J'utilise ensuite cette statistique sur *BDFRues* afin de déterminer les tronçons de rues les plus représentatifs.

Dans ce chapitre, après avoir présenté et illustré la notion de représentativité dans un échantillon d'un élément, j'exposerai le calcul des dissimilarités localisationnelles, temporelles, orientationnelles et globales entre objets archéologiques de *BDFRues*. Je visualiserai dans un SIG les représentativités de chaque objet de *BDFRues* en affectant à chaque objet une couleur selon son rang moyen en fonction de l'indice de dissimilarité. Je m'attacherai enfin à extraire les éléments les plus représentatifs. Pour cela, j'exposerai la définition du Vecteur de Meilleur Rang Moyen et extrairai, dans le contexte des rues de Durocortorum, les tronçons de rues trouvés les plus représentatifs de leur ensemble.

8.1 Représentativité d'un objet : rang moyen

La notion de représentativité, introduite dans [Blanchard, 2005], est associée à des données multidimensionnelles pour lesquelles on dispose d'un indice de dissimilarité. La représentativité d'une donnée multidimensionnelle x est déterminée de la manière suivante : chaque donnée de l'ensemble attribue à x un rang selon un classement par dissi-

milarité croissante ; le rang moyen de x sur l'ensemble détermine sa représentativité. Plus le rang moyen d'une donnée est petit, plus celle-ci est représentative de l'ensemble des données. L'élément rangé en premier sera le plus représentatif.

Plus formellement, considérons un échantillon de données multidimensionnelles Ω ($\Omega = \{x_1, x_2, \ldots, x_n\}$) dans un espace E de dimension $d \in \mathbb{N}$. On suppose que l'on dispose, sur cet échantillon, d'un indice de dissimilarité. Autrement dit, on suppose que l'on dispose d'un moyen de quantifier la dissimilarité entre deux éléments quelconques de l'échantillon Ω. On notera $\delta(x_i, x_j)$ la dissimilarité entre x_i et x_j ($i, j \in [1..n]$). La distance euclidienne est un exemple d'indice de dissimilarité.

8.1.1 Rangs d'une donnée

Considérons maintenant les n classements (i.e. les n tris) obtenus en utilisant les dissimilarités par rapport à chaque x_i. Autrement dit, pour chaque élément x_i, on classe l'ensemble de l'échantillon par ordre de dissimilarité croissante avec x_i. Soit $Rg_{x_i}(x_j)$ le rang de la donnée x_j dans le classement par dissimilarité croissante à x_i. La valeur de $Rg_{x_i}(x_j)$ représente ainsi la position de x_j dans le classement des données les moins dissimilaires à x_i. Par exemple, $Rg_{x_i}(x_j) = k$ ($k \in [1..n]$) signifie que x_j est la k-ième donnée de Ω la moins dissimilaire à x_i, c'est à dire $x_{(k)}$ dans l'ordre induit par la donnée x_i.

On obtient donc, sur l'ensemble de l'échantillon, n classements des données.

8.1.2 Rang moyen d'une donnée

On calcule ensuite, pour chaque donnée x_j de Ω, la moyenne des rangs $Rg_{x_i}(x_j)$ obtenus par la donnée au cours de ces n classements. On note ce rang moyen : $\overline{Rg}(x_j)$ et on a :

$$\overline{Rg}(x_j) = \frac{1}{n} \times \sum_{i=1}^{n} Rg_{x_i}(x_j).$$

Le rang moyen est un critère qui permet alors d'évaluer le potentiel d'une donnée à représenter l'échantillon auquel elle appartient. En effet, cette valeur moyenne traduit la façon dont une donnée est *la moins dissimilaire* à l'ensemble des autres. On appelle cette notion la *représentativité d'une donnée dans son échantillon*. Plus le rang moyen d'une donnée est petit, plus la donnée est représentative de l'échantillon.

8.1.3 Exemple

Soit l'échantillon $\Omega = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ constitué de 6 vecteurs de $\mathbb{R}^4$; les composantes des vecteurs sont c_1, c_2, c_3, c_4. Les valeurs numériques des 6 vecteurs correspondants sont présentées dans le tableau 8.1a. En prenant la distance euclidienne comme indice de dissimilarité entre les données (tableau 8.1b), on obtient les rangs présentés dans le tableau

Tableau 8.1 – Rangs moyens : Exemple sur un échantillon de données

a : Echantillon exemple

	c_1	c_2	c_3	c_4
x_1	10	-15	10	30
x_2	15	-60	20	50
x_3	55	-10	20	30
x_4	15	-25	15	25
x_5	30	-200	500	200
x_6	15	-60	20	75

b : Dissimilarités
(Distance Euclidienne)

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0	80	60	25	865	105
x_2	80	0	110	65	785	25
x_3	60	110	0	65	865	135
x_4	25	65	65	0	850	90
x_5	865	785	865	850	0	760
x_6	105	25	135	90	760	0

c : Rangs

	Rg_{x_1}	Rg_{x_2}	Rg_{x_3}	Rg_{x_4}	Rg_{x_5}	Rg_{x_6}
x_1	1	4	2	2	5	4
x_2	4	1	4	3	3	2
x_3	3	5	1	4	6	5
x_4	2	3	3	1	4	3
x_5	6	6	6	6	1	6
x_6	5	2	5	5	2	1

d : Rangs moyens

	$\overline{Rg}$
x_1	3
x_2	2.83
x_3	4
x_4	2.67
x_5	5.17
x_6	3.33

8.1c. La colonne Rg_{x_i} de ce tableau représente les rangs induits par les dissimilarités avec la donnée x_i. Autrement dit, si l'élément de la $j^{\text{ième}}$ ligne et de la colonne Rg_{x_i} vaut k alors cela signifie que x_j est le $k^{\text{ième}}$ moins dissimilaire vecteur à x_i dans Ω. Les rangs moyens de chaque donnée de l'échantillon sont calculés à partir des rangs dans le tableau 8.1d.

Ainsi, dans cet exemple, la donnée x_1 est moins représentative de Ω que la donnée x_2. Cependant, x_1 est plus représentative de Ω que la donnée x_3.

Comme on le voit, par définition, la représentativité d'un élément dans un ensemble est fonction de l'indice de dissimilarité choisi. Dans la section suivante, je présenterai les indices choisis pour le calcul des dissimilarités entre objets archéologiques issus de *BDFRues*.

8.2 Dissimilarités entre objets archéologiques

L'étude de la dissimilarité entre les objets archéologiques amène à définir quatre indices. Le premier détermine la dissimilarité temporelle, le deuxième porte sur la dissimilarité orientationnelle, le troisième évalue la dissimilarité de localisation et le quatrième conjugue les trois premiers afin d'évaluer la dissimilarité globale.

Pour calculer ces dissimilarités, je me base sur une métrique classique proposée dans [Grzegorzewski, 1998]. Cette métrique n'évalue pas l'aire séparant les fonctions d'apparte-

nance de deux nombres flous à l'instar de la distance dite de Hamming, mais détermine la distance séparant les extrémités de toutes les α-coupes. Ce choix permet de considérer les distances séparant les parties montantes et celles séparant les parties descendantes. C'est en considérant ces distances que, dans le cadre de ce chapitre, je détermine la dissimiliarité entre objets archéologiques.

8.2.1 Dissimilarités temporelles et orientationnelles

Soit F et G deux nombres flous, soit $F_{\alpha-}$ (resp. $G_{\alpha-}$) et $F_{\alpha+}$ (resp. $G_{\alpha+}$) la borne inférieure et la borne supérieure de l'α-coupe F_α de F (resp. G_α de G), alors la distance entre F et G est obtenue par :

$$D(F,G) = \int_0^1 |F_{\alpha-} - G_{\alpha-}| + |F_{\alpha+} - G_{\alpha+}| d\alpha.$$

J'utilise cette métrique pour le calcul de la dissimilarité d'orientation (D_{orien}) et de période d'activité entre éléments (D_{date}). Ainsi, pour deux objets archéologiques A_1 et A_2 issus de *BDFRues* ayant respectivement pour périodes d'activité $A_1.fDate$ et $A_2.fDate$ et pour orientations $A_1.fOrien$ et $A_2.fOrien$, l'indice de dissimilarité temporelle entre A_1 et A_2 est :

$$D_{date}(A_1,A_2) = \int_0^1 |(A_1.fDate)_{\alpha-} - (A_2.fDate)_{\alpha-}| + |(A_1.fDate)_{\alpha+} - (A_2.fDate)_{\alpha+}| d\alpha$$

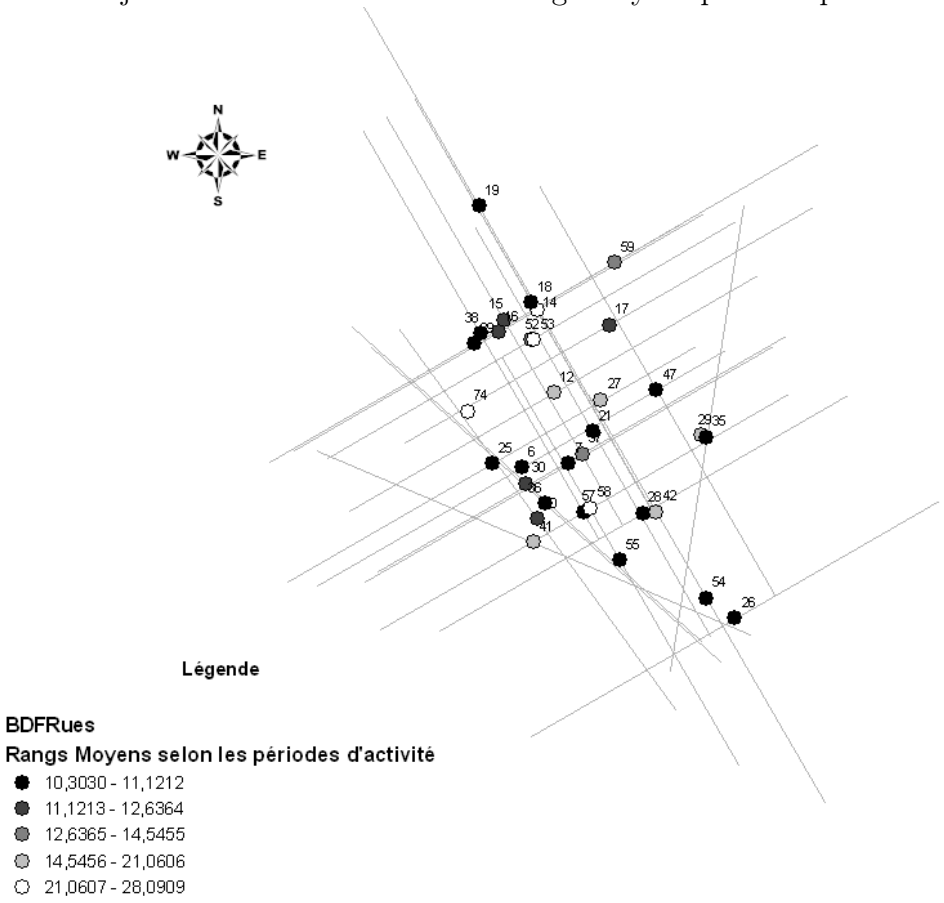
tandis que l'indice d'orientation est :

$$D_{orien}(A_1,A_2) = \int_0^1 |(A_1.fOrien)_{\alpha-} - (A_2.Orien)_{\alpha-}| + |(A_1.fOrien)_{\alpha+} - (A_2.fOrien)_{\alpha+}| d\alpha.$$

En utilisant l'indice de dissimilarité temporelle, j'obtiens, pour les objets de *BDFRues*, les rangs moyens présentés dans la Figure 8.1. Du rang moyen de chaque objet, on peut déduire sa représentativité temporelle. Ainsi les éléments les plus représentatifs temporellement sont ceux ayant les rangs moyens les plus petits, c'est-à-dire pour les objets de *BDFRues* les objets ayant un rang moyen prenant valeur entre 10.3 et 11.12 ; ils sont représentés dans la figure en noir. Les éléments les moins représentatifs temporellement obtiennent les rangs moyens les plus élevés — à partir de 21 — et sont coloriés en blanc.

La Figure 8.2 présente les rangs moyens des objets de *BDFRues* calculés à l'aide de l'indice de dissimilarité orientationnelle. En utilisant la relation liant le rang moyen à la représentativité, les éléments les plus représentatifs selon leurs orientations sont en noir, les moins représentatifs en blanc. Ainsi, les objets de *BDFRues* les moins représentatifs selon l'orientation ont des rangs moyens supérieurs à 20, tandis que les plus représentatifs ont des rangs moyens inférieurs à 8.7.

Figure 8.1 – Objet de *BDFRues* selon leurs rangs moyens pour les périodes d'activité



8.2.2 Dissimilarité de localisation

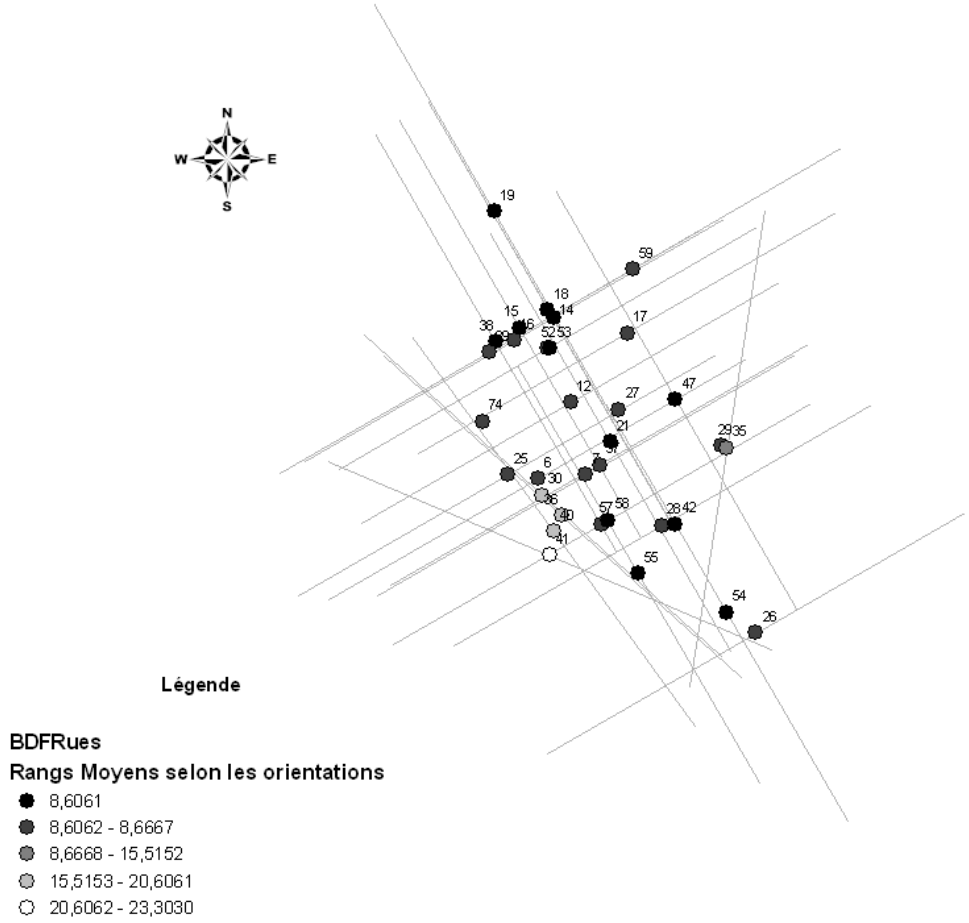
Pour le calcul de la dissimilarité de localisation, en raison du caractère cylindrique de la fonction d'appartenance des ensembles flous représentant les localisations associées aux données, je calcule la métrique de dissimilarité D_{loc} à partir de leurs projections sur le plan passant par les centres des localisations (voir Figure 8.3).

Ainsi l'indice de dissimilarité de localisation entre deux objets A_1 et A_2 de *BDFRues* est le suivant :

$$D_{loc}(A_1, A_2) = \int_0^1 |P(A_1.fLoc)_{\alpha-} - P(A_2.fLoc)_{\alpha-}| + |P(A_1.fLoc)_{\alpha+} - P(A_2.fLoc)_{\alpha+}| d\alpha.$$

La figure 8.2 dépeint les rangs moyens de chaque objet de *BDFRues* obtenus en utilisant l'indice de dissimilarité de localisation. Ainsi les objets de *BDFRues* ayant les localisations les plus représentatifs sont en noir — leurs rangs moyens sont inférieurs à 11 — et les moins représentatifs en blanc, c'est à dire les objets dont le rang moyen est supérieur à 19).

Figure 8.2 – Objet de *BDFRues* selon leurs rangs moyens pour les orientations



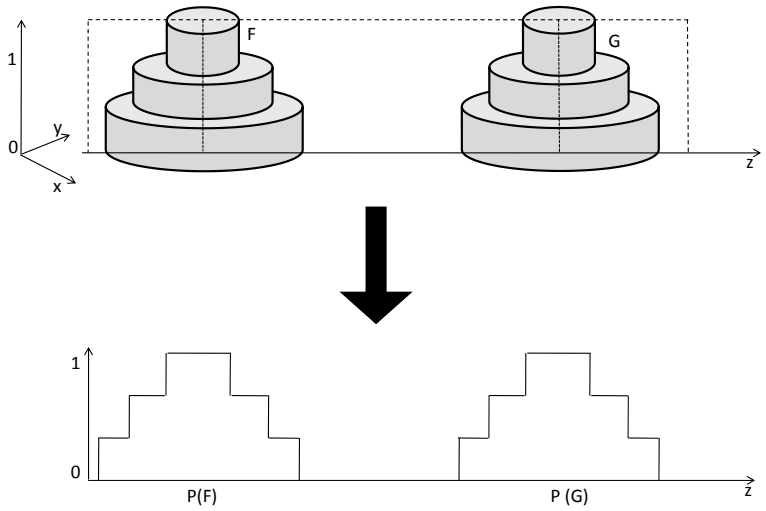
8.2.3 Dissimilarité globale

A l'instar des métriques de dissimilarité pour le calcul des rangs moyens en termes de localisation, d'orientation ou de période d'activité, on a besoin d'une métrique de dissimilarité pour l'ensemble des caractéristiques. Afin de l'obtenir, on normalise la dissimilarité liant un objet à un autre par la dissimilarité maximale du premier objet aux objets de l'échantillons. J'effectue cette normalisation des dissimilarités pour toutes les caractéristiques. La dissimilarité résultant de la moyenne de ces dissimilarités normalisées sera considérée comme la dissimilarité globale. Ainsi, soit deux objets A_1 et A_2 de *BDFRues* alors :

$$D_{global}(A_1, A_2) = \frac{1}{3} \times \sum_{i \in \{loc, orien, date\}} \frac{D_i(A_1, A_2)}{\max\limits_{A_j \in BDFRues} D_i(A_1, A_j)}$$

Avec l'indice de dissimilarité globale, je propose dans la Figure 8.5 d'affecter à chaque objet de *BDFRues* une couleur fonction de sa représentativité globale par le prisme de

Figure 8.3 – Projections pour le calcul de la dissimilarité de localisation



son rang moyen. Ainsi, les objets plus représentatifs sont coloriés en noir — leurs rangs moyens sont inférieurs à 13 — et les moins représentatifs en blanc ; leurs rangs moyens sont supérieurs à 18.

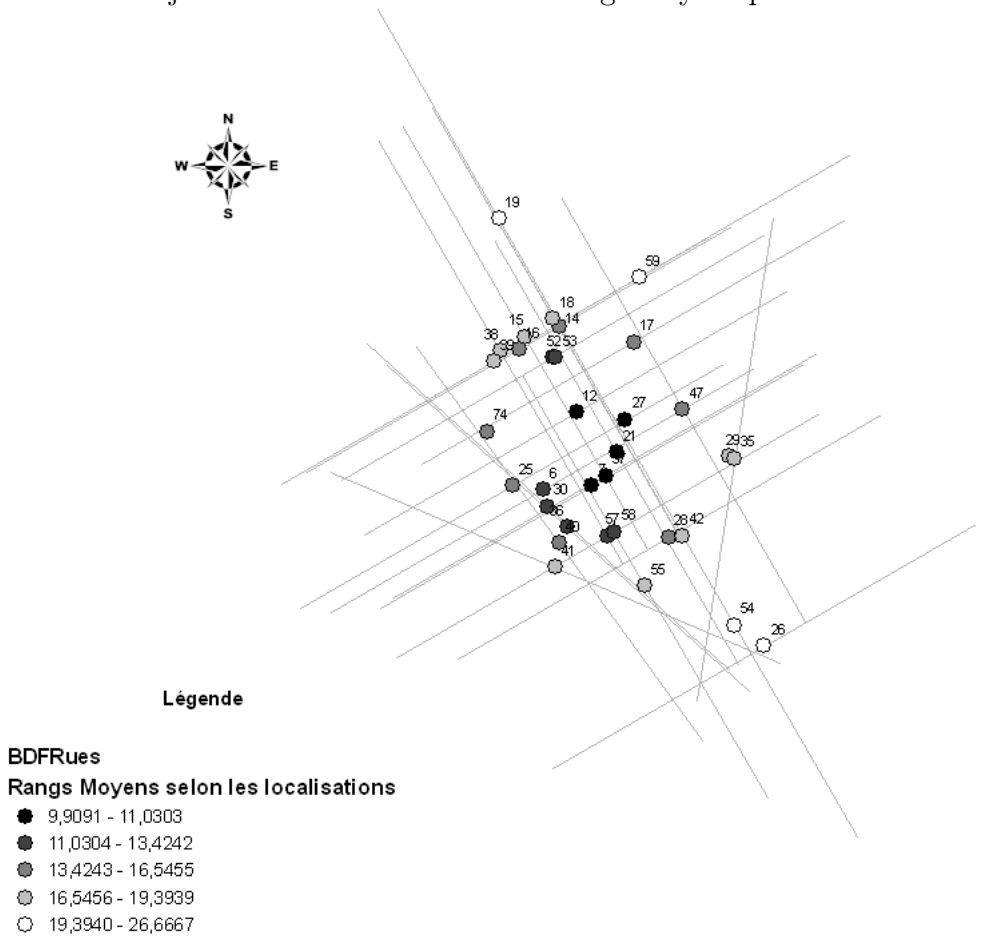
Dans l'optique de la classification des données, l'extraction des éléments les plus représentatifs d'une base de données archéologiques peut s'avérer significative. En effet, les objets les plus représentatifs d'une base de données sont « emblématiques » de la base ; les mettre en évidence donne une indication sur le positionnement global de la base de données. De plus, ces éléments sont des représentants réels et non virtuels de l'ensemble des objets étudiés et présentent donc un intérêt majeur pour la classification des objets.

8.3 Vecteur de Meilleur Rang Moyen

Dans le cadre statistique, les éléments $x_1, \ldots, x_n$ de Ω sont considérés comme les observations. Les variables aléatoire associées à Ω sont notées $X_1, \ldots, X_n$. Les statistiques de rangs associées sont les $X_{(1)}$, $X_{(2)}$, …, $X_{(n)}$ triées par ordre croissant. Par définition, les statistiques de rangs sont intrinsèquement liées à la façon dont sont triées les variables aléatoires. Dans le cas multidimensionnel, le tri n'est pas trivial [68].

68. On se reportera à [Barnett, 1976] pour une étude des différentes techniques pour trier des vecteurs multidimensionnels

Figure 8.4 – Objet de *BDFRues* selon leurs rangs moyens pour les localisations



8.3.1 Vecteur de meilleur rang moyen

La représentativité d'un élément dépend de son rang moyen. Plus petit est le rang moyen, plus la donnée est représentative. Le Vecteur de Meilleur Rang Moyen (VMRM) est une statistique qui extrait d'un ensemble l'élément le plus représentatif.

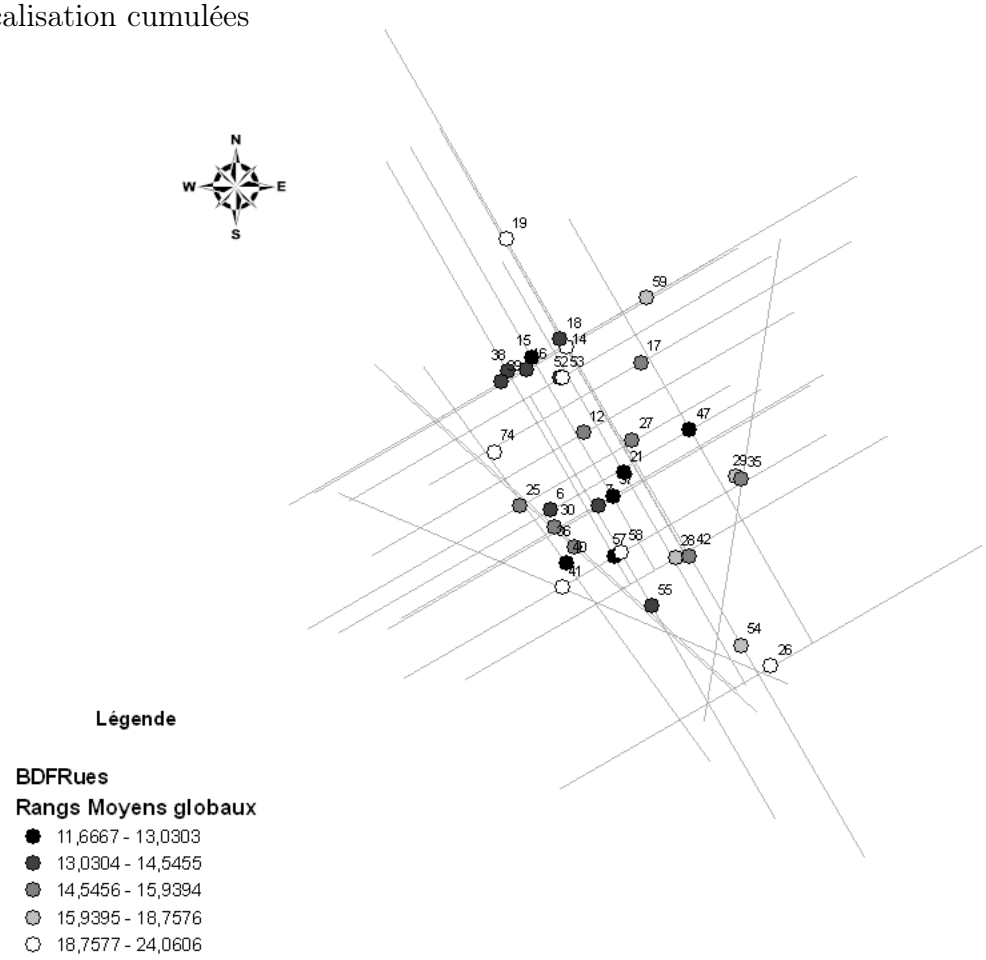
Dans un échantillon d'observation, la donnée la plus représentative correspond à la donnée ayant le plus petit rang moyen.

À l'aide des rangs moyens, je définis une statistique exprimée comme une fonctionnelle des statistiques de rangs et notée *VMRM* (Vecteur de Meilleur Rang Moyen) :

$$VMRM : (X_1, X_2, \dots X_n) \mapsto \underset{X_i, i=1..n}{argmin} \left(\overline{Rg}(X_j)\right)$$

Cette statistique associe à un échantillon l'élément qui le représente le mieux. De plus, au même titre que la médiane, cette statistique est un estimateur robuste de position de l'échantillon. En effet, la donnée de meilleur rang moyen est un élément typique et

Figure 8.5 – Objet de *BDFRues* selon leurs rangs moyens pour l'orientation, la datation et la localisation cumulées

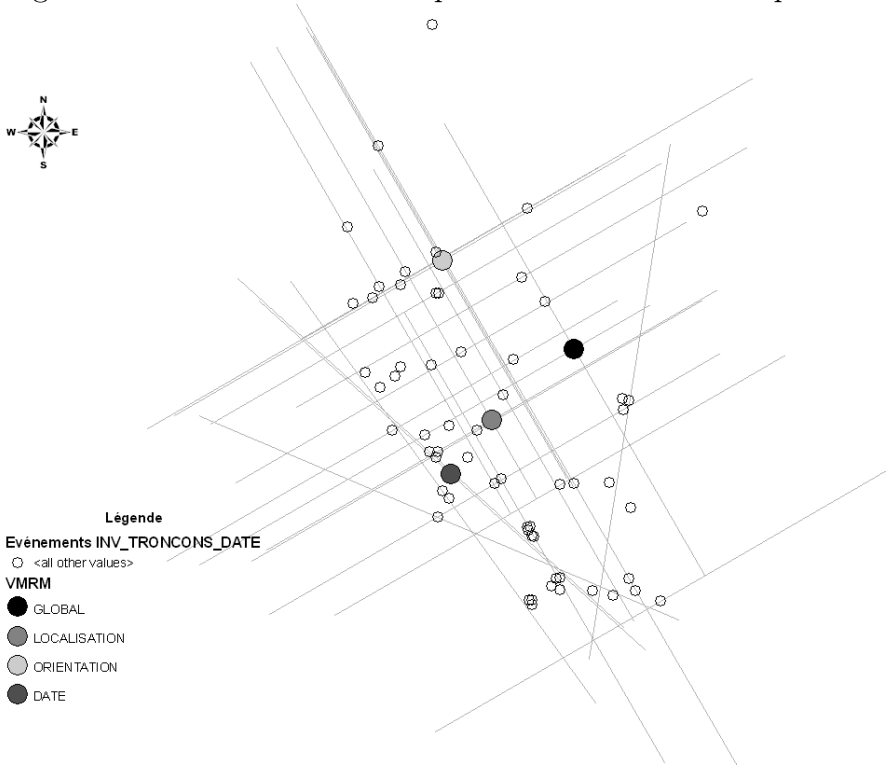


représentatif de l'échantillon.

La statistique *VMRM* de vecteur de meilleur rang moyen appliquée à l'échantillon Ω (voir 8.1.3 et tableau 8.1) donne finalement l'élément x_4 dont le rang moyen est minimal. On peut remarquer que la donnée x_5 est aberrante, mais qu'elle ne perturbe pas le résultat. En effet, on peut facilement vérifier que le vecteur x_4 est plus proche de la moyenne de l'ensemble des vecteurs privé de la donnée aberrante que le Vecteur Médian [Astola *et al.*, 1990] qui est ici x_2. Cette propriété de robustesse est une caractéristique des statistiques de rangs [David et Nagaraja, 2003].

L'objectif est d'utiliser le *VMRM* dans le contexte des rues de Durocortorum afin de déterminer les tronçons de rues emblématiques que sont les tronçons plus représentatifs de *BDFRues* en terme de localisation, d'orientation et/ou de période d'activité.

Figure 8.6 – Différents VMRM pour *BDFRues* : vision spatiale



8.3.2 Calcul des VMRM sur les objets archéologiques

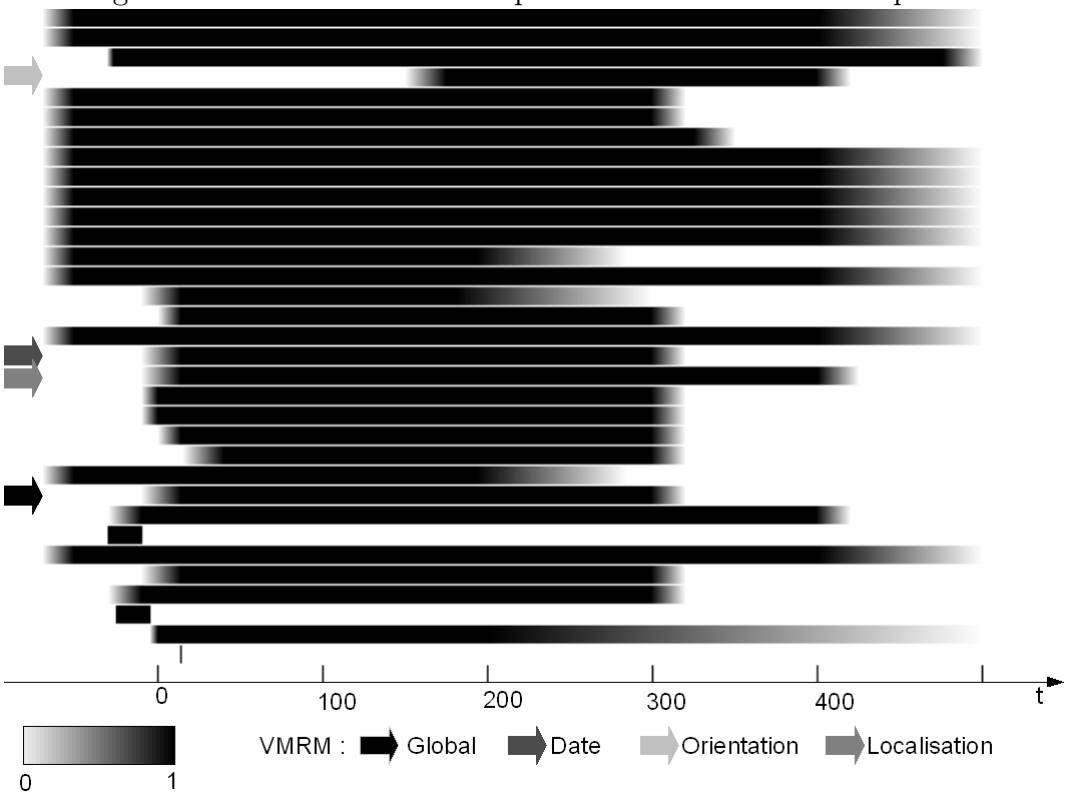
En calculant le *VMRM* pour la localisation, pour l'orientation, pour la période d'activité et les trois conjuguées on obtient les plans de la Figure 8.6.

On peut s'apercevoir que les tronçons représentatifs sont différents en fonction de la recherche effectuée, et en cela reflètent bien l'influence de chacune des caractéristiques des données archéologiques. Ainsi, on observe que le *VMRM* Global n'a pas la même orientation que le *VMRM* Orientation. Cela est dû au fait que le nombre de tronçons de rues dont l'orientation est proche de celle du *VMRM* Global est presque égal à celui des tronçons de rues dont l'orientation est proche de celle du *VMRM* Orientation (17 contre 16).

Enfin, c'est la période d'activité des tronçons de rues qui a entrainé le décalage au centre du *VMRM* Global par rapport au *VMRM* Localisation. Si on ne regarde que le *VMRM* Global, il est celui qui à la fois : est au centre, a l'une des deux orientations principales et a une période d'activité (voir Figure 8.7) qui se situe dans l'âge d'or de la période romaine de Reims (Haut-Empire).

◇◇◇◇◇

Figure 8.7 – Différents VMRM pour *BDFRues* : vision temporelle



Ainsi, dans ce chapitre, en utilisant la notion de représentativité, j'ai étudié la représentativité des objets archéologiques à l'aide d'indices de dissimilarité en termes de localisation, d'orientation et de périodes d'activité.

Une fois les représentativités des objets déterminées, je me suis intéressé aux éléments les plus représentatifs. Pour cela, j'ai défini la statistique de vecteur de meilleur rang moyen. Grâce à cette statistique, j'ai extrait les tronçons de rues représentant le mieux l'ensemble des tronçons de rues romaines en termes de localisation, d'orientation et/ou de périodes d'activité. Les tronçons ainsi extraits sont des représentants réels de l'ensemble des tronçons de rues trouvés à Reims. Ils sont donc utiles pour la classification des tronçons et pour la généralisation de l'information archéologique. Cette extraction a été effectuée en tenant compte de l'aspect spatial, temporel et imparfait de l'information archéologique.

Ce travail a fait l'objet de deux communications nationales : la première a porté sur l'utilisation du VMRM dans le filtrage d'images couleurs [69] [de Runz *et al.*, 2007d], et l'autre à son application dans le cadre des objets archéologiques [de Runz *et al.*, 2008b].

Dans le chapitre suivant, je présente un outil pour l'exploration des objets par l'inter-

69. Cette utilisation est présentée dans l'annexe C.

médiaire d'une image couleur. Grâce à cet outil, je proposerai de visualiser l'information temporelle imparfaite mais aussi l'adéquation spatiotemporelle des objets par rapport à un objet sélectionné.

Chapitre 9

Visualisation : image couleur des objets

Sommaire

L'étude intuitive et visuelle de l'ensemble des données associées aux objets d'une base de données archéologiques est complexe dans les SIG. En effet, bien que l'on puisse actuellement avoir une légende combinant un certain nombre d'attributs, ce nombre est limité. Dans cet objectif, l'exploration nécessite d'utiliser une technique de visualisation. Pour cela, il faut considérer ces composantes de l'information archéologique comme une collection de données multidimensionnelles [Guptill, 2005].

L'approche générale de l'exploration de grands volumes de données multicomposantes consiste à présenter un résumé en image de ces informations à l'instar de la démarche proposée par Keim (voir [Keim, 1996] et [Keim, 2000]). Afin de visualiser la plus grande quantité d'information possible, j'utilise une technique de visualisation qui construit une image couleur à partir de ces informations et qui fut introduite dans [Blanchard et Herbin, 2004]
.

Cette technique orientée-pixel consiste à représenter une collection par une image où chaque pixel correspond à une et une seule donnée. Les couleurs des pixels sont déterminées « objectivement [70] ». La couleur et la spatialisation fournissent alors une image qui constitue un résumé des données en permettant de voir de manière intuitive les principales structures. Ce travail a montré son efficacité sur des bases de données classiques [Blanchard *et al.*, 2005a].

Pour cela, les données multidimensionnelles sont préalablement réduites par une Analyse en Composantes Principales (voir [Rao, 1964]) à des données tridimensionnelles regroupant les trois composantes principales. En utilisant ces données réduites, à chaque donnée est associée un pixel d'une image couleur ; la couleur est affectée objectivement par la transformée inverse de celle proposée dans [Ohta *et al.*, 1980]. Afin de regrouper au maximum, dans l'image de visualisation, les données proches selon leurs trois premières composantes principales, les pixels représentant les données multidimensionnelles sont organisés spatialement à l'aide d'une courbe de remplissage dite de Peano-Hilbert. Cette technique de visualisation permet de dégager visuellement des informations structurelles (proximité, regroupement) sur les données visualisées. Cette méthode se place dans le champs des techniques de fouille de données et de l'extraction de connaissance.

Dans ce chapitre, deux processus d'exploration visuelle sont étudiés. Le premier a pour but de visualiser les composantes temporelles de l'information archéologique. Ces informations sont difficiles à visualiser dans le cas de grand volume de données, mais surtout l'exploration intuitive et directe de ces composantes est difficile. Dans le second processus, l'objectif est de visualiser les dissimilarités à un objet sélectionné dans la base de données.

Dans le cas de l'information archéologique, notamment celui des objets de *BDFRues*, les composantes temporelles et l'information orientationnelle sont modélisées en tenant compte de leurs imperfections. Il faut donc pré-traiter l'information afin d'en dégager des évaluations quantitatives. Ces évaluations quantitatives seront dès lors considérées comme des vecteurs multidimensionnels. Ainsi, la technique proposée est appliquée aux vecteurs multidimensionnels pour visualiser les objets archéologiques.

Je propose, dans ce chapitre, de visualiser les composantes temporelles des objets de *BDFRues*. Pour cela, comme énoncé précédemment, il est nécessaire de quantifier les données avant même le lancement du processus de visualisation. Dans ce but, je propose de construire un vecteur d'évaluation pour chaque nombre flou représentant la période d'activité d'un objet archéologique. Les différentes valeurs de ces vecteurs seront déterminées par différents estimateurs de quantités floues.

Pour visualiser les dissimilarités entre objets archéologiques, les vecteurs multidimensionnels nécessaires sont issus de trois indices de dissimilarité entre objets archéologiques proposés dans le chapitre précédent. Ces indices de dissimilarités permettent d'évaluer les dissimilarités temporelles, d'orientation et de localisation entre objets archéologiques.

70. La couleur est calculée à partir des données sans intervention d'une légende.

Dans l'image couleur de visualisation, le processus exploratoire regroupe spatialement autour du pixel représentant A_i les pixels représentant les objets les moins dissimilaires à A_i d'un point de vue spatial, directionnel et temporel.

Je vais, en premier lieu, présenter la technique de visualisation choisie. Ensuite, je présenterai le processus permettant de visualiser les composantes temporelles associées aux tronçons de rues trouvés à Reims datant de l'époque romaine. Enfin, j'exposerai le processus permettant de visualiser la dissimilarité d'un objet de la base vis-à-ivs des autres objets.

9.1 Visualisation de données par une image couleur

Afin de fournir une image couleur des objets, je m'intéresse plus particulièrement à la visualisation statique et plane de données multidimensionnelles quantitatives.

L'utilisation des outils de visualisation se heurte alors à deux difficultés principales : la dimension et l'effectif des échantillons de données.

La dimension de l'espace dans lequel se situent les données peut être importante et, dans certain cas, dépasser largement 100. Ceci conduit à un ensemble de phénomènes dissimulant l'information pertinente que l'on recherche. Ces phénomènes sont connus sous le nom de « malédiction de la dimensionalité » (curse of dimensionality [Bellman, 1961]).

Par ailleurs, l'effectif de l'échantillon peut être considérable et dépasser le million. Les techniques de visualisation ont alors tendance à masquer l'information pertinente du fait de cet effectif considérable.

Dans ce cadre, la méthode de [Blanchard *et al.*, 2005a] est une approche orientée pixels qui résume les données, et en fournit un résumé sous forme d'une image couleur. Cette approche permet de s'affranchir de la première difficulté signalée précédemment par la réduction de l'information et la seconde par l'association d'un pixel à chaque donnée ce qui autorise la visualisation d'autant de données que de pixels affichables.

9.1.1 Réduction de la dimensionalité

L'analyse de données multidimensionnelles nécessite une réduction de dimensionalité pour des raisons pratiques liées aux représentations des données [Healey et Enns, 1999] et théoriques liées à la malédiction de la dimensionalité [71]. Dans l'approche de visualisation présentée ici, les données sont dans un espace initial de dimension supérieure à trois.

Une approche classique, simple et généralement efficace de la réduction de dimensionalité est utilisée : on conserve les trois premières composantes générées par une Analyse en Composantes Principales [72] (ACP) [Rao, 1964].

71. Des exemples illustrant la malédiction de la dimensionalité sont présentées dans [Donoho, 2000]

72. Analyse en Composantes Principales (ACP) : des revues des techniques d'ACP sont proposées dans [Jolliffe, 1986], [Cardoso et Comon, 1996] et [Hyvärinen, 1999]. L'ACP est considérée comme une approche statistique usuelle en géographie pour synthétiser l'information. Elle est par exemple utilisée

Le principe est de projeter les données dans un sous-espace de dimension trois, les axes de projections étant orthogonaux et décorrélés dans le cas particulier de la Transformée de Karhunen-Loève (TKL). L'avantage de la TKL est que les composantes sont présentées par ordre décroissant de contenance d'information. Ainsi, la première composante contient plus d'information que la seconde, qui en contient plus que la troisième, et ainsi de suite. Ainsi, en réduisant les données de dimension $n > 3$ à des données de dimension 3 par la sélection des trois premières composantes (C_1, C_2, C_3) de la TKL, on maximise l'information contenue dans ces trois composantes.

Les données réduites guident ensuite le processus de visualisation. À chaque donnée de dimension trois est affecté un pixel que l'on place spatialement dans l'image à l'aide d'une courbe de Peano-Hilbert.

9.1.2 Remplissage de l'image de visualisation

Pour construire une image d'un échantillon de données, chaque donnée est associée à un pixel de l'image. Cette approche de la visualisation orientée-pixel permet de représenter des échantillons de grande taille [Keim, 2000]. La construction de l'image consiste à déterminer les coordonnées des pixels (i.e. des représentations des données) dans l'espace image.

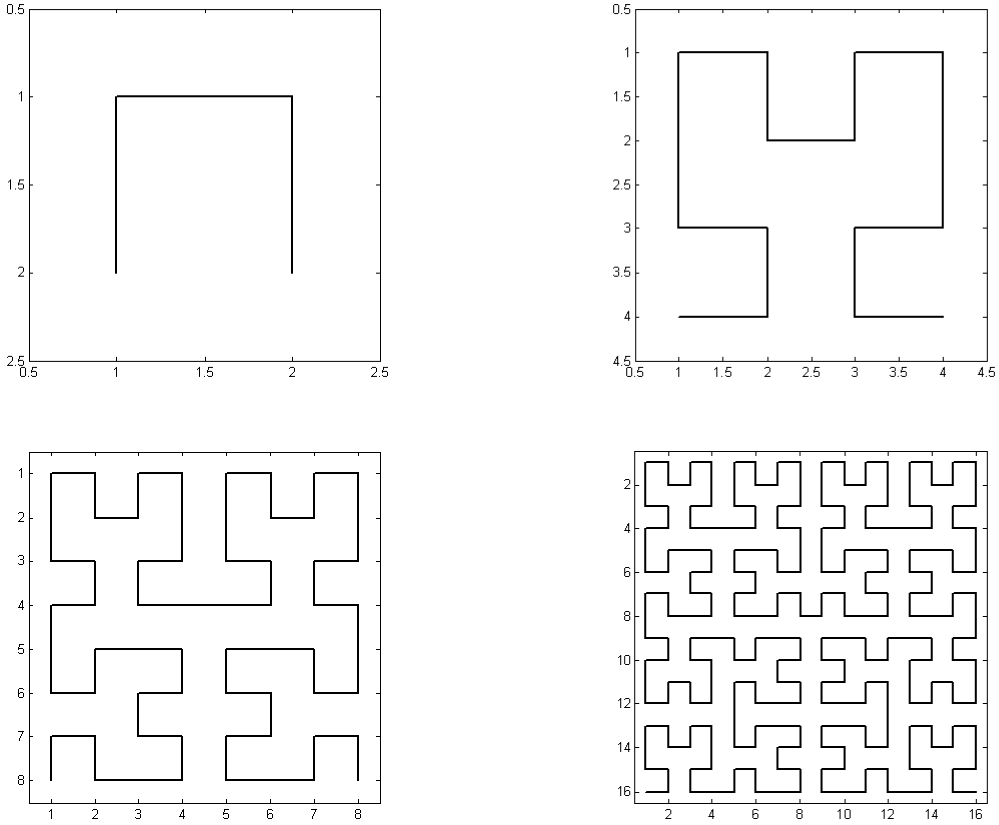
Si les pixels sont placés arbitrairement ou dispersés dans l'image, il devient difficile d'effectuer des rapprochements entre les données. Pour que l'image soit un outil de visualisation efficace, lisible au premier coup d'oeil de manière très intuitive, il faut que les proximités entre données soient faciles à déterminer. Pour cela, il faut que dans l'image résultat des données similaires soient spatialement très proches. La construction de l'image s'effectue en deux étapes : les pixels (i.e. les représentations des données) seront d'abord triés de manière à former une suite de pixels successifs ; ensuite cette ligne sera utilisée pour remplir l'image.

Ainsi la première étape du remplissage de l'image de visualisation consiste à trier les pixels associés aux données afin de produire une liste de pixels. Le tri des pixels est effectué en utilisant les résultats de la réduction de dimension des données. Les trois composantes obtenues (C_1, C_2, C_3) donnent trois clefs pour effectuer un tri sur l'ensemble des données de l'échantillon. En effet, les composantes issues de la TKL sont classées selon la quantité d'information qu'elles fournissent. Ainsi, le tri se fera en majeur sur la première composante car elle contient le plus d'information, puis sur la seconde composante et enfin en mineur sur la troisième composante.

L'étape suivante consiste à remplir l'image avec cette liste de pixels successifs. La courbe de Peano-Hilbert constitue le moyen le plus classique pour effectuer cette construction [Moon *et al.*, 2001] (voir sur la Figure 9.1 la description de la procédure récursive de construction d'une telle courbe). Le principal avantage de cette courbe est de préserver

dans [Jacquemot *et al.*, 2004]

Figure 9.1 – Étapes de la construction d'une courbe de Peano-Hilbert



au mieux la connexité des classes de données [Sasov, 1992]. En effet, cette courbe tend à minimiser les écarts entre d'une part les distances entre pixels dans l'image (distances dans l'espace image 2D) et d'autre part les différences de rangs des pixels sur la ligne initiale. Ces écarts seraient plus importants si l'on utilisait un parcours de l'image ligne par ligne ou bien colonne par colonne.

Avec ces deux étapes de tri des données puis le remplissage de l'image par une courbe de Peano-Hilbert, on évite de disperser les pixels dans l'image construite. Cette approche tend à préserver la cohérence spatiale des données permettant ainsi une visualisation très intuitive des échantillons de données.

Il faut maintenant déterminer la couleur de chaque pixel en fonction des valeurs contenues dans la donnée réduite associée au pixel.

9.1.3 A propos de la couleur

En imagerie, la couleur est souvent définie par un triplet (R,V,B) de trois valeurs Rouge, Vert et Bleu, codées sur 8 bits (entre 0 et 255). Après avoir réduit la dimension de l'échantillon, chaque donnée est représentée par un triplet (C_1, C_2, C_3). Malheureusement

on ne peut considérer que les trois composantes principales forment directement le triplet RVB car on nierait les quantités relatives d'information contenues par les composantes résultantes de la TKL.

La technique de visualisation proposée se basent pour l'affectation de la couleur sur l'étude statistique de la couleur de [Ohta *et al.*, 1980]. Ces derniers proposent d'approximer (de donner des valeurs proche de) la transformation de Karhunen-Loève (TKL) d'une image couleur par une transformation linéaire. A partir des données (R, V, B) des pixels couleur, ils calculent les triplets (C_1, C_2, C_3) qui approximent les trois composantes de la TKL [Ohta *et al.*, 1980]. La technique de visualisation exposée ici cherche à associer une couleur à chaque triplet (C_1, C_2, C_3). Par la transformation inverse de celle de Ohta *et al.*, elle associe à chaque donnée une couleur (R, V, B). Ainsi les composantes couleurs R, V et B sont liées aux composantes C_1, C_2 et C_3 par la relation suivante :

$$\begin{cases} R &= (6 \times C_1 + 3 \times C_2 - 2 \times C_3)/6 \\ V &= (3 \times C_1 + 2 \times C_3)/3 \\ B &= (6 \times C_1 - 3 \times C_2 - 2 \times C_3)/6 \end{cases}$$

Ce type d'approche présente l'avantage d'être objectif et non supervisé contrairement aux méthodes traditionnelles de détermination de palettes ou d'échelles de couleurs. Cette approche de la couleur dépend de l'échantillon de données. Si l'échantillon change, les couleurs changent. Elle ne propose qu'un résumé coloré associé à un échantillon. Cette technique de visualisation a été appliquée avec succès à des bases de données classiques, des données simulées et à des images de fluorescence X.

L'idée dans l'approche de la fouille de données proposée dans ce chapitre est d'utiliser la technique de visualisation exposée précédemment afin de visualiser les objets selon l'information des composantes temporelles, et les dissimilarités des objets archéologiques en tenant compte de leurs imperfections.

9.2 Visualisation des objets selon leurs périodes d'activité par une image couleur

Les données archéologiques sont spatiotemporelles et imparfaites. Afin de donner une lecture intuitive des données, il est nécessaire de les visualiser. Lorsqu'elles sont stockées dans une BDG associée à un SIG, une visualisation classique consiste à la production d'une ou plusieurs cartes thématiques. Cependant ces cartes ne permettent pas de rapprocher spatialement les objets aux localisations éloignées, et les couleurs dépendent d'une échelle fiée (d'une légende). Afin de rapprocher les données archéologiques selon le temps en prenant en considération l'imperfection, il est nécessaire d'utiliser une autre approche pour la visualisation.

La visualisation d'un grand nombre de données spatiotemporelles imparfaites est difficile. En effet, dans le cadre de données représentées par des quantités floues, représenter l'ensemble des fonctions d'appartenance sur un même repère complique la lecture des quantités floues à considérer. Cependant, de nombreuses techniques de visualisation de données multi-composantes, à l'instar de celle présentée précédemment, utilisent les informations quantitatives contenues dans ces données afin de les visualiser. Une solution à la visualisation de quantités floues consiste à visualiser les données par l'intermédiaire d'évaluations des différentes quantités floues. Ces évaluations forment des données quantitatives multi-composantes décrivant l'information à visualiser.

L'analyse de données floues nécessite généralement une défuzzification des données. La défuzzification est le processus qui amène à produire un résultat quantifiable à partir de données floues. Ainsi, par exemple, les méthodes de comparaison de quantités floues rangent le plus souvent celles-ci par le biais d'évaluations [Wang et Kerre, 2001a].

Le principe de la visualisation consiste donc à d'abord décrire une quantité floue par plusieurs évaluations quantitatives obtenues avec différentes méthodes de défuzzification. Une quantité floue est alors représentée par un vecteur d'évaluations. Les données sont ensuite visualisées à l'aide d'une technique utilisant les vecteurs d'évaluations qui les décrivent.

Dans BDFRues, les objets archéologiques sont temporellement modélisés par des nombres flous. Le but de la visualisation est alors d'associer à chaque objet archéologique un pixel couleur de l'image résultat en fonction du nombre flou représentant sa période d'activité.

L'objectif ici est d'abord d'évaluer chaque nombre flou séparément puis de le positionner par rapport aux autres par la technique de visualisation précédente via ses évaluations. Les évaluations des nombres flous doivent donc ne prendre en compte que le nombre flou devant être visualisé.

9.2.1 Méthodes de défuzzification des nombres flous

Les méthodes de défuzzification présentées ici ne considèrent que le nombre flou à évaluer. Dans [VanLeekwijck et Kerre, 1999], Van Leekwijck et Kerre les classent en trois classes. Bien que chaque méthode ait ses particularités, les classes proposées invitent à des utilisations différentes. Le choix de l'utilisation de l'une de ces méthodes dépend donc fortement de l'analyse voulue.

Les méthodes de type maxima et les méthodes dérivées forment la première classe. Elles sélectionnent un élément du cœur de la quantité à évaluer comme valeur de défuzzification. Selon Van Leekwijck et Kerre, l'utilisation première de ces méthodes se situe dans le cadre des systèmes de connaissances floues. De plus, ces méthodes sont efficaces d'un point de vue calculatoire.

Dans la seconde classe, les opérateurs de défuzzification convertissent en premier les fonctions d'appartenance en distribution de probabilités afin de calculer la valeur espérée.

Au regard du manque de fondement théorique de ces conversions, la principale raison de leur utilisation est que ces méthodes vérifient l'hypothèse de continuité, si désirable pour les contrôleurs flous [73].

Dans la troisième classe, les méthodes utilisent les aires sous les fonctions d'appartenance afin d'évaluer les quantités floues. Comme pour les méthodes de la seconde classe, elles sont principalement utilisables dans le cadre du contrôle flou.

Je souhaite, afin d'explorer visuellement les objets archéologiques selon les représentations de leurs périodes d'activité, définir pour chacun des nombres flous, modélisant la période d'activité de ces objets, un vecteur d'évaluation le représentant dans le processus de visualisation.

9.2.2 Construction du vecteur multidimensionnel d'évaluation de la représentation d'une période d'activité

Afin de construire un vecteur, simple de conception, constitué de méthodes de chaque classe, je propose de n'utiliser que des méthodes de défuzzification sélectionnées parmi celles présentées dans [VanLeekwijck et Kerre, 1999] afin qu'elles ne prennent pas en considération de paramètre autre que l'ensemble à considérer, l'objectif final étant de proposer une visualisation temporelle non supervisée.

Pour la première classe, je choisis les méthodes suivantes : le "first of maximum" (FOM) qui retourne le plus petit élément du cœur d'un nombre flou ; le "last of maximum" (LOM) qui renvoie le plus grand élément du cœur d'un nombre flou ; le "middle of maximum" (MOM) qui permet de récupérer l'élément médian du cœur d'un nombre flou.

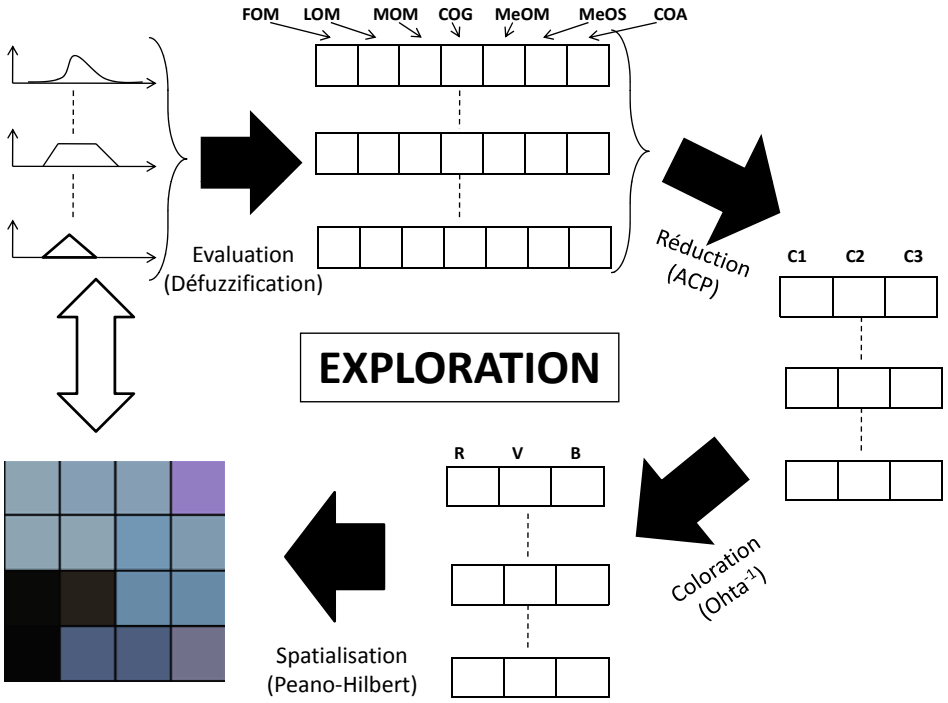
Pour la seconde classe, je sélectionne les méthodes suivantes : le "center of gravity" (COG) qui donne en sortie le centre de gravité de la fonction d'appartenance d'un nombre flou ; le "mean of maxima' (MeOM) qui calcule la moyenne du cœur d'un nombre flou ; le "mean of support" (MeOS) par lequel on obtient la moyenne du support d'un nombre flou.

Pour la dernière classe, je prends le "center of area" (COA) car celui-ci permet d'obtenir l'élément du support minimisant la différence des aires de la fonction d'appartenance avant et après ce dernier.

J'associe donc à chaque période d'activité un vecteur d'évaluation de dimension 7. C'est par le prisme de ce vecteur que j'explore les données. Pour cela, j'utilise l'ensemble des vecteurs en entrée au processus de visualisation de Blanchard et Herbin. Le processus général de l'exploration est présenté dans la figure 9.2.

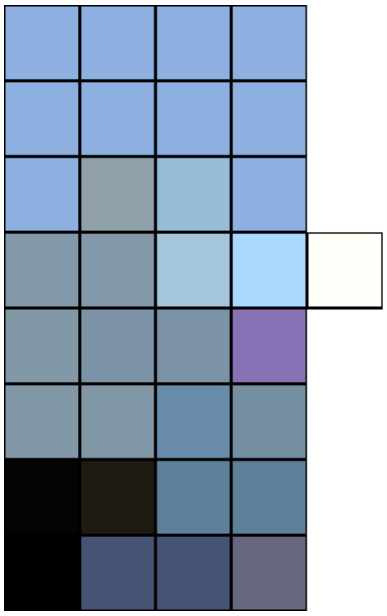
73. Le contrôle flou tire son nom des applications de contrôle ou de commande en automatique. Le principe de l'algorithme de contrôle — le controleur — est très simple, il consiste à réaliser une « interpolation » entre un petit nombre de situations connues données par un expert sous la forme de règles floues du genre « si x est petit et y est modéré, alors u doit être très grand ».

Figure 9.2 – Visualisation des objets selon les représentations flous des périodes d'activité — schéma récapitulatif



9.2.3 Visualisation des objets selon les représentations de leurs périodes d'activité

Figure 9.3 – Visualisation des objets de *BDFRues* par une image couleur les représentations de leurs périodes d'activité



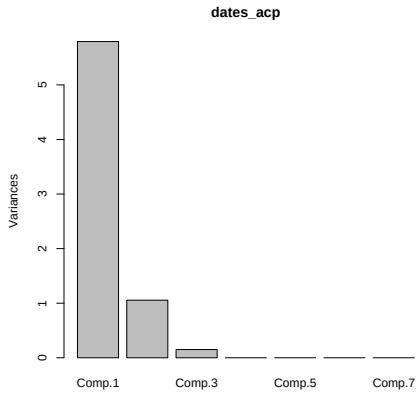
L'image résultante est présentée sur la figure 9.3. Elle contient 33 pixels en couleurs. Chaque pixel représente un objet selon le nombre flou associé à la période d'activité de l'objet. L'organisation spatiale et couleur des pixels permet d'observer de façon immédiate des informations de structuration de cet ensemble de périodes. Cette image suggère des regroupements des données par couleurs semblables.

Les regroupements observés correspondent à des objets ayant des représentations de leurs périodes d'activité de profils proches. En effet, les objets, dont les représentations des périodes d'activité ont les plus larges supports, sont visualisés par des pixels de couleur bleue claire (haut de l'image de visualisation), tandis que ceux dont les cœurs des représentations des périodes d'activité sont de cardinalité moyenne, sont coloriés dans les gris (milieu de l'image).

On observe par ailleurs que les composantes principales calculées sur l'ensemble des vecteurs décrivant les représentations des périodes d'activité des objets issus de *BDFRues* permettent d'expliquer plus de 99% de la variance totale (voir figure 9.4). Cette opération de projection conserve donc la quasi totalité de l'information apportée par les différentes évaluations. La visualisation porte donc sur l'essentiel de l'information temporelle contenue dans *BDFRues*.

Ainsi, l'image couleur résultant de la visualisation permet une lecture intuitive — par proximité spatiale et de couleur — d'un grand nombre d'objets archéologiques selon la

Figure 9.4 – Histogramme de l'ébouli des valeurs propres de l'ACP



proximité entre les représentations de leurs périodes d'activité. C'est un outil facilitant l'analyse exploratoire des données.

La section suivante porte sur une visualisation analogue (par le biais d'un vecteur d'évaluation) des dissimilarités entre les objets archéologiques de la base et un objet sélectionné.

9.3 Visualisation des dissimilarités à un objet sélectionné

L'objectif de ce processus exploratoire est de visualiser par une image couleur les objets selon leur dissimilarité à un objet sélectionné. On utilise pour cela les mêmes indices de dissimilarité que dans la section 8.2 en termes d'orientation, de localisation, de période d'activité, c'est à dire D_{date}, D_{orien} et D_{loc}.

9.3.1 Construction du vecteur multidimensionnel d'évaluation des dissimilarités à un objet sélectionné

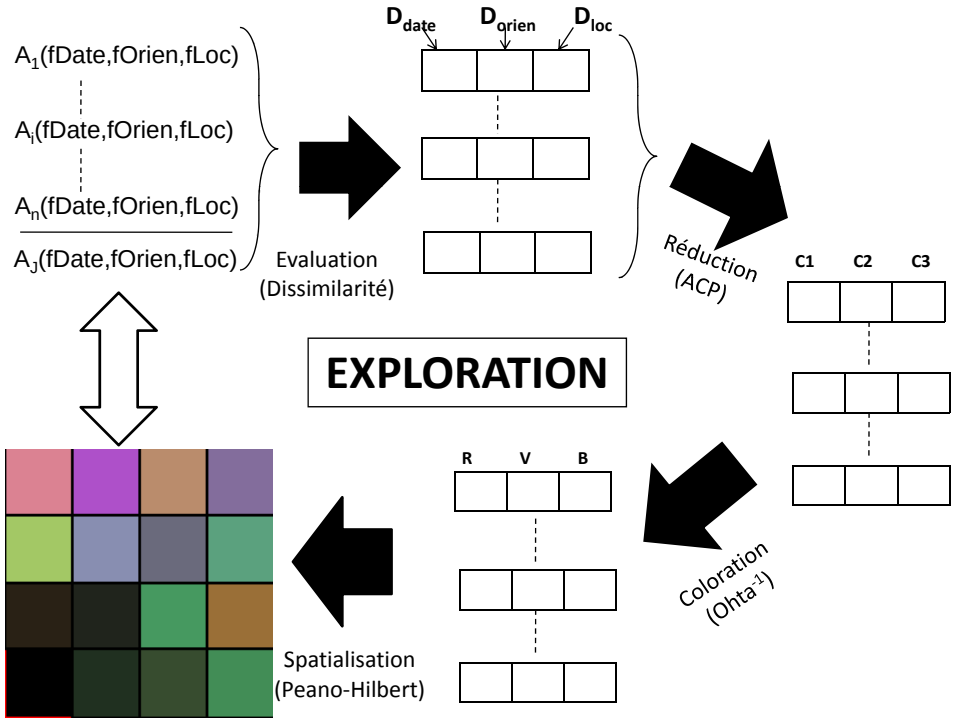
Dans *BDFRues*, une fois l'objet A_j sélectionné, le vecteur d'évaluation $v_{A_j}(A_i)$ de chaque objet A_i à évaluer est déterminé par les dissimilarités de cet objet avec A_i. Ainsi, $v_{A_j}(A_i)$ est défini de la manière suivante :

$$v_{A_j}(A_i) = (D_{date}(A_j, A_i), D_{orien}(A_j, A_i), D_{loc}(A_j, A_i).$$

Ce vecteur contient trois informations *a priori* décorrélées — D_{date}, D_{orien} et D_{loc}.

Je propose de visualiser la dissimilarité des objets à un objet sélectionné dans la base, en donnant en entrée du processus de visualisation ces vecteurs de dimension 3. Le

Figure 9.5 – Visualisation des objets selon leurs dissimilarités à un objet sélectionné — schéma récapitulatif

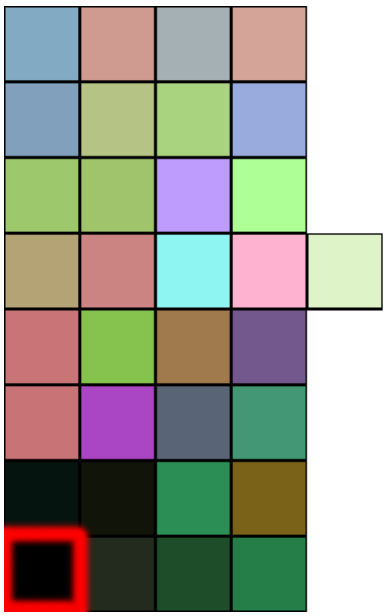


processus exploratoire est présenté dans la figure 9.5.

9.3.2 Visualisation des dissimilarités

Le processus de visualisation des dissimilarités des objets à un objet sélectionné donne la Figure 9.6.

Figure 9.6 – Visualisation de la dissimilarité des objets à un objet sélectionné par une image couleur



Comme cette application prend à la fois en compte les dissimilarités d'orientation, de localisation et de datation, celle-ci ne fait plus apparaître les classes observées dans la visualisation strictement temporelle.

Les objets les plus similaires à l'objet sélectionné sont dans l'image les plus proches spatialement de celui sélectionné (de contour rouge dans la figure). Il y a donc *a priori* trois objets très similaires à celui sélectionné.

◇◇◇◇◇

J'ai présenté dans ce chapitre une méthode originale de visualisation des objets archéologiques permettant une exploration intuitive de l'information archéologique. Cette méthode s'est basée sur la construction de vecteurs dont les valeurs furent obtenues soit

par plusieurs défuzzifications des représentations (nombres flous) de la composante observée, soit par calcul des indices de dissimilarité des objets à un objet en entrée. L'étape de visualisation a consisté à affecter à chaque objet archéologique une couleur pour obtenir des pixels que l'on organise spatialement dans une image. Dans ce but, j'ai réduit les vecteurs d'évaluations par une ACP à des vecteurs de dimension 3. Par la transformée inverse de celle d'Ohta *et al.* (1980), j'ai déterminé les couleurs des pixels représentant les objets. L'image fut alors construite en utilisant une courbe de Peano-Hilbert.

Cette visualisation est strictement exploratoire. Elle permet de faire des rapprochements entre données et de les regrouper pour aider à les interpréter. C'est un outil qui présente d'autant plus d'intérêt que le nombre de données augmente (il offre la possibilité de visualiser plusieurs millions de données).

L'image résultante fournit une carte synthétique de la base archéologique étudiée en fonction de l'objectif du processus exploratoire. Cette image peut être considérée comme une légende organisée de l'information multidimensionnelle visualisée qui associe à chaque objet une couleur de manière objective.

Le travail présenté dans ce chapitre est l'objet d'une communication dans une conférence d'audience nationale [de Runz *et al.*, 2008a]. Cette soumission porte sur la visualisation de quantités floues via la construction d'un vecteur d'évaluation pour chacune d'entre elles.

Conclusion

Cette partie développe plusieurs outils pour l'analyse des relations entre données spatiotemporelles imparfaites dans un SIG ; les données ont été préalablement modélisées selon la démarche proposée dans la partie I. Dans ce contexte, des outils spécifiques adaptés aux données imparfaites décrivent le positionnement de chaque objet contenu dans une base de données par rapport aux autres objets de la base. Ces outils ont pour objectif de faciliter les analyses de données archéologiques. Ces analyses ont besoin d'explorer temporellement la base de données, de déterminer la représentativité de chaque objet dans la base afin d'en extraire des objets emblématiques et aussi de visualiser des informations quantitatives relatives aux objets. Pour cela, j'ai proposé deux processus exploratoires différents ainsi qu'un processus de visualisation.

Le premier, exposé dans le chapitre 7, permet la représentation des positions temporelles des objets relativement aux autres. Pour cela, à partir de l'indice d'antériorité présenté dans le chapitre 4, un graphe orienté pondéré est construit, permettant ainsi une vision schématique des relations temporelles entre objets. Par l'intermédiaire de ce graphe, le potentiel d'antériorité, celui de postériorité et la position temporelle de chaque objet dans l'échantillon sont déterminés. Dans ce contexte, l'objet le plus antérieur (resp. le plus postérieur) à l'ensemble des objets est celui dont la position temporelle est la plus petite (resp. la plus grande). Ce processus exploratoire permet donc de regarder la position temporelle de chaque objet dans un ensemble mais aussi de mettre en valeur des objets aux caractéristiques temporelles particulières.

Le second, présenté dans le chapitre 8, détermine, dans un premier temps, la représentativité de chaque objet, et extrait, dans un deuxième temps, les objets les plus représentatifs qui sont emblématiques de la base. La représentativité de chaque objet est déterminée à l'aide du rang moyen qui lui est attribué en fonction de sa dissimilarité à chaque objet de la base. Pour calculer cette dissimilarité, plusieurs indices sont utilisés : un indice temporel, un indice d'orientation et un indice de localisation. Les éléments les plus représentatifs sont alors ceux de meilleurs rangs moyens. Ce processus de fouille de données regarde donc la représentativité de chaque objet à travers une étape d'abstraction (le passage par les rangs) et d'indices de dissimilarité. Par cette méthode, l'extraction des objets les plus représentatifs de l'ensemble permet de valoriser le patrimoine par l'obtention d'objets caractéristiques des données. Elle fournit aussi des représentants réels utilisables par les processus de classification des données.

Dans le chapitre 9, j'ai visualisé les objets archéologiques dans une image couleur en fonction soit des représentations de leurs périodes d'activité, soit de leurs similarités à un objet sélectionné . Le processus de visualisation utilise la définition de vecteurs d'évaluations quantitatives associées aux objets afin d'obtenir une image couleur construite à l'aide de la technique de [Blanchard, 2005]. Cette technique nécessite la transformation des données par une Analyse en Composantes Principales et la sélection des trois premières composantes principales issues de l'ACP, puis le calcul par la transformée inverse de [Ohta *et al.*, 1980] des couleurs des pixels de l'image et enfin le remplissage de l'image par une courbe de Peano-Hilbert. Dans le but de visualiser les informations temporelles, les vecteurs d'évaluation résultent de méthodes de défuzzification de nombres flous. Pour visualiser les proximités de localisation, d'orientation et de datation, les valeurs d'indices de dissimilarité associées à ces proximités composèrent les vecteurs. La visualisation proposée est un processus pré-exploratoire fournissant une carte synthétique des objets.

Les outils détaillés dans cette partie ont été présentés à la communauté scientifique à travers trois communications dans des conférences d'audience nationale. Dans la communication [de Runz *et al.*, 2007d], j'ai présenté l'application du *VMRM* pour le filtrage d'image couleur (voir annexe C). La communication [de Runz *et al.*, 2008b] a porté sur l'extraction à l'aide du *VMRM* des tronçons de rues, trouvés à Reims et datant de l'époque romaine, les plus représentatifs en termes de localisation, datation et/ou orientation. Dans la troisième communication, [de Runz *et al.*, 2008a], je présente une méthode pour l'exploration visuelle de quantités floues que j'applique aux représentations floues des périodes d'activité des données archéologiques contenues dans *BDFRues*. Ces outils constituent une avancée pour le traitement et l'analyse des données imparfaites en particulier dans les SIG archéologiques.

Conclusion générale

-

« Einstein published Special Relativity in 1905... the conclusion was that the universe is a lot weirder than common sense tells us, although they probably didn't use that actual word. »

Ian Stewart, Jack Cohen, Terry Pratchett, *The science of the Discworld* (1999).

Face aux enjeux urbains actuels, à la patrimonialisation des ressources archéologiques, et grâce au développement de l'informatique, l'information issue de fouilles archéologiques préventives est de plus en plus disponible en format numérique, afin de pouvoir répondre aux besoins de la société en terme de valorisation du territoire et d'aménagement urbain.

En effet, le contexte, une génération après les années de reconstruction d'après-guerre, est celui d'une transformation structurelle des formes urbaines, tributaires de nouvelles fonctionnalités suscitées par une concentration démographique et culturelle sans précédent. Dans cette perspective, les mutations liées à la politique de la ville conduisent, sur la base des restructurations dédiées à la durabilité et à la qualité environnementales, à patrimonialiser la ressource archéologique mise au jour. Désormais, vecteurs d'identité et de légitimité historique pour les génération urbaines, les vestiges composent des éléments fragmentés dont on cherche la cohérence dans le temps et dans l'espace, comme autant des preuves de filiation territoriale.

Sous leur forme numérique, les données archéologiques deviennent plus faciles à manipuler, gérer, stocker, analyser et visualiser. L'information archéologique étant attachée à un sol, un espace, l'utilisation de systèmes d'information géographique devient essentielle pour l'exploitation des données. Cependant, les différences entre informations archéologiques et géographiques demandent le développement d'une approche propre à l'information archéologique.

Les travaux de recherches, rapportés dans cette thèse et effectués dans le cadre du projet *SIGRem* et du « Centre Image » portés par l'Université de Reims Champagne-Ardenne, avaient pour problématique la généralisation des données archéologiques et l'étude de la structure d'un espace urbain antique (Durocortorum). Pour cela, la modélisation, l'analyse et la visualisation de l'information archéologique dans un SIG en prenant en considération l'aspect temporel, l'aspect spatial et les imperfections des données archéologiques s'avèrent nécessaires.

L'objectif visé était de répondre à trois questions principales portant d'abord sur le choix de la théorie de la représentation et de modélisation des données spatiotemporelles imparfaites, puis sur l'élaboration de méthodes nécessaires à l'analyse des données précédemment modélisées et enfin sur la construction de techniques de visualisation devant faciliter la compréhension des données.

Bilan et contributions

Dans la première partie « *Représentation et modélisation de données imparfaites dans un SIG archéologique* », j'ai exposé une démarche portant sur le choix de la représentation et de la modélisation des données archéologiques en tenant compte de leurs imperfections. Cette démarche est structurée en trois étapes : l'étude de la nature de l'imperfection des données, la construction d'une taxonomie pour associer aux imperfections de l'information une théorie de formalisation et l'utilisation des modèles en adéquation aux théories de

représentation choisies.

Afin d'exposer les principes de cette approche, j'ai étudié, dans le chapitre 1, la nature des imperfections de l'information archéologique, puis j'ai présenté, dans le chapitre 2, les principales théories de formalisation de l'imparfait nécessaires à la représentation des données imparfaites. Enfin, dans le chapitre 3, la nature des imperfections de l'information archéologique et les spécificités des théories de l'imperfection m'ont amené à proposer une taxonomie de l'imperfection spécifique à l'information archéologique, et à présenter l'étape de modélisation nécessaire à l'exploitation des données.

Le cadre applicatif issu du projet SIGRem (*BDRues*) donnent corps à ces données, qui sont à la fois vagues, imprécises et incomplètes. Le choix de la représentation s'est porté, en suivant la taxonomie proposée, sur la théorie des ensembles flous car elle est la plus classique et semble la plus appropriée pour la formalisation de l'imprécis et du vague. L'étape de modélisation a fait appel pour ces données à des ensembles flous convexes et normalisés. Cette démarche a été présentée à la communauté scientifique lors de la conférence internationale CAA'2008. Elle m'a permis de répondre à la première question.

Dans la deuxième partie « *Analyse de données spatiotemporelles imparfaites dans un SIG archéologique* », je me suis attaché à l'analyse des interactions des données présentes dans un SIG archéologique avec des informations externes au système. Dans ce cadre, deux nouvelles méthodes d'interrogation des données et une étude sur l'appariement de données archéologiques ont été proposées.

La première méthode d'interrogation, présentée dans le chapitre 4, porte sur le positionnement temporel des données vis-à-vis d'une période et pour cela utilise un nouvel indice, appelé indice d'antériorité, quantifiant l'antériorité entre deux périodes. Grâce à cette analyse, l'étude de l'évolution des objets en fonction du temps en prenant en considération l'imperfection de l'information est possible. La seconde méthode, exposée dans le chapitre 5, propose des scénarios de reconstruction en fonction d'une forme et d'une période recherchée, et, dans ce sens, adapte une méthode classique de reconnaissance de forme : la transformée floue de Hough. Ces scénarios présentent une complétion partielle de l'information et lient des données issues de fouilles différentes. Cette analyse participe donc à la généralisation des données spatiotemporelles. Elle constitue de plus une approche très originale permettant de fusionner des informations temporelles et spatiales grâce à des hypothèses de formes. Dans le chapitre 6, l'étude de l'appariement des données archéologiques m'a conduit à privilégier l'approche d'Olteanu et à en proposer une évolution mais aussi à proposer des critères spécifiques à la structure de l'information archéologique. Grâce à cette étude, les mises en correspondance de données spatiotemporelles imparfaites issues de différentes sources deviennent possibles et forment ainsi une analyse multi-source.

Les méthodes exposées ont été présentées à la communauté scientifique durant trois conférences internationales (IPMU'06, RCIS'07 et ISSDQ-07) et une conférence francophone (Conférence Francophone ESRI). De plus, les résultats du traitement spatiotempo-

rel sous critère de forme ont été présentés aux experts à travers le numéro 111 du magazine « Culture et Recherche » publié par le Ministère de la Culture. Par ailleurs, deux articles ont été soumis à des revues internationales (GeoInformatica [de Runz *et al.*, 2008d] et Soft Computing [de Runz *et al.*, 2008c]).

Les méthodes d'analyse des interactions des données avec des informations externes forment, de par leurs résultats, des outils pertinents pour l'expertise archéologique et l'analyse géographique des données. Elles considèrent les données modélisées à l'aide de la démarche présentée dans la partie I, c'est-à-dire qu'elles prennent en compte les données dans leur spatialité, leur temporalité et leur imperfection. Par la démarche analytique proposée à travers ces trois méthodes, j'ai développé une partie de la réponse à la deuxième question de la problématique de ce mémoire. Ces méthodes, que j'ai appliquées au contexte archéologique, peuvent aussi être pertinentes pour l'analyse de données en géographie et en informatique.

À travers la troisième et dernière partie « *Analyse exploratoire et visualisation de données spatiotemporelles imparfaites dans un SIG* », j'ai exposé des méthodes permettant l'analyse des interactions spatiotemporelles entre objets archéologiques. Pour cela, j'ai développé trois nouveaux outils ayant pour objectif de positionner, dans une base de données, les objets contenus selon l'aspect temporel de l'information ou selon la représentativité des objets. Les deux premiers outils présentés explorent les données spatiotemporelles imparfaites, le troisième permet de les visualiser.

Le premier outil, détaillé dans le chapitre 7, propose de construire un graphe orienté pondéré à partir de l'indice d'antériorité afin d'étudier le positionnement temporel des objets dans la base. À partir de ce positionnement, j'ai réparti les objets selon trois classes (« plutôt antérieur », « plutôt postérieur » et « anté-postérieur ») et j'ai aussi extrait les objets suivants : le plus antérieur, le plus postérieur et le médian temporellement. Les classes d'objets ainsi définies et les objets extraits sont précieux pour l'étude de la structure des données. La seconde technique, exposée dans le chapitre 8, considère la représentativité de chaque objet afin d'en dégager les objets les plus représentatifs. J'ai recherché dans ce chapitre, dans la base des données sur les rues de Reims datant de l'époque romaine, les tronçons de rues les plus représentatifs pour la localisation, pour l'orientation, et pour le temps. Ces derniers forment des représentants réels de l'échantillon et sont utiles à la classification des données disponibles. L'apport de ces travaux n'est pas à visée purement applicative, il participe à un thème scientifique plus large de recherche du meilleur représentant d'un échantillon de données. Ce thème est d'une importance cruciale en analyse de données et en statistiques. Le dernier outil, développé dans le chapitre 9, permet, par une approche orientée-pixel, la visualisation par une image couleur des objets archéologiques. Cette approche m'a permis de visualiser par une image couleur les objets selon les représentations de leurs périodes d'activité, ainsi que selon leurs dissimilarités à un objet sélectionné dans la base. Pour cela, il fut nécessaire d'évaluer par des vecteurs quantitatifs, en fonction de la recherche, soit les représentations des périodes d'activité des objets

de *BDFRues*, soit les dissimilarités à un objet particulier. Ces visualisations fournissent des cartes synthétiques de l'information observée. À ce titre, elles constituent des outils exploratoires des masses de données qui sont destinés à faire émerger des rapprochements originaux ou des connaissances nouvelles. Elles complètent donc la technique mise en place pour déterminer les meilleurs représentants d'un échantillon.

Les outils précédents ont été présentés à la communauté scientifique à travers trois communications dans des conférences d'audience nationale : GRETSI'07, EGC'08 ainsi que LFA-08. Les deux premiers outils exposés m'ont permis de compléter ma réponse à la deuxième question du sujet de ces recherches. De plus, on peut remarquer que les différentes méthodes d'analyse ont fourni des résultats visualisés par des cartes thématiques dans le SIG. Ces cartes thématiques facilitent, à l'instar de l'image résultant de la technique de visualisation, la compréhension des données par l'utilisateur. J'ai ainsi proposé une réponse à la troisième question.

Cette thèse a donc permis d'élaborer une démarche pour la modélisation, des méthodes pour l'analyse et une approche à la visualisation de données spatiotemporelles imparfaites dans un SIG, à l'exemple des données archéologiques. Les travaux effectués contribuent à une meilleure gestion des imperfections dans l'information archéologique tant pour sa représentation que pour son traitement. Pour cela, les concepts théoriques de taxonomie de l'imperfection dans le contexte archéologique, de représentation des données imparfaites, d'indice d'antériorité, d'accumulation dans un espace des paramètres, d'appariement, de rangs, de graphes orientés pondérés, de représentativité, de visualisation par une image couleur ont été mis en œuvre et rapportés à leur contexte scientifique très général en informatique.

J'ai proposé dans ce mémoire une approche interdisciplinaire se plaçant dans des thématiques informatiques, une thématique géographique et une thématique commune aux deux sciences. Pour l'informatique, les problèmes abordés sont celui de la gestion des connaissances imparfaites et celui de l'analyse de données spatialisées. Pour la géographie, la thématique porte sur l'analyse des données affectées de changements d'échelles. Le domaine commun à l'informatique et la géographique au coeur de ce travail de thèse est la représentation de l'information spatiotemporelle dans un SIG. Les démarches, les méthodes, les outils et les techniques exposés dans ce mémoire contribuent à ces différentes thématiques.

Les travaux présentés dans cette thèse ont été exposés aux différentes communautés scientifiques englobant ce travail : la communauté de l'intelligence artificielle, la communauté du traitement des images, celle des sciences de l'information géographique, et celle des archéologues s'intéressant aux outils informatiques.

Perspectives

J'ai donc, durant cette thèse, pris le parti de penser les informations archéologiques dans leur complexité en tenant compte du temps, de l'espace et de l'imperfection. Cependant, les recherches menées ont été effectués en prenant pour cas d'étude la base de données *BDRues* contenant relativement peu de données. Il serait donc intéressant de travailler sur de nouvelles données afin de rendre plus opérationnels les travaux présentés.

Durant mes recherches, j'ai été amené à étudier l'appariement de données archéologiques. J'ai, pour cela, proposé l'utilisation de critères propres à l'informations archéologiques. Il serait important d'avoir plusieurs études d'une même réalité archéologique afin de vérifier l'apport de mes considérations pour l'appariement. C'est pour cela que j'envisage d'étudier l'évolution des voiries rémoises au cours des époques grâce à l'appariement des données de *BDFRues* avec des données décrivant le Moyen-Âge et celles issues des parcellaires napoléonien et actuel. Cette démarche nécessitera d'utiliser des techniques de traitement d'images pour récupérer les rues du parcellaire napoléonien et de modéliser les données archéologiques datant du Moyen-Âge en fonction de leurs imperfections.

On peut aussi s'orienter vers le développement d'une méthode de visualisation schématique des villes selon l'influence interne des activités selon la période étudiée : c'est à dire une approche chrono-chorème à l'instar de celle proposée dans [Rodier et Galinié, 2006]. La méthode étudierait alors l'influence des secteurs d'activités (religieux, commerçant, administratif) dans la ville. Ces influences seraient alors déterminées à l'aide des données issues de fouilles archéologiques ou d'ouvrages. Dans l'approche de [Rodier et Galinié, 2006], ni l'imperfection de l'information ni les modèles de diffusions des influences dans une ville ne sont pris en compte. La méthode de schématisation automatique devra inévitablement y répondre. Cette méthode ferait donc appel à l'archéologie pour les modèles d'influence, la géographie pour la généralisation, le traitement d'images pour la visualisation et l'intelligence artificielle pour la gestion des connaissances et la modélisation des données.

Dans un soucis d'exploiter et de diffuser au mieux les nouvelles connaissances, issues des méthodes présentées dans ce mémoire, et de travailler sur les données modélisées avec leurs imperfections, des efforts devront être effectués pour la formation des utilisateurs à ces nouvelles approches de l'information.

Enfin, cette thèse a ouvert une collaboration entre le CReSTIC (laboratoire d'informatique, EA-3804, URCA) et Habiter (laboratoire de géographie, EA-2076, URCA) qui se développe actuellement autour d'autres projets. L'un d'entre eux, le projet *AQUAL*, porte sur l'étude des pollutions diffuses dans le bassin de la Vesle. Dans ce projet, une attention particulière est portée à la gestion des données récoltées dans un SIG. Les approches méthodologiques et techniques proposées dans ce travail seraient d'un apport essentiel dans l'évaluation des pollutions diffuses.

Annexes

Annexe A

Indice d'antériorité et comparaison de nombres flous

Cette annexe est dédiée à l'utilisation de l'indice d'antériorité pour comparer deux nombres flous. Pour deux nombres flous F et G, l'indice d'antériorité $I(F,G)$, indique que F est « plus petit que » G avec une confiance de $I(F,G)$. L'indice d'antériorité de (G,F), $I(G,F)$, indique que G est « plus petit que » F avec une confiance $I(G,F)$. Par construction, on a $I(G,F) = 1 - I(F,G)$, excepté si $F = G$.

Le fait que $I(F,G) \geq I(G,F)$ veut donc dire que la confiance pour le fait que F est « plus petit que » G est supérieure à celle pour le fait G est « plus petit que » F. Dans ce cas, il apparait logique de dire que F est plus petit que G, la confiance en ce résultat ne sera que de $I(F,G)$. On notera la relation ainsi défini de la manière suivante : $F \preccurlyeq_{I(F,G)} G$.

Ainsi, on peut définir la position relative de F et G selon les valeurs de l'indice. Soit λ la valeur maximale de l'indice pour les couples (F,G) et (G,F), c'est à dire

$$\lambda = max(I(F,G), I(G,F))$$

En utilisant λ, on définit la relation valuée F est « plus petit que » G comme suit :

$$(F \preccurlyeq_{I(F,G)} G) \Leftrightarrow (I(F,G) > I(G,F)),$$

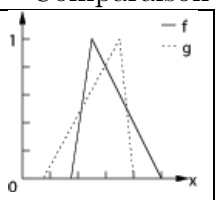
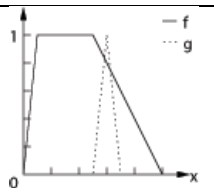
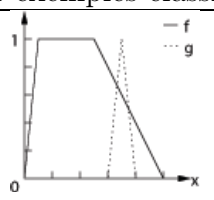
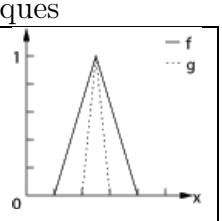
Ce qui donne :

- $G \preccurlyeq_{I(G,F)} F \Leftrightarrow 0.5 \leq I(G,F) \leq 1$,
- $F \preccurlyeq_{I(F,G)} G \Leftrightarrow 0.5 \leq I(F,G) \leq 1$.

Ainsi, si et seulement si $F = G$ alors $G \preccurlyeq_1 F$ et $F \preccurlyeq_1 G$. La relation ainsi définie n'est malheureusement pas transitive et ce n'est donc pas une relation d'ordre sur l'ensemble des nombres flous.

Il existe de nombreuses méthodes de comparaison de nombres flous. Je propose de comparer sur quatre exemples classiques (voir Tableau A.1) issus de [Bortolan et Degani, 1985] les résultats de la comparaison utilisant les indices d'antériorité avec des méthodes clas-

Tableau A.1 – Comparaison ensembles flous : exemples classiques

	 (f)	 (g)	 (h)	 (l)
Methodes				
Relation proposée	$F \preccurlyeq_{0.54} G$	$F \preccurlyeq_{0.81} G$	$F \preccurlyeq_{0.92} G$	$F \preccurlyeq_{0.50} G$ & $G \preccurlyeq_{0.50} F$
Adamo, $\alpha = 0.9$	$F \prec G$	$F \prec G$	$F \prec G$	$G \prec F$
Adamo, $\alpha = 0.5$	$F \sim G$	$G \prec F$	$F \sim G$	$G \prec F$
Adamo, $\alpha = 0.1$	$G \prec F$	$G \prec F$	$G \prec F$	$G \prec F$
FR	$F \sim G$	$F \prec G$	$F \prec G$	$F \sim G$
Kerre	$F \prec G$	$F \prec G$	$F \prec G$	$F \sim G$
Chang	$G \prec F$	$G \prec F$	$G \prec F$	$G \prec F$
PSD,ND,PD **PSD,PD,ND**	$G \prec F$	$G \prec F$	$F \prec G$	$G \prec F$
ND,PSD,PD **ND,PD,PSD**	$G \prec F$	$F \prec G$	$F \prec G$	$F \prec G$
PD,ND,PSD	$F \prec G$	$F \prec G$	$F \prec G$	$F \prec G$
PD,PSD,ND	$F \prec G$	$F \prec G$	$F \prec G$	$G \prec F$

siques de classement de nombres flous. Pour cela j'utilise les indices de comparaison suivants :

- l'indice d'Adamo, issu de [Adamo, 1985], donne la borne supérieure d'une α-coupe,
- l'indice de Fortemps et Roubens (FR), issu de [Fortemps et Roubens, 1996], renvoie la somme des bornes supérieures et inférieures de toutes les α-coupes,
- l'indice de Kerre, issu de [Kerre, 1982], est défini dans le chapitre 4,
- l'indice de Chang, issu de [Chang, 1981], renvoie la somme des produits, pour tout x du support, $x \times$ valeur d'appartenance du nombre flou étudié pour x,
- les trois indices de Dubois et Prade, issus de [Dubois et Prade, 1983], que l'on combine pour la comparaison : PD est la possibilité de dominance, PSD est la possibilité stricte de dominance, et ND est la nécessité de dominance.

À l'aide de ces indices, des méthodes de comparaison sont comparées dans le tableau A.1. On observe que les résultats des différentes méthodes sont souvent contradictoires. L'utilisation des indices d'antériorité sur les relations permet de relativiser les décisions. Elle quantifie donc la confiance que l'on peut avoir sur la décision. Cette utilisation donne aux experts des informations permettant de mettre en exergue la relativité de leurs interprétations temporelles des objets de fouilles archéologiques.

Annexe B

Détection de contours en traitement numérique des images

Cette annexe a pour objet la présentation de méthodes classiques pour la détection de contours dans une image : le filtre de Roberts, le filtre de Sobel et le filtre de Prewitt.

Ces filtres utilisent un masque de dimension $k * k$ qui est appliqué à tous les pixels (x, y) d'une image I sur leur voisinage $k \times k$, x et y étant les coordonnées du pixel. Chaque masque et chaque pixel (x, y) de valeur $I(x, y)$ permettent de calculer la valeur $I_{filtre}(x, y)$ dans l'image filtrée I_{filtre}.

Le filtre de Roberts utilise les deux masques $\frac{dI}{dx}$ et $\frac{dI}{dy}$ pour un voisinage 2×2 suivants :

$$\frac{dI}{dx} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \quad \text{et} \quad \frac{dI}{dy} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}.$$

La valeur $I_{Roberts}(x, y)$ d'un pixel (x, y) après l'utilisation d'un filtre de Roberts est la valeur maximale des valeurs issues des masques $\frac{dI}{dx}$ et $\frac{dI}{dy}$, c'est à dire :

$$I_{Roberts}(x, y) = \max\left(\frac{dI}{dx}(x, y)\right), \left(\frac{dI}{dy}(x, y)\right).$$

Le filtre de Sobel utilisent les deux masques $\frac{dI}{dx}$ et $\frac{dI}{dy}$ pour un voisinage 3×3 définis comme suit :

$$\frac{dI}{dx} = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} /4 \quad \text{et} \quad \frac{dI}{dy} = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} /4.$$

La valeur $I_{Sobel}(x, y)$ d'un pixel (x, y) après l'utilisation d'un filtre de Sobel résulte de la racine carrée de la somme des carrés des valeurs obtenues à l'aide des masque $\frac{dI}{dx}$ et $\frac{dI}{dy}$,

c'est à dire :

$$I_{Sobel}(x,y) = \sqrt{(\frac{dI}{dx}(x,y))^2 + (\frac{dI}{dy}(x,y))^2}.$$

Le filtre de Prewitt utilise, sur un voisinage 3×3, les quatre masques M_1, M_2, M_3 et M_4 suivants :

$$M_1 = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix} /3, \quad M_2 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & -1 \end{bmatrix} /3,$$

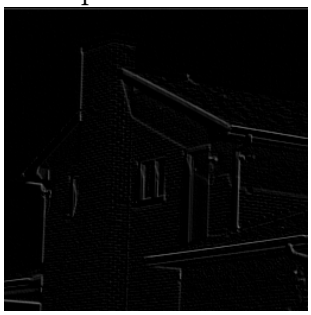
$$M_3 = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix} /3, \quad M_4 = \begin{bmatrix} 0 & 1 & 1 \\ -1 & 0 & 1 \\ -1 & -1 & 0 \end{bmatrix} /3.$$

La valeur $I_{Prewitt}(x,y)$ d'un pixel (x,y) après l'utilisation d'un filtre de Prewitt est obtenue par la valeur absolue maximale des valeurs obtenues à l'aide des quatre masques précédents :

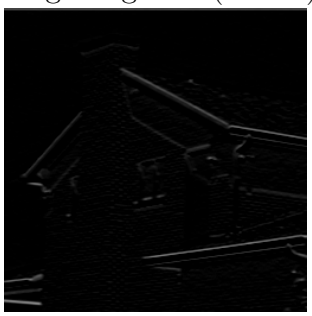
$$I_{Prewitt}(x,y) = \max_{i \in [1;4]} |M_i(x,y)|.$$

Figure B.1 – Illustration de méthodes classiques de détection de contours

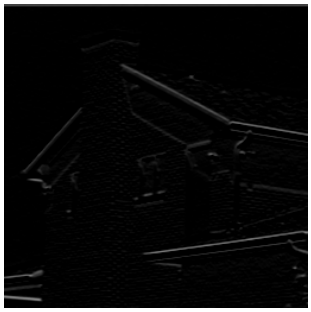
Image originale (House)



Filtre de Roberts



Filtre de Prewitt



Filtre de Sobel

Les images résultantes de l'application de ces trois filtres sur l'image House [74] sont présentées dans la figure B.1.

74. House ou House-garden : image classique en traitement d'image

Ces filtres renvoient une image des contours de l'image originale. La conversion des images résultats des filtrages en images binaires se fait usuellement à l'aide de seuillages suivis d'une fermeture des contours, en général par hystérésis.

Annexe C

Filtrage d'images couleurs : comparaison Vecteur de Meilleur Rang Moyen - Vecteur Médian

L'objet de cette annexe est l'étude comparative entre le filtre construit grâce au Vecteur Médian [Astola *et al.*, 1990] et le filtre basé sur le Vecteur de Meilleur Rang Moyen pour le débruitage d'images couleurs. Les statistiques de rangs sont classiquement utilisées en filtrage d'images [Pitas et Tsakalides, 1991]. Elles donnent lieu à de nombreuses applications en débruitage d'images couleurs à l'instar de [Lukac *et al.*, 2006] et [Vautrot *et al.*, 2006].

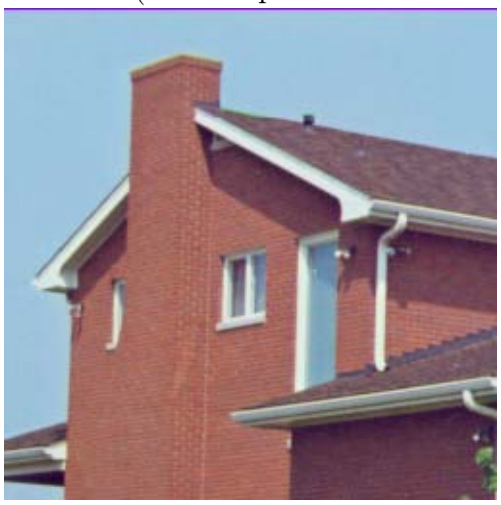
Dans le contexte du filtrage d'images en couleur, il suffit de considérer chaque voisinage où l'on opère le filtrage, comme l'échantillon S dont on cherche un représentant par notre statistique. Autrement dit, lorsque l'on applique le filtre sur un voisinage $t \times t$, on considère l'échantillon dont les éléments sont t^2 pixels de ce voisinage (ou fenêtre), avec t entier naturel impair. On remplace alors le pixel central de cette fenêtre par le pixel le plus représentatif, c'est à dire le pixel correspondant au vecteur de meilleur rang moyen dans cet échantillon. On nommera *VMRM* le filtre ainsi construit.

Afin de montrer l'efficacité de cette statistique, je la compare à une méthode statistique reconnue comme efficace dans ce contexte — le vecteur médian [Astola *et al.*, 1990] — sur les images House [75] et Lena [76].

Le filtre Vecteur Median [Astola *et al.*, 1990] est classiquement utilisé pour l'élimination du bruit d'images couleurs. On a utilisé un jeu classique d'images sur lesquelles du bruit impulsionnel a été préalablement ajouté. Pour des fenêtres 5×5, l'efficacité de notre filtrage est similaire à celle issue du filtre Vecteur Median tant du point visuel (Figure C.1c et Figure C.1d), que numérique. Le tableau C.1 donnent l'erreur quadratique moyenne (MSE : *Mean Square Error*), l'erreur absolue moyenne (MAE : *Mean Absolute Error*), ainsi que l'erreur maximale (ME : *Max Error*).

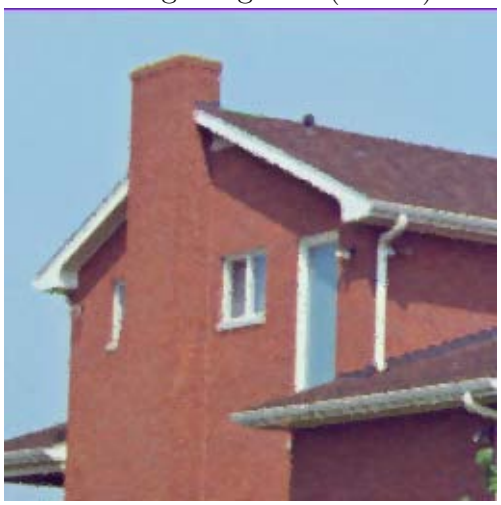
75. House ou House-garden : image classique en traitement d'image
76. Lena : Image classique en traitement d'image, voir [Munson, 1996]

Figure C.1 – Comparaison visuelle par filtrage d'une image couleur du VMRM et du Vecteur Médian (bruit impulsionnel 5% - voisinage 5×5)



a : image originale (House)

b : image bruitée



c : après VMRM

d : après Vecteur Median

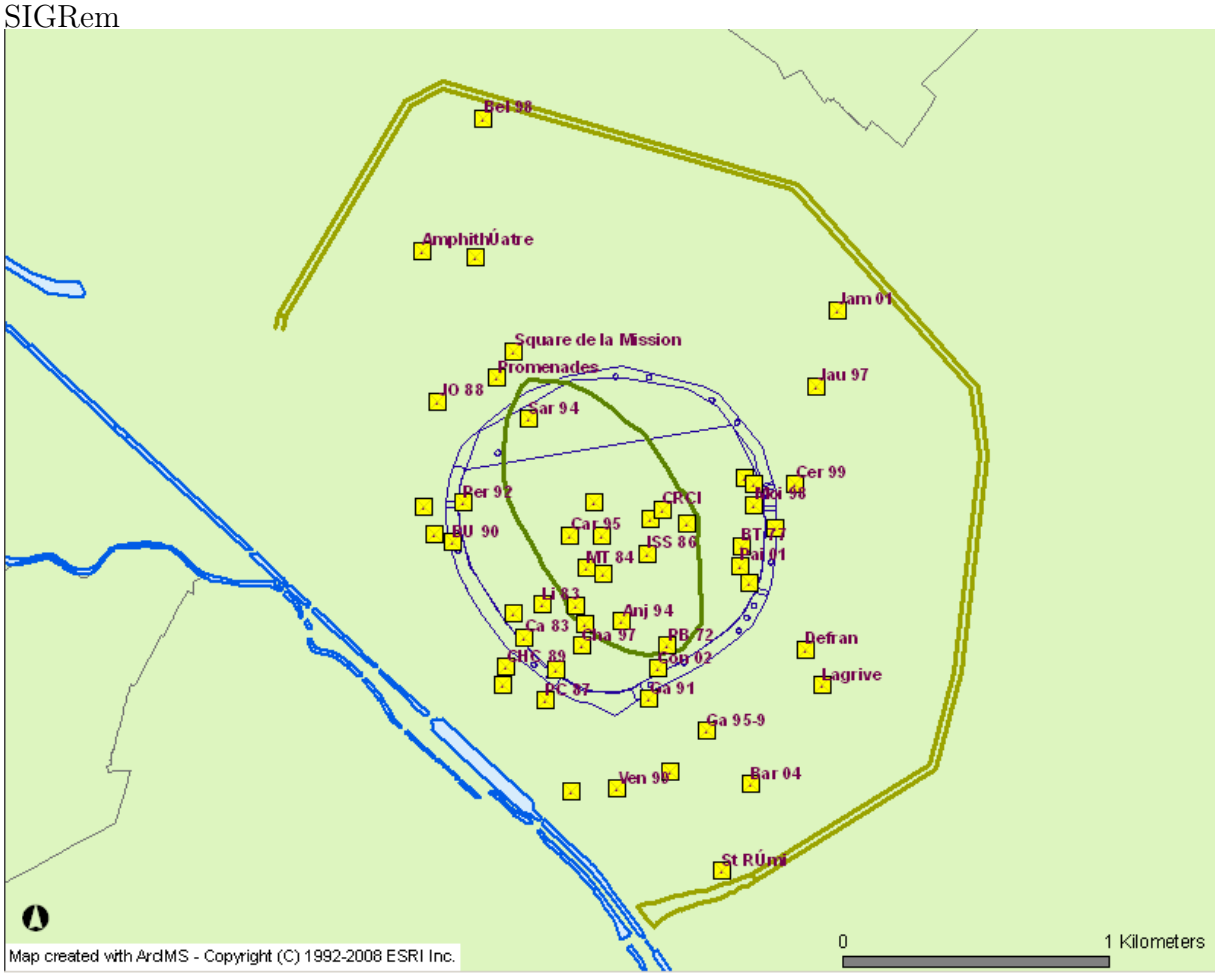
Tableau C.1 – Comparaison numérique pour du bruit impulsionnel du filtre Vecteur Median et Filtre VMRM sur un voisinage 3×3

Image	Bruit (%)	MAE V. Médian	MAE VMRM	MSE V. Médian	MSE VMRM	ME V. Médian	ME VMRM
House	0	5.16	5.10	97.41	90.62	107.67	105.33
House	5	5.17	5.08	95.87	86.95	107.67	105.33
House	10	5.18	5.32	96.32	97.16	109.67	109.67
House	20	5.28	6.01	98.19	120.77	115.00	115.00
Lena	0	8.11	8.03	223.03	212.36	146.67	146.67
Lena	5	8.16	8.15	223.81	212.13	146.67	146.67
Lena	10	8.24	8.39	224.93	216.80	146.67	146.67
Lena	20	8.45	9.18	229.08	243.74	151.00	120.00

Annexe D

Sites de fouilles référencés dans le programme SIGRem

Figure D.1 – Cartes des sites de fouilles archéologiques référencés dans le programme SIGRem



Publications personnelles

Soit :
- 1 article dans une revue nationale [1],
- 4 communications internationales [2, 3, 4, 5],
- 3 communications nationales [6, 7, 8],
- 1 communications nationales sans comité de lecture [9],

J'ai de même soumis
- 2 articles dans des revues internationales [10, 11],

[1] C. de Runz, E. Desjardin, M. Herbin, D. Pargny, F. Piantoni et F. Berthelot, « Simulation de cartes et prédictivité à partir de données archéologiques traitées par SIG », *Culture et Recherche*, vol. 111, 2007, p. 35.

[2] C. de Runz, E. Desjardin, M. Herbin et F. Piantoni, « A new Method for the Comparison of two fuzzy numbers extending Fuzzy Max Order », dans *Information Processing and Managment of Uncertainty in Knowledge-Based Systems - IPMU'06*, p. 127–133, Paris, France, 2006. Editions EDK.

[3] C. de Runz, E. Desjardin, F. Piantoni et M. Herbin, « Management of multimodal data using the Fuzzy Hough Transform : Application to archaeological simulation », dans C. Rolland, O. Pastor et J.-L. Cavarero, éditeurs, *First International Conference on Research Challenges in Information Science*, p. 351–356, Ouarzazate, Maroc, 2007.

[4] C. de Runz, E. Desjardin, F. Piantoni et M. Herbin, « Using fuzzy logic to manage uncertain multi-modal data in an archaeological GIS », dans *International Symposium on Spatial Data Quality - ISSDQ'07*, Enschede, Pays-Bas, 2007.

[5] C. de Runz, E. Desjardin, F. Piantoni et M. Herbin, « Toward handling uncertainty of excavation data into a GIS », dans *36th Annual Conference on Computer Applications and Quantitative Methods in Archaeology - CAA*, Budapest, Hongrie, 2008.

[6] C. de Runz, M. Herbin, F. Blanchard, L. Hussenet, V. Vrabie et P. Vautrot, « Le vecteur de meilleur rang moyen : une statistique pour l'analyse de données multidimensionnelles - Application au filtrage d'images couleurs », dans *GRETSI*, p. 97–100, Troyes, France, 2007.

[7] C. de Runz, F. Blanchard, E. Desjardin et M. Herbin, « Fouilles archéologiques : à la recherche d'éléments représentatifs », dans *Atelier Fouilles de Données Complexes - Conférence Extraction et Gestion des Connaissances - EGC'08*, p. 95–103, Sophia Antipolis, France, 2008.

[8] C. de Runz, F. Blanchard, E. Desjardin et M. Herbin, « Exploration d'un ensemble de quantités floues », dans *Logique floue et ses applications - LFA'08*, 2008. soumis le 15 mai.

[9] C. de Runz, D. Pargny, E. Desjardin, M. Herbin et F. Piantoni, « Aide à la décision en archéologique préventive : Les rues de la Cité des Rèmes », dans *Conférence Francophone ESRI*, Issy les Moulineaux, France, 2006.

[10] C. de Runz, E. Desjardin, F. Piantoni et M. Herbin, « Management of locations, dates and shapes in archaeological GIS », *GeoInformatica*, 2008. Soumis le 28 septembre 2007, en évaluation depuis le 8 octobre 2007.

[11] C. de Runz, E. Desjardin, F. Piantoni et M. Herbin, « Anteriority index for managing fuzzy dates in archaeological GIS », *Soft Computing*, 2008. Soumis le 30-10-2007, en révisions majeures depuis le 09-06-2008.

Bibliographie

« The relevant equation is :
Knowledge = power = energy = matter = mass ; a good bookstore is just a genteel Black Hole that knows how to read. »

Terry Pratchett, *Annals of the DiscWorld — Guards ! Guards !* (1989).

[Abadie *et al.*, 2007] Nathalie Abadie, Ana-Maria Olteanu, et Sébastien Mustière. Comparaison de la nature d'objets géographiques. Dans *Ontologies et Gestion de l'Hétérogénéité Sémantique*, Grenoble, 2007. Plate-forme AFIA. (Cité p. 75 et 113)

[Accary *et al.*, 2003] Tiphaine Accary, Aurélien Bénel, et Sylvie Calabretto. Modélisation de connaissances temporelles en archéologie. *Extraction des connaissances et apprentissage*, 17(1) :503–508, 2003. (Cité p. 56)

[Adamo, 1985] J.M. Adamo. Fuzzy decision trees. *Fuzzy Sets and Systems*, 4 :207–219, 1985. (Cité p. 187)

[Astola *et al.*, 1990] Jaakko Astola, Petri Haavisto, et Yrjo Neuvo. Vector median filters. *Proceedings of the IEEE*, 78(4) :678–689, 1990. (Cité p. 156 et 191)

[Barnett, 1976] V. Barnett. The Ordering of Multivariate Data. *Journal of the Royal Statistical Society, Series A (General)*, 139(3) :318–355, 1976. (Cité p. 147 et 154)

[Beeri *et al.*, 2005] Catriel Beeri, Yerach Doytsher, Yaron Kanza, Eliyahu Safra, et Yehoshua Sagiv. Finding corresponding objects when integrating several geo-spatial datasets. Dans *GIS '05 : Proceedings of the 13th annual ACM international workshop on Geographic information systems*, pages 87–96, Bremen, Allemagne, 2005. ACM. (Cité p. 112)

[Beeri *et al.*, 2004] Catriel Beeri, Yaron Kanza, Eliyahu Safra, et Yehoshua Sagiv. Object fusion in geographic information systems. Dans *The Thirtieth international conference on Very large data bases*, Toronto, Canada, 2004. VLDB Endowment. (Cité p. 109)

[Bejaoui *et al.*, 2007] Lotfi Bejaoui, Yvan Bédard, François Pinet, Mehrdad Salehi, et Michel Schneider. Logical consitency for vague spatioteporal objects and relations. Dans *International Symposium on Spatial Data Quality - ISSDQ'07*, Enschede,Pays-Bas, 2007. (Cité p. 13, 56 et 57)

[Bel Hadj Ali, 2001] Atef Bel Hadj Ali. *Qualité géométrique des entités géographiques surfaciques — Application à l'appariement et définition d'une typologie des écarts géométriques*. PhD thesis, Université de Marne-La-Vallée, 2001. (Cité p. 109)

[Bellman, 1961] R. Bellman. *Adaptive control processes : a guide tour*. Princeton University Press, 1961. (Cité p. 162)

[Béguin et Pumain, 2003] Michel Béguin et Denise Pumain. *La représentation des données géographiques - statistique et cartographie*. Cursus. Armand Colin, 2003. (Cité p. 7 et 8)

[Blanchard, 2005] Frédéric Blanchard. *Visualisation et classification de données multidimensionnelles ;Application aux images multicomposantes*. Thèse de doctorat, Université de Reims Champagne-Ardenne, France, Reims, France, 2005. (Cité p. 19, 128, 146, 147, 148 et 175)

[Blanchard et Herbin, 2004] Frédéric Blanchard et Michel Herbin. L'image couleur pour visualiser des données multidimensionelles. *Traitement du Signal*, 21 :453–460, 2004. (Cité p. 160 et 161)

[Blanchard *et al.*, 2005a] Frédéric Blanchard, Michel Herbin, et Laurent Lucas. A New Pixel-Oriented Visualization Technique Through Color Image. *Information Visualization*, 4(4) :257–265, 2005. (Cité p. 16, 19, 129, 161 et 162)

[Blanchard *et al.*, 2005b] Frédéric Blanchard, Michel Herbin, et Françis Rousseaux. Compendium de données multidimensionnelles par une image couleur. Dans *EGC 2005 Atelier Visualisation et extraction de connaissances*, Paris, 2005. (Cité p. 19)

[Bloch, 1996] Isabelle Bloch. Incertitude, imprécision et additivité en fusion de données : point de vue historique. *Traitement du Signal*, 13(4) :267–288, 1996. (Cité p. 38)

[Bonnet, 2002] Noël Bonnet. An unsupervised generalized Hough transform for natural shapes. *Pattern Recognition*, 35 :1192–1196, 2002. (Cité p. 96)

[Bordin, 2002] Patricia Bordin. *SIG - concepts, outils, et données*. Hermes, 2002. (Cité p. 6, 7 et 8)

[Bortolan et Degani, 1985] Giovanni Bortolan et Rosanna Degani. A review of some methods for ranking fuzzy subsets. *Fuzzy Sets and Systems*, 15 :1–19, 1985. (Cité p. 186)

[Bouchon-Meunier, 1993] Bernadette Bouchon-Meunier. *La logique floue*. QUE SAIS-JE ? PUF, 1993. (Cité p. 11, 26, 45 et 46)

[Bouchon-Meunier, 1995] Bernadette Bouchon-Meunier. *La logique floue et ses applications*. EddAddison-Wesley, 1995. (Cité p. 12, 26, 45 et 46)

[Bow, 2002] Sing-Tze Bow. *Pattern Recognition and Image Preprocessing*. Marcel Dekker Inc., 2 edition, 2002. (Cité p. 95)

[Brown, 1992] Lisa G. Brown. A survey of image registration techniques. *ACM Computing Surveys*, 24(4) :325–376, 1992. (Cité p. 112)

[Burrough et Frank, 1996] Peter A. Burrough et Andrew U. Frank, Éditeurs. *Geographic objects with indeterminate boundaries.* GISDATA. Taylor & Francis, 1996. (Cité p. 13, 62, 199 et 202)

[Burrough et McDonnell, 1998] Peter A. Burrough et Rachael McDonnell. *Principle of Geographical Information Systems.* Oxford University Press, 1998. (Cité p. 3, 6 et 7)

[Canny, 1986] John F. Canny. A Computational Approach To Edge Detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8 :679–714, 1986. (Cité p. 95)

[Cardoso et Comon, 1996] Jean-François Cardoso et Pierre Comon. Independent Component Analysis, a Survey of Some Algebraic methods. Dans *ISCAS Conference*, volume 2, pages 93–96, 1996. (Cité p. 19, 162 et 163)

[Chang, 1981] W. Chang. Ranking of fuzzy utilities with triangular membership functions. Dans *Proceedings of International Conference in Policy Analysis and System*, pages 263–272, 1981. (Cité p. 187)

[Chen, 1985] Shan-Huo Chen. Ranking fuzzy numbers with maximizing set and minimizing set. *Fuzzy Sets and Systems*, 17 :113–129, 1985. (Cité p. 135)

[Cheylan *et al.*, 1999] Jean-Paul Cheylan, Denis Gautier, Sylvie Lardon, Thérèse Libourel, Hélène Mathian, Serge Motet, et Lenna Sanders. Les mots du traitement de l'information spatio-temporelle. *Revue Internationale de Géomatique*, 9(1) :11–23, 1999. (Cité p. 9)

[Choquet, 1954] Gustave Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5 :131–295, 1954. (Cité p. 101)

[Clementini et Di Felice, 1996] Eliso Clementini et Paolino Di Felice. An Algebraic Model for Spatial Objects with Indeterminate Boundaries. Dans [Burrough et Frank, 1996], Chapitre 11, pages 155–170. (Cité p. 13)

[Conolly et Lake, 2006] James Conolly et Mark Lake. *Geographic Information System in Archaeology.* Cambridge University Press, 2006. (Cité p. 4)

[Couclelis, 2005] Helen Couclelis. Space, Time, Geography. Dans [Longley *et al.*, 2005b], Chapitre 2, pages 29–38. Seconde Edition. (Cité p. 62)

[Cowen, 1998] David J. Cowen. GIS versus CAD versus DBMS : what are the differences ? *Photogrammetric Engineering & Remote Sensing*, 54(11) :1551–1555, 1998. (Cité p. 7)

[David et Nagaraja, 2003] Herbert A. David et Haikady N. Nagaraja. *Order Statistics.* Wiley, third edition, 2003. (Cité p. 120, 147 et 156)

[de Ruffray, 1999] Sophie de Ruffray. Mise en évidence de l'organisation spatiale d'un espace de marges. L'exemple de l'interface Moselle Est-Alsace du Nord-Ouest. *Mosella*, XXIV(3-4) :41–64, 1999. (Cité p. 28)

[de Ruffray, 2004] Sophie de Ruffray. Le Grand Est : un espace différencié, interface marginale aux portes de l'Europe. *Revue Géographique de l'Est*, XLIV(3-4) :97–106, 2004. (Cité p. 28)

[De Ruffray, 2007] Sophie De Ruffray. The application of fuzzy logic to school districting : the example of Mozelle, France. Dans *Applied Geography Conference*, 2007. (Cité p. 63)

[de Runz *et al.*, 2008a] Cyril de Runz, Frédéric Blanchard, Eric Desjardin, et Michel Herbin. Exploration d'un ensemble de quantités floues. Dans *Logique floue et ses applications - LFA'08*, 2008. (Cité p. 173 et 175)

[de Runz *et al.*, 2008b] Cyril de Runz, Frédéric Blanchard, Eric Desjardin, et Michel Herbin. Fouilles archéologiques : à la recherche d'éléments représentatifs. Dans *Atelier Fouilles de Données Complexes - Conférence Extraction et Gestion des Connaissances - EGC'08*, pages 95–103, Sophia Antipolis, France, 2008. (Cité p. 158 et 175)

[de Runz *et al.*, 2007a] Cyril de Runz, Eric Desjardin, Michel Herbin, Dominique Pargny, Frédéric Piantoni, et François Berthelot. Simulation de cartes et prédictivité à partir de données archéologiques traitées par SIG. *Culture et Recherche*, 111 :35, 2007. (Cité p. 106 et 107)

[de Runz *et al.*, 2006] Cyril de Runz, Eric Desjardin, Michel Herbin, et Frédéric Piantoni. A new Method for the Comparison of two fuzzy numbers extending Fuzzy Max Order. Dans *Information Processing and Managment of Uncertainty in Knowledge-Based Systems - IPMU'06*, pages 127–133, Paris, France, 2006. Editions EDK. (Cité p. 91 et 123)

[de Runz *et al.*, 2007b] Cyril de Runz, Eric Desjardin, Frédéric Piantoni, et Michel Herbin. Management of multi-modal data using the Fuzzy Hough Transform : Applicationto archaeological simulation. Dans Colette Rolland, Oscar Pastor, et Jean-Louis Cavarero, Éditeurs, *First International Conference on Research Challenges in Information Science*, pages 351–356, Ouarzazate, Maroc, 2007. (Cité p. 106, 123 et 107)

[de Runz *et al.*, 2007c] Cyril de Runz, Eric Desjardin, Frédéric Piantoni, et Michel Herbin. Using fuzzy logic to manage uncertain multi-modal data in an archaeological GIS. Dans *International Symposium on Spatial Data Quality - ISSDQ'07*, Enschede, Pays-Bas, 2007. (Cité p. 106, 123 et 107)

[de Runz *et al.*, 2008c] Cyril de Runz, Eric Desjardin, Frédéric Piantoni, et Michel Herbin. Anteriority index for managing fuzzy dates in archaeological GIS. *Soft Computing*, 2008. Soumis le 30-10-2007, en révision majeure depuis le 09-06-2008. (Cité p. 91, 123 et 181)

[de Runz *et al.*, 2008d] Cyril de Runz, Eric Desjardin, Frédéric Piantoni, et Michel Herbin. Management of locations, dates and shapes in archaeological GIS. *GeoInformatica*, 2008. Soumis le 28 septembre 2007, en évaluation depuis le 8 octobre 2007. (Cité p. 107, 123 et 181)

[de Runz *et al.*, 2008e] Cyril de Runz, Eric Desjardin, Frédéric Piantoni, et Michel Herbin. Toward handling uncertainty of excavation data into a GIS. Dans *36th Annual*

Conference on Computer Applications and Quantitative Methods in Archaeology - CAA, Budapest, Hongrie, 2008. (Cité p. 67)

[de Runz *et al.*, 2007d] Cyril de Runz, Michel Herbin, Frédéric Blanchard, Laurent Hussenet, Valeriu Vrabie, et Philippe Vautrot. Le vecteur de meilleur rang moyen : une statistique pour l'analyse de données multidimensionnelles - Application au filtrage d'images couleurs. Dans *GRETSI*, pages 97–100, Troyes, France, 2007. (Cité p. 158 et 175)

[de Wit et de Bruin, 2006] Allard de Wit et Sytze de Bruin. Simulating space-time uncertainty in continental-scal gridded precipitationfields for agrometeorological modelling. Dans *7th International Symposium on Spatial Accuracy in Natural Ressources and Environmental Sciences*, pages 367–376, Lisbone, Portugal, 2006. (Cité p. 64)

[Dempster, 1968] Arthur P. Dempster. A generalization of Bayesian inference. *Journal of the Royal Statistical Society*, 30(2) :205–247, 1968. (Cité p. 13, 42 et 43)

[Denègre et Salgé, 1996] Jean Denègre et François Salgé. *Les Systèmes d'Information Géographique*. QUE SAIS-JE ? PUF, 1996. (Cité p. 8 et 77)

[Deriche, 1987] Rachid Deriche. Using Canny's criteria to derive an optimal edge detector recursively implemented. *International Journal of Computer Vision*, 1 :167–187, 1987. (Cité p. 95)

[Detyniecki, 2000] Martin Detyniecki. *Mathematical aggregation operators and video querying their applicationto video querying*. PhD thesis, Université Paris 6, France, 2000. (Cité p. 100 et 101)

[Devillers et Jeansoulin, 2005] Rodolphe Devillers et Robert Jeansoulin, Éditeurs. *Qualité de l'information géographique*. Traité IGAT. Hermes, 2005. (Cité p. 11 et 202)

[Devogele, 1997] Thomas Devogele. *Processus d'intégration et d'appariement de Bases de Données Géographiques - Application à une base de données routières multi-échelles*. PhD thesis, Université de Versailles, 1997. (Cité p. 109, 110 et 112)

[Dilo *et al.*, 2004] Arta Dilo, Rolf A. de By, et Alfred Stein. Definition and implementation of vague objects. Dans Andrew U. Frank et Eva Grum, Éditeurs, *International Symposium on Spatial Data Quality - ISSDQ'04*, volume 1 of *GeoInfo*, pages 139–145, 2004. (Cité p. 62, 63 et 209)

[Donoho, 2000] D. L. Donoho. High-Dimensional Data Analysis : The Curses and Blessings of Dimensionality. Dans *AMS Conference Mathematical Challenges of the 21st Century*, 2000. (Cité p. 162)

[Dragicevic et Marceau, 2000a] Suzana Dragicevic et Danielle J. Marceau. A fuzzy set approach for modelling time in GIS. *International Journal of Geographical Information Science*, 14(3) :225–245, 2000. (Cité p. 64)

[Dragicevic et Marceau, 2000b] Suzana Dragicevic et Danielle J. Marceau. An application of fuzzy logic reasoning for GIS temporal modeling of dynamic processes. *Fuzzy Sets and Systems*, 113 :69–80, 2000. (Cité p. 56)

[Dubois *et al.*, 2003] Didier DUBOIS, Allel HADJ ALI, et Henri PRADE. Fuzzyness and Uncertainty in Temporal Reasoning. *Journal of Universal Computer Science*, 9, 9 :1168–1194, 2003. (Cité p. 64)

[Dubois et Prade, 1983] Didier DUBOIS et Henri PRADE. Ranking fuzzy numbers in the setting of possibility theory. *Information Sciences*, 30 :183–224, 1983. (Cité p. 187)

[Dubois et Prade, 1985] Didier DUBOIS et Henri PRADE. *Théorie des Possibilités. Applications à la Représentation des Connaissances en Informatique*. Collection Méthode + Programmes. Masson, 1985. (Cité p. 13 et 50)

[Dubois et Prade, 1988] Didier DUBOIS et Henri PRADE. *Possibility Theory : An Approach to Computerized Processing of Uncertainty*. Plenum Press, 1988. (Cité p. 13, 50 et 51)

[Dubois et Prade, 2004] Didier DUBOIS et Henri PRADE. On the use of aggregation operations in information fusion processes. *Fuzzy Sets and Systems*, 142 :143–161, 2004. (Cité p. 100)

[Duda et Hart, 1972] Richard O. DUDA et Peter E. HART. Use of the Hough transform to detect lines and curves in pictures. *Comm. ACM*, 15(1) :pp 11–15, 1972. (Cité p. 18, 94 et 96)

[Duda et Hart, 1973] Richard O. DUDA et Peter R. HART. *Pattern Classification And Scene Analysis*. Wiley, New York, 1973. (Cité p. 95)

[Dupin de Saint-Cyr et Prade, 2006] Florence Dupin de SAINT-CYR et Henri PRADE. Multiple-source data fusion problems in spatial information systems. Dans *Information Processing and Managment of Uncertainty in Knowledge-Based Systems - IPMU'06*, Paris, France, 2006. (Cité p. 13)

[Ecobichon, 1994] CLaude ECOBICHON. *L'information géographique - Nouvelles techniques, nouvelles pratiques*. Perspectives. Hermes, 1994. (Cité p. 3)

[Fisher, 1996] Peter FISHER. Boolean and Fuzzy Regions. Dans [Burrough et Frank, 1996], Chapitre 6, pages 87–94. (Cité p. 13 et 62)

[Fisher *et al.*, 2005] Peter FISHER, Alexis COMBER, et Richard WADSWORTH. Nature de l'incertitude pour les données spatiales. Dans [Devillers et Jeansoulin, 2005], Chapitre 4, pages 49–64. (Cité p. 13, 17, 24, 25, 27, 31, 35, 56, 57, 58, 59, 60, 65, 116 et 209)

[Fisher, 2005] Peter F. FISHER. Models of uncertainty in spatial data. Dans [Longley *et al.*, 2005b], Chapitre 13, pages 191–205. Seconde Edition. (Cité p. 13, 24, 26, 27, 32, 35, 55, 57, 58, 59, 209 et 212)

[Fortemps et Roubens, 1996] P. FORTEMPS et M. ROUBENS. Ranking fuzzy sets : a decision theoretic approach. *Fuzzy Sets and Systems*, 82 :319–330, 1996. (Cité p. 187)

[Galambos, 1975] Janos GALAMBOS. Order Statistics of Samples from Multivariate Distributions. *Journal of the American Statistical Association*, 70(351) :674–680, 1975. (Cité p. 147)

[Goodchild, 2006] Michael F. Goodchild. Foreword. Dans Rodolphe Devillers et Robert Jeansoulin, Éditeurs, *Fundamentals of Spatial Data Quality*, GIS, pages 13–16. ISTE, 2006. (Cité p. 11)

[Grzegorzewski, 1998] Przemyslaw Grzegorzewski. Metrics and orders in space of fuzzy numbers. *Fuzzy Sets and Systems*, 97 :83–94, 1998. (Cité p. 114 et 150)

[Guptill, 2005] Stephen C. Guptill. Metadata and data catalogues. Dans [Longley *et al.*, 2005b], Chapitre 49, pages 677–692. Seconde Edition. (Cité p. 129 et 160)

[Hamming, 1950] Richard W. Hamming. Error Detecting and Error Correcting Codes. *Bell System Technical Journal*, 26(2) :147–160, 1950. (Cité p. 82 et 115)

[Han *et al.*, 1993] Joon H. Han, Laszlo T. Koczy, et Timothy Poston. Fuzzy Hough Transform. Dans *Second IEEE International Conference on Fuzzy Systems*, volume 2, pages 803 – 808, 1993. (Cité p. 75, 94 et 97)

[Han *et al.*, 1994] Joon H. Han, Laszlo T. Koczy, et Timothy Poston. Fuzzy Hough Transform. *Pattern Recognition Letters*, 15 :649–648, 1994. (Cité p. 15, 18, 94 et 97)

[Healey et Enns, 1999] C. G. Healey et J. T. Enns. Large Datasets at a Glance : Combining Textures and Colors in Scientific Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 5(2) :145–167, 1999. (Cité p. 162)

[Hjelmas et Low, 2001] Erik Hjelmas et Boon Kee Low. Face Detection : A Survey. *Computer Vision and Image Understanding*, 83 :236–274, 2001. (Cité p. 95)

[Hough, 1962] Paul V. C. Hough. Method and means for recognizing complex patterns. Rapport technique, 1962. US 3 069 654. (Cité p. 18, 94, 95 et 96)

[Hyvärinen, 1999] Aapo Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2 :94–128, 1999. (Cité p. 19, 162 et 163)

[Illingworth et Kittler, 1988] J. Illingworth et J. Kittler. A survey of the Hough transform. *Information Control*, 44 :87–116, 1988. (Cité p. 96)

[Jacquemot *et al.*, 2004] Thomas Jacquemot, Jeannine Corbonnois, et Sophie de Ruffray. Hiérarchisation des processus de la dynamique fluviale par le traitement statistique des données. *Mosella*, XXIX(3-4) :363–370, 2004. (Cité p. 163)

[Jain, 1977] R. Jain. A procedure for multiple-aspect decision making using fuzzy set. *Internat. J. Systems Sci.*, 8 :1–7, 1977. (Cité p. 128, 131, 132, 134 et 140)

[Jolliffe, 1986] I. T. Jolliffe. *Principal Component Analysis.* Springer Verlag, 1986. (Cité p. 19, 162 et 163)

[Keim, 1996] Daniel A. Keim. Pixel-oriented Visualization Techniques for Exploring Very Large Databases. *Journal of Computational and Graphical Statistics*, 5(1) :58–77, 1996. (Cité p. 160)

[Keim, 2000] Daniel A. KEIM. Designing Pixel-oriented Visualization Techniques : Theory and Applications. *IEEE Transaction on Visualization and Computer Graphics (TVCG)*, 6(1) :59–78, 2000. (Cité p. 160 et 163)

[Kerre, 1982] Etienne E. KERRE. The use of fuzzy set theory in electrocardiological diagnostics. Dans M. GUPTA et E. SANCHEZ, Éditeurs, *Approximate Reasoning in Decision-Analysis*, pages 277–282. North-Holland Publishing Company, 1982. (Cité p. 18, 74, 82, 122, 128, 131, 132, 140 et 187)

[Klir et Yuan, 1995] George J. KLIR et Bo YUAN. *Fuzzy Sets and Fuzzy Logic : Theory and Application*. Prentice Hall PTR, 1995. (Cité p. 13, 35, 36, 49, 55, 56 et 209)

[Le Ber *et al.*, 2007] Florence LE BER, Gérard LIGOZAT, et Odile PAPINI, Éditeurs. *Raisonnements sur l'espace et le temps*. Traité IGAT. Hermes, 2007. (Cité p. 206 et 205)

[Leavers, 1993] V.F. LEAVERS. Which Hough Transform. *CVGIP*, 58 :250–264, 1993. (Cité p. 96)

[Levenshtein, 1966] Vladimir LEVENSHTEIN. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10 :707, 1966. (Cité p. 115)

[Liou et Wang, 1992] Tian-Shy LIOU et Mao-Jiun J. WANG. Ranking fuzzy numbers with integral value. *Fuzzy Sets and Systems*, 50(3) :247–255, 1992. (Cité p. 135)

[Lock et Stancic, 1995] Gary R. LOCK et G. STANCIC, Éditeurs. *Archaeology And Geographic Information Systems : A European Perspective*. Taylor and Francis, 1995. (Cité p. 4)

[Loncaric, 1998] Sven LONCARIC. A survey of shape analysis techniques. *Pattern Recognition*, 31(8) :983–1001, 1998. (Cité p. 95)

[Longley *et al.*, 2005a] Paul A. LONGLEY, Michael F. GOODCHILD, David J. MAGUIRE, et David W. RHIND. *Geographic Information Systems and Science*. Wiley, 2005. Seconde Edition. (Cité p. 3 et 14)

[Longley *et al.*, 2005b] Paul A. LONGLEY, Michael F. GOODCHILD, David J. MAGUIRE, et David W. RHIND, Éditeurs. *Geographical Information Systems. Principles, Techniques, Management and Applications*. Wiley, 2005. Seconde Edition. (Cité p. 3, 199, 202, 203 et 204)

[Longley *et al.*, 2005c] Paul A. LONGLEY, Michael F. GOODCHILD, David J. MAGUIRE, et David W. RHIND. Introduction. Dans [Longley *et al.*, 2005b], Chapitre 1, pages 1–20. Seconde Edition. (Cité p. 3)

[Lowen et Peeters, 1998] R. LOWEN et W. PEETERS. Distances between fuzzy sets representing grey level images. *Fuzzy Sets and Systems*, 99(2) :135–150, 1998. (Cité p. 114)

[Lukac *et al.*, 2006] Rastislav. LUKAC, Bogdan SMOLKA, Konstantinos N. PLATANIOTIS, et Anastasios N. VENETSANOPOULOS. Vector sigma filters for noise detection and removal in color images. *J. Vis. Commun. Image R.*, 17 :1–26, 2006. (Cité p. 191)

[Mantel et Lipeck, 2004] Daniela MANTEL et Udo LIPECK. Matching cartographic objects in spatial databases. Dans *XXth ISPRS Congress - Geo-Imagery Bridging Continents*, pages 172–176, Istanbul, Turkey, 2004. ISPRS Commision IV. (Cité p. 109 et 112)

[McDonnell et Kemp, 1996] Rachael MCDONNELL et Karen K. KEMP. *The International GIS Dictionary.* Wiley, 1996. (Cité p. 7)

[Mehrer et Wescott, 2005] Mark W. MEHRER et Konnie L. WESCOTT, Éditeurs. *GIS and Archaeological Site Location Modeling.* Taylor and Francis, 2005. (Cité p. 4)

[Modersitzki, 2004] Jan MODERSITZKI. *Numerical methods for image registration.* Numerical mathematics and scientific computation. Oxford University Press, 2004. (Cité p. 112)

[Moon *et al.*, 2001] B. MOON, H. V. JAGADISH, C. FALOUTSOS, et J. H. SALTZ. Analysis of the Clustering Properties of the Hilbert Space-Filling Curve. *IEEE Transactions on Knowledge and Data Engineering*, 13(1) :124–141, 2001. (Cité p. 163 et 164)

[Morin, 2005] Edgar MORIN. *Introduction à la pensée complexe.* Essais. Points, 2005. (Cité p. 10)

[Munson, 1996] David C. MUNSON. A Note on Lena. *IEEE Transactions on Image Processing,*, 5(1), 1996. (Cité p. 191)

[Navratil, 2007] Gerhard NAVRATIL. Modeling data quality with possibility-distributions. Dans *International Symposium on Spatial Data Quality - ISSDQ'07*, Enschede, Paysbas, 2007. (Cité p. 63)

[Ohta *et al.*, 1980] Yulchi OHTA, Takeo KANADE, et Toshiyuki SAKAI. Color Information for Region Segmentation. *Computer Graphics and Image Processing*, 13 :222–241, 1980. (Cité p. 19, 161, 165 et 175)

[Olteanu, 2007a] Ana-Maria OLTEANU. A multi-criteria fusion for geographical data matching. Dans *International Symposium on Spatial Data Quality - ISSDQ'07*, Enschede, 2007. (Cité p. 15, 18, 75, 109, 116, 117 et 210)

[Olteanu, 2007b] Ana-Maria OLTEANU. Appariement de données géographique utilisant la théorie des fonctions de croyance. Dans *LFA'07*, Sohpia-Antipolis, France, 2007. (Cité p. 75, 109, 116, 117 et 210)

[Olteanu *et al.*, 2006] Ana-Maria OLTEANU, Sébastien MUSTIÈRE, et Anne RUAS. Matching imperfect spatial data. Dans *7th International Symposium on Spatial Accuracy Assessment in Natural Ressources and Environmental Sciences*, pages 694–704, Lisbonne, Portugal, 2006. (Cité p. 75, 109, 112 et 113)

[Pandazis *et al.*, 1999] Jean-Charles PANDAZIS, Claus DORENBECK, Luc HERES, Monterie MARCEL, et Kees WEVERS. EVIDENCE - Algorithms and implementation guidelines - final report. PU - public usage of the result TR 4011, ERTICO and Navigation Technologies and Tele Atlas and DDoT-RWS-AVV and AND Mapping, 1999. (Cité p. 112)

[Papini, 2007a] Odile Papini. Raisonnement en logique modale. Dans [Le Ber *et al.*, 2007], Chapitre 6, pages 151–180. (Cité p. 64)

[Papini, 2007b] Odile Papini. Représentation en logique modale. Dans [Le Ber *et al.*, 2007], Chapitre 3, pages 71–100. (Cité p. 64)

[Pargny et Piantoni, 2005a] Dominique Pargny et Frédéric Piantoni. Méthodologie pour la gestion, la représentation et la modélisation des données archéologiques. Dans *Conférence Francophone ESRI*, 2005. (Cité p. 5)

[Pargny et Piantoni, 2005b] Dominique Pargny et Frédéric Piantoni. SIGRem : un Système d'Information Géographique pour l'archéologie en Champagne-Ardenne. Dans *colloque Archéologie en Champagne-Ardenne*. INRAP, 2005. (Cité p. 5)

[Pawlak, 1982] Zdsislaw Pawlak. Rough sets. *International Journal of Information and Computer Sciences*, 11(5) :341–356, 1982. (Cité p. 52)

[Pawlak, 1991] Zdsislaw Pawlak. *Rough Sets. Theoretical Aspects of Reasoning about Data*. Kluwer, Dordrecht, Pays-Bas, 1991. (Cité p. 13 et 52)

[Pawlak *et al.*, 1995] Zdsislaw Pawlak, Jerry Grzymala-Busse, Roman Slowinski, et Wojciech Ziarko. Rough Sets. *Communications of the ACM*, 35(11) :88–95, 1995. (Cité p. 52)

[Piantoni, 2005] Frédéric Piantoni. Le SIGRem. Problématique et méthodologie. Dans *séminaire de recherche du CIRAR*, 2005. (Cité p. 5)

[Piantoni *et al.*, 2005] Frédéric Piantoni, Dominique Pargny, Robert Neiss, et Claire Prévôtat. Méthodologie pour la gestion, la représentation et la modélisation des données archéologiques en milieu urbain. Le SIGRem. Dans *colloque du réseau Information Spatiale et Archéologie*, 2005. (Cité p. 5)

[Pitas et Tsakalides, 1991] I. Pitas et P. Tsakalides. Multivariate Ordering in Color Image Filtering. *IEEE Transactions on Circuits and Systems for Video Technology*, 1(3) :247–259, 1991. (Cité p. 148 et 191)

[Ramik et Rimanek, 1985] J. Ramik et J. Rimanek. Inequality relation between fuzzy numbers and its use in fuzzy optimization. *Fuzzy Sets and Systems*, 16 :123–138, 1985. (Cité p. 18, 74, 78, 80, 81 et 122)

[Rao, 1964] C. Radhakrishna Rao. The use and interpretation of principal component analysis in applied research. *Sankya serie A*, 26 :329–358, 1964. (Cité p. 19, 161, 162 et 163)

[Rodier et Galinié, 2006] Xavier Rodier et Henri Galinié. Figurer l'espace/temps de Tours pré-industriel : essai de chrono-chorématique urbaine. *M@ppemonde*, 83(3), 2006. (Cité p. 183)

[Rodier et Saligny, 2007] Xavier Rodier et Laure Saligny. Modélisation des objets urbains pour l'étude des dynamiques urbaines dans la longue durée. Dans *SAGEO'07*, Clermont-Ferrand, France, 2007. (Cité p. 9, 16, 27 et 56)

[Rolland-May, 2000] Christiane Rolland-May. *Évaluation des territoires*. Hermes, 2000. (Cité p. 25 et 63)

[Sasov, 1992] A. Sasov. Non-raster isotropic scanning for analytical instruments. *Journal of Microscopy*, 165, 1992. (Cité p. 164)

[Scherzer, 2006] Otmar Scherzer, Éditeur. *Mathematical models for registration and applications to medical imaging*. Mathematics in industry. Springer, 2006. (Cité p. 112)

[Shafer, 1976] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J., USA, 1976. (Cité p. 13, 23, 42 et 43)

[Sheeren, 2005] David Sheeren. *Méthodologie d'évaluation de la cohérence inter-représentations pour l'integration de bases de données spatiales*. PhD thesis, Université Paris 6, 2005. (Cité p. 108, 111 et 210)

[Sheeren *et al.*, 2008] David Sheeren, Sébastien Mustière, et Jean-Daniel Zucker. A data mining approach for assessing consistency between multiple representations in spatial databases. *International Journal of Geographical Information Systems*, 2008. A paraitre. (Cité p. 110)

[Shi, 2007] Wenzhong Shi. Four advances in handling uncertainties in spatial data and analysis. Dans *International Symposium Spatial Data Quality - ISSDQ'07*, 2007. (Cité p. 56, 62 et 209)

[Shokri *et al.*, 2006] T. Shokri, Mahmoud Reza Delavar, M. R. Malek, et Andrew U. Frank. Modeling uncertainty in spatiotemporal objects. Dans M. Caetano et M. Painho, Éditeurs, *7th International Symposium on Spatial Accuracy Assessment in Natural Ressources and Environmental Sciences*, pages 469–478, Lisbone, Portugal, 2006. (Cité p. 64)

[Smets, 1990] Philippe Smets. The Combination of Evidence int the Transferable Belief Model. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12 :447–458, 1990. (Cité p. 44)

[Smets, 1994] Philippe Smets. The Transferable Belief Model. *Artificial Intellignence*, 66 :191–234, 1994. (Cité p. 44)

[Smith *et al.*, 1987] T. R. Smith, S. Menon, J. L. Starr, et J. E. Estes. Requirements and principles for the implementation and construction of large-scale geographic information systems. *International Journal of Geographical Information Systems*, 1 :13–31, 1987. (Cité p. 7 et 8)

[Thériault et Claramunt, 1999] Mariue Thériault et Christophe Claramunt. La représentation du temps et des processus dans les SIG : une nécessité pour la recherche interdisciplinaire. *Revue Internationale de Géomatique*, 1 :67–99, 1999. (Cité p. 9)

[Torre et Poggio, 1984] V. Torre et T. Poggio. On Edge Detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8(2) :147–163, 1984. (Cité p. 95)

[VanLeekwijck et Kerre, 1999] Werner VanLeekwijck et Etienne E. Kerre. Defuzzification : criteria and classification. *Fuzzy Sets and Systems*, 108 :159–178, 1999. (Cité p. 47, 129, 166, 167 et 48)

[Vautrot *et al.*, 2006] Philippe Vautrot, Laurent Hussenet, et Michel Herbin. A Robust Filtering Method Using OWA Filters : Application to ColorImages. Dans *3rd European Conference on Colour in Graphics, Imaging and Vision*, University of Leeds, Royaume-Uni, 2006. (Cité p. 120 et 191)

[Volz, 2006] S. Volz. An iterative approach for matching multiple representations of streetdata. Dans *ISPRS Workshop on Multiple Representation and Interoperability of Spatial Data*, volume XXXVI, Hannover, Allemagne, 2006. ISPRS. (Cité p. 109 et 112)

[Wang et Kerre, 2001a] Xuzhu Wang et Etienne E. Kerre. Reasonable properties for the ordering of fuzzy quantities (I). *Fuzzy Sets anf Systems*, 118 :375–385, 2001. (Cité p. 78, 82, 132, 133 et 166)

[Wang et Kerre, 2001b] Xuzhu Wang et Etienne E. Kerre. Reasonable properties for the ordering of fuzzy quantities (II). *Fuzzy Sets anf Systems*, 118 :387–405, 2001. (Cité p. 78 et 82)

[Wang *et al.*, 1995] Xuzhu Wang, Etienne E. Kerre, B. Cappelle, et D. Ruan. Transitivity of Fuzzy Orderings Based on Pairwise Comparis. *The Journal of Fuzzy Mathematics*, 3(2) :455–463, 1995. (Cité p. 78, 130, 135 et 131)

[Webb, 2002] Andrew R. Webb. *Statistical Pattern Recognition*. Wiley, 2002. (Cité p. 95)

[Xiong et Sperling, 2004] Demin Xiong et Jonathan Sperling. Semiautomated matching for network database integration. *ISPRS Journal of Photogrammetry and Remote Sensing*, 59(1-2) :35–46, 2004. (Cité p. 109 et 113)

[Xu et Oja, 1993] Lei Xu et Erkki Oja. Randomized Hough Transform (RHT) : Basic Mechanisms, Algorithms, and Computational Complexities. *CVGIP : Image Understanding*, 57(2) :131–154, 1993. (Cité p. 94)

[Zadeh, 1965] Lotfi A. Zadeh. Fuzzy Sets. *Information Control*, 8 :338–353, 1965. (Cité p. 13, 23, 45, 74, 78, 79, 99, 100, 132, 46 et 133)

[Zadeh, 1978] Lotfi A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1 :3–28, 1978. (Cité p. 13 et 50)

Table des figures

Liste des tableaux

Liste des algorithmes

Résumé

Face aux enjeux urbains actuels, à la patrimonialisation des ressources archéologiques et grâce au développement de l'informatique, l'utilisation des systèmes d'information géographique devient essentielle pour l'exploitation des données archéologiques. Pour cela, il s'avère nécessaire de modéliser, d'analyser et de visualiser l'information archéologique en prenant en considération l'aspect temporel et spatial mais surtout les imperfections des données archéologiques. Cette thèse élabore une démarche globale pour l'utilisation de données spatiotemporelles imparfaites dans un SIG archéologique. Cette démarche contribue à une meilleure gestion de celles-ci tant pour leur représentation que pour leur traitement. Dans cette démarche scientifique, les concepts théoriques de taxonomie de l'imperfection et de représentation des données imparfaites permettent d'abord la modélisation des données archéologiques. Ce mémoire propose ensuite des méthodes d'analyse des données d'un SIG archéologique. La spécificité de leur caractère temporel implique une gestion plus flexible du temps par un indice quantifiant l'antériorité. L'aspect lacunaire de l'information est aussi considéré à travers une méthode d'interrogation sous critère de forme. Enfin, des outils originaux d'exploration et de visualisation de données archéologiques sont exposés afin de mieux définir les éléments les plus représentatifs. Par une approche interdisciplinaire liant informatique et géographie, cette thèse développe une vision transversale autour de la gestion des connaissances imparfaites dans le temps et l'espace. Cette approche est illustrée par l'utilisation de données archéologiques dans un SIG.

Mots-clés: Système d'information géographique, archéologie, données spatiotemporelles, imperfection, modélisation, analyse, exploration, visualisation

Abstract

This thesis develops a global approach for the handling of spatiotemporal and imperfect data in an archaeological GIS. This approach allows us a better management of those data in order to model or to represent them. In this approach, a new taxonomy of imperfection is proposed for the modeling of archaeological information. Using the modeling, this work presents some new methods for data analysis in an GIS. The temporal aspect of archaeological data implies to define an index which quantifies the anteriority. The lacunar aspect is also exploited through an interrogation method using a geometrical form. This work finally explores and visualizes archaeological dataset to extract the most representative elements. This thesis, which gives an approach on the management of imperfect knowledge in time and space, links computer science and geography. The use-case of this thesis is an archaeological database associated to a GIS.

Keywords: Geographical Information System, archaeology, spatiotemporal data, imperfection, modeling, analysis, data mining, visualization