Exploration visuelle de données spatiotemporelles imprécises : application en archéologie

Cyril de Runz*, Frédéric Blanchard*, Philippe Vautrot*
Eric Desjardin*, Michel Herbin**

* CReSTIC
IUT de Reims Châlons Charleville
Rue des Crayères, BP 1035, 51687 Reims Cedex 2, France
cyril.de-runz@univ-reims.fr
**CReSTIC
Antenne de Châlons, IUT de Reims Châlons Charleville
Chaussée du port, BP 541, 51012 Châlons-en-Champagne cedex, France
michel.herbin@univ-reims.fr

Résumé. Dans cet article, nous proposons d'exploiter une technique spécifique d'exploration visuelle d'un ensemble d'objets archéologiques dont les composantes spatiales et temporelles sont représentées par des ensembles flous convexes et normalisés. Pour cela, en nous basant sur la définition de vecteurs multidimensionnels issus de défuzzifications ou de comparaisons entre deux nombres flous, nous construisons une image couleur dans laquelle chaque pixel représente un objet. L'image couleur donne un rendu synthétique de l'information permettant à l'utilisateur de l'observer et de l'analyser.

1 Introduction

L'étude intuitive et visuelle de l'ensemble des données associées aux objets d'une base de données archéologiques est complexe dans les systèmes d'information géographique (SIG). En effet, bien que l'on puisse actuellement avoir une légende combinant un certain nombre d'attributs, ce nombre est limité. L'exploration visuelle nécessite alors d'utiliser une technique spécifique de visualisation permettant de présenter un résumé de l'information (contenue dans les données) sans réduire le nombre de données visualisées. Guptill (2005) considère que l'information géographique peut être vue comme une collection de données multidimensionnelles. C'est ce principe que nous exploiterons afin de nous permettre d'utiliser une méthode de visualisation de bases de données multidimensionnelles.

L'approche générale de l'exploration visuelle de grands volumes de données multidimensionnelles consiste à présenter un résumé en image de ces informations à l'instar de la démarche proposée par Keim (2000) et celle introduite dans Auber et al. (2007). Afin de visualiser la plus grande quantité d'information possible, nous utilisons une technique de visualisation sans *a priori* sur les données qui construit une image couleur à partir de ces informations et qui fut introduite dans Blanchard et al. (2005).

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 [1] ». La couleur et la spatialisation fournissent alors une image qui constitue un résumé des données et permet de voir de manière intuitive les principales structures. Ce travail a montré son efficacité sur des bases de données classiques (Blanchard et al., 2005).

Pour cela, les données 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, on associe un pixel de l'image couleur à chaque donnée ; la couleur est affectée objectivement et calculée 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 sont organisés spatialement à l'aide d'une courbe de remplissage dite de Peano-Hilbert (Moon et al., 2001). Cette technique permet de dégager visuellement des informations structurelles (proximité, regroupement) sur les données. Cette méthode se place dans le champs des techniques de fouille de données et de l'extraction de connaissance.

Dans cet article, 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 grands volumes de données et rendent presque impossible l'exploration intuitive et directe de ces composantes. Dans le second processus, l'objectif est de visualiser les dissimilarités à un objet sélectionné dans la base de données.

Dans le cas du projet SIGRem (présenté par exemple dans Desjardin et de Runz (2009)), les objets de *BDFRues* représentent des tronçons de rues romaines trouvés à Reims. La théorie des ensembles flous étant une des principales théories permettant de représenter l'imprécision de manière graduée sur l'espace des possibles, les composantes temporelles, spatiales et orientationnelles des données sont modélisées en tenant compte de leurs imprécisions par des ensembles flous convexes et normalisés (de Runz et al., 2008). Il faut donc pré-traiter l'information afin d'en dégager des évaluations quantitatives qui seront dès lors considérées comme des vecteurs multidimensionnels. Ainsi, la technique proposée est appliquée aux vecteurs multidimensionnels pour visualiser les objets archéologiques.

Nous proposons, dans cet article, de visualiser les composantes temporelles des objets de *BDFRues*. Pour cela, comme énoncé précédemment, il est nécessaire de quantifier les données avant même de lancer le processus de visualisation. Dans ce but, nous construisons 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 nombres flous.

Pour visualiser les dissimilarités entre objets archéologiques, les vecteurs multidimensionnels nécessaires sont issus de trois indices de dissimilarité entre objets archéologiques présentés dans de Runz et al. (2008). Ces indices de dissimilarité permettent d'évaluer les dissimilarités temporelles, d'orientation et de localisation entre objets archéologiques. Pour un objet sélectionné, l'image couleur de visualisation regroupe alors spatialement les pixels associés aux objets qui lui sont le moins dissimilaires d'un point de vue spatial, directionnel et temporel.

Nous présenterons dans la section 2 le contexte applicatif. La section 3 sera dédiée à la description de la technique de visualisation choisie. La section 4 exposera le processus de

1. La couleur est calculée à partir des données sans utilisation d'une échelle de couleurs.

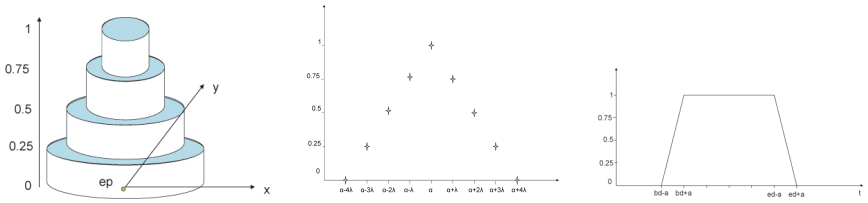


FIG. 1 – *Modèles flous pour la localisation, l'orientation et les périodes d'activité des tronçons de rues romaines*

visualisation des composantes temporelles associées aux tronçons de rues trouvés à Reims datant de l'époque romaine. Dans la section 5, nous proposerons de visualiser la dissimilarité des objets de la base vis-à-vis d'un objet présélectionné.

2 Projet SIGRem

Dans la problématique de la valorisation et de la gestion du patrimoine archéologique, la démarche développée par l'Université de Reims Champagne Ardenne, l'Institut National de Recherches Archéologiques Préventives et Ministère de la Culture et de la Communication dans le Centre Interinstitutionnel de Recherches Archéologiques de Reims 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 [2], le projet SIGRem, soutenu par la région Champagne Ardenne, l'État et la ville de Reims, et cadre applicatif de ce travail, porte sur la mise en place d'un SIG pluridisciplinaire intégrant les données archéologiques recueillies depuis les 30 dernières années. Dans cet article, nous proposons d'appliquer le processus exploratoire proposé sur la base de données *BDFRues*, partie intégrante du projet SIGRem. Cette base est dédiée aux éléments de rues romaines à Reims. Elle est constituée de 33 objets à l'heure actuelle. Son enrichissement est en cours. Les tronçons de rues sont caractérisés par des points ayant une orientation et une période d'activité.

La datation de la période d'activité des objets est généralement issue d'interprétations ou d'estimations dépendantes de l'environnement de la découverte (lieux de fouilles, stratigraphie, comparaison aux objets se situant dans la même pièce…). Elle est donc largement imprécise. Le géoréférencement est lui aussi sujet à de l'imprécision liée à différents facteurs : positionnement du point de fouilles, position par rapport à la route, référentiel utilisé, mouvement de terrain. L'orientation de la route est aussi à redéfinir dans ce cadre. En effet, l'orientation est notamment dépendante de la technique d'estimation utilisée à l'époque de la fouille.

Nous représentons les orientations, les périodes d'activité, et les localisations par des ensembles flous convexes et normalisés à savoir respectivement par des nombres flous, des intervalles flous et des ensembles flous spatiaux (2D). On peut ainsi prendre en compte cette incertitude (voir Figure 1).

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

Afin d'obtenir un rendu synthétique visuel et pertinent de ces données, nous proposons d'exploiter une technique de visualisation orientée-pixel présentée dans la section suivante.

3 Visualisation de données par une image couleur

Afin de fournir une image couleur des objets, nous nous intéressons 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 (la dimensionnalité peut être supérieure à 100 dans certains cas). 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 dimensionnalité » (Bellman, 1961; Donoho, 2000). Par ailleurs, l'effectif de l'échantillon peut être considérable : il peut dépasser le million d'individus. Les techniques de visualisation ont alors tendance à masquer l'information pertinente du fait de cet effectif.

Dans ce cadre, la méthode de Blanchard et al. (2005) est une approche orientée-pixel 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 par la réduction de la dimensionnalité et de la seconde par l'association d'un pixel à chaque donnée permettant ainsi de visualiser autant de données qu'il y a de pixels affichables.

3.1 Réduction de la dimensionnalité

L'analyse de données multidimensionnelles nécessite une réduction de dimensionnalité pour des raisons pratiques et théoriques (représentations des données, malédiction de la dimensionnalité). 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 réduction de dimensionnalité est utilisée : on conserve les trois premières composantes générées par une Analyse en Composantes Principales (Rao, 1964). Une revue des techniques d'ACP est proposée dans Jolliffe (1986), Cardoso et Comon (1996) et Hyvärinen (1999).

L'ACP est considérée comme une approche statistique usuelle en science de l'information géographique pour synthétiser l'information. Une étude de son usage en géographie est proposée dans Wang (2009). Elle est par exemple utilisée dans Jacquemot et al. (2004).

Le principe est de projeter les données dans un sous-espace de dimension trois, les axes de projection étant orthogonaux et décorrélés. L'avantage de l'ACP est de déterminer itérativement les composantes par ordre décroissant de l'information qu'elles portent. 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 l'ACP, on maximise l'information contenue dans ces trois composantes. De nombreux travaux proposent des approches alternatives pour réduire la dimension. Des techniques dites de poursuite de projection (Nason, 1995) constituent un autre moyen d'obtenir des projections orthogonales en optimisant un index de projection. Si l'orthogonalité n'est pas nécessaire, l'Analyse en Composantes

Indépendantes (ACI) permet d'obtenir des composantes « aussi statistiquement indépendantes que possible » en maximisant une fonction de contraste Comon (1994). Il faut cependant noter que la relation liant la couleur aux trois premières composantes de l'ACP mise en avant dans (Ohta et al., 1980) est le fondement de notre choix porté sur l'ACP.

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.

3.2 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). Un ensemble de N données sera représenté par une image couleur (R,V,B) ayant N pixels. Ainsi, chaque donnée est représentée par une couleur : un triplet (R,V,B). 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 ; 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. Trouver un ordre sur l'ensemble des pixels équivaut à projeter les données sur un espace de dimension 1 (une ligne). Les représentations des données sont alors ordonnées (ou rangées). Or la dimension est réduite par la projection des données dans un espace 3D (du fait du choix des 3 premières composante de l'ACP sur l'ensemble 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 l'ACP 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. La récupération des trois premières composantes de l'ACP et leur tri sont d'ordre polynomial.

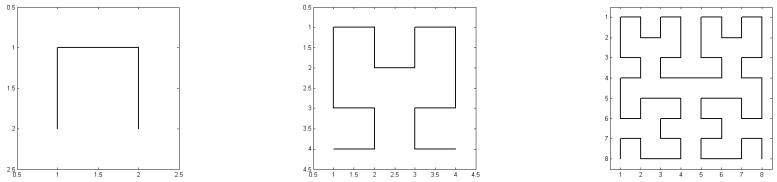


FIG. 2 – *Étapes de la construction d'une courbe de Peano-Hilbert*

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 2 la description de la procédure récursive de construction d'une telle courbe). Le principal avantage de cette courbe est de préserver au mieux les regroupements des classes de données (Sasov, 1992). En effet, deux points qui sont proches sur la ligne initiale sont proches dans l'image construite. Ces écarts seraient plus importants si l'on utilisait un parcours de l'image ligne par ligne ou bien colonne par colonne, créant des discontinuités et des sauts.

Avec ces deux étapes de tri des données et de remplissage de l'image par une courbe de Peano-Hilbert, on évite de disperser les pixels dans l'image construite (voir figure 3). Cette approche tend à préserver la cohérence spatiale des données permettant ainsi une visualisation très intuitive des échantillons de données.

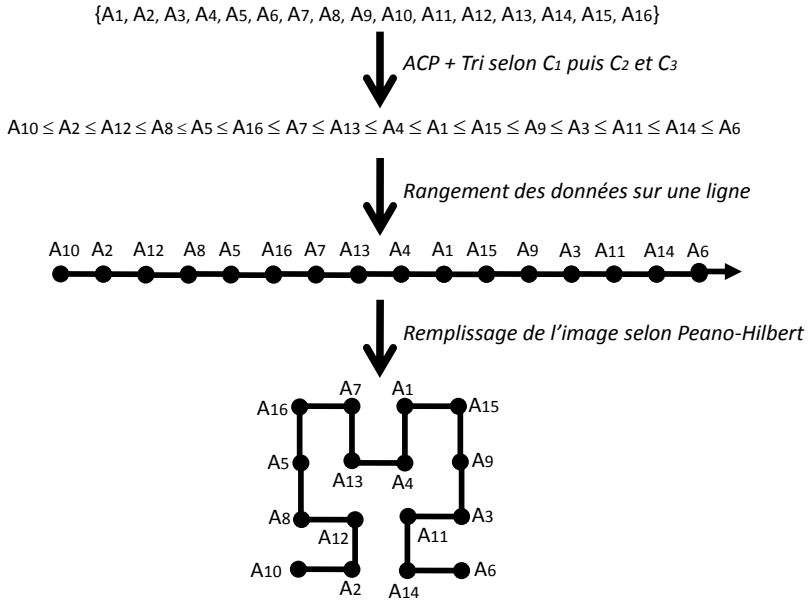


FIG. 3 – *Illustration du processus de remplissage de l'image pour 16 données multidimensionnelles quelconques* $A_1, A_2, \ldots, A_{16}$

Notre approche permettant la visualisation des données par une image couleur, nous devons maintenant déterminer la couleur de chaque pixel en fonction des valeurs contenues dans la donnée réduite associée au pixel.

3.3 À 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). Toutefois, on ne peut considérer pas

que les trois composantes principales forment directement le triplet RVB car on nierait les quantités relatives d'information contenues par les composantes résultantes de l'ACP (*i.e.* on introduirait arbitraitmenent des importances différentes aux trois couleurs).

La technique de visualisation proposée se base pour l'affectation de la couleur sur l'étude statistique de la couleur de Ohta et al. (1980). Ces derniers proposent d'approximer l'ACP 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 l'ACP (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 calculés à partir des composantes C_1, C_2 et C_3 de la manière suivante :

$$\begin{cases} R &= (6 \times C_1 + 3 \times C_2 - 2 \times C_3)/6 \\ V &= (3 \times C_1 + 2 \times C_3)/3 \\ B &= (6 \times C_1 - 3 \times C_2 - 2 \times C_3)/6 \end{cases} \quad (1)$$

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. Ohta et al. ont proposé leur transformation pour permettre une segmentation plus aisée d'une image de couleurs naturelles. La transformation inverse doit permettre d'obtenir des couleurs respectant naturellement les classes des données. Cette approche de la couleur dépend de l'échantillon de données. Si l'échantillon change, les couleurs changent. Elle propose un résumé coloré associé à un échantillon.

Notre idée est d'utiliser cette technique pour visualiser les objets selon l'information des composantes temporelles, et les dissimilarités des objets archéologiques en tenant compte de leurs imperfections.

4 Visualisation des objets selon leurs périodes d'activité par une image couleur

Les données archéologiques sont spatiotemporelles et imprécises. Afin de donner une lecture intuitive des données, il est nécessaire de les visualiser. Lorsqu'elles sont stockées dans une base de données 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 fixée (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 imprécises est difficile. En effet, dans le cadre de données représentées par des nombres flous, représenter l'ensemble des fonctions d'appartenance sur un même repère complique la lecture des nombres flous à considérer. Par exemple, si l'on regarde la figure 4, l'extension à plus de 10 nombres flous des modèles proposés par le logiciel *Mathematica* de *Wolfram Research* devient peu lisible même en 3 dimensions.

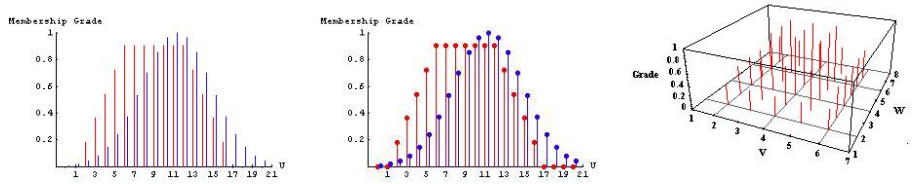
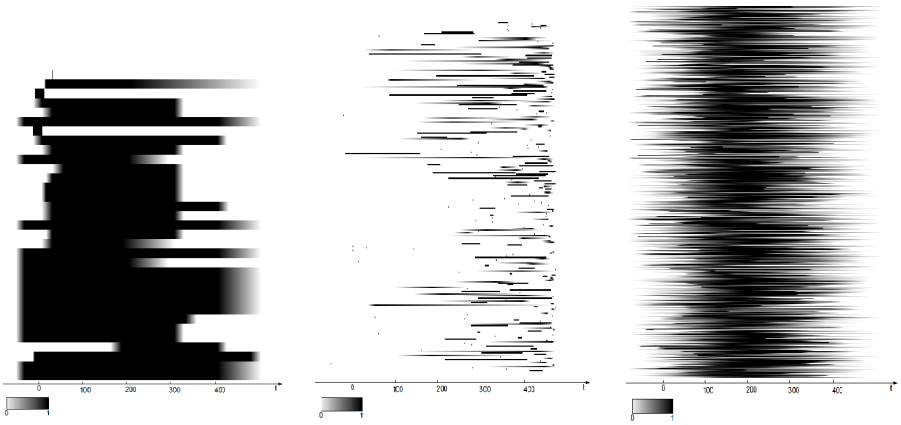


FIG. 4 – *Modèles pour la visualisation d'ensembles et relations floues proposés par le logiciel Mathematica Wolfram*

Une autre visualisation est cependant possible pour un nombre restreint d'ensemble à représenter (figure 5a). Dans celle-ci on représente un ensemble flou par une bande dans une image et on fait varier le niveau de gris en fonction des degrés d'appartenances (noir pour 1, blanc pour 0). Malheureusement, lorsque l'on souhaite visualiser 256 nombres flous, le rapprochement visuel (ou regroupement visuel) est difficile (figure 5b). Avec 1024 objets, la tâche est impossible (figure 5c).



a : 33 nombres flous issues de BDFRues

b : 256 nombres flous générés aléatoirement

c : 1024 nombres flous générés aléatoirement

FIG. 5 – *Visualisation en bande de nombres flous dans une image en niveau de gris*

Le problème de la visualisation est donc lié au nombre et à la nature complexe des objets à observer. Notre idée est de les décrire plus simplement et d'organiser spatialement ces descriptions par la méthode présentée précédemment.

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.

Pour construire le pixel associé à chaque objet, on choisit de décrire les nombres flous en les évaluant à l'aide de différentes méthodes de défuzzification.

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 nombres flous rangent le

plus souvent celles-ci par le biais d'évaluations (Wang et Kerre, 2001).

Le principe de la visualisation consiste alors à décrire un nombre flou par plusieurs évaluations quantitatives obtenues avec différentes méthodes de défuzzification. Un nombre flou est alors représenté par un vecteur d'évaluations que l'on assimile à une donnée quantitative multidimensionnelle. Les données sont ensuite visualisées en utilisant les vecteurs d'évaluations.

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.

Ainsi, la représentation visuelle dépendra des méthodes d'évaluations choisies. Pour explorer un ensemble de nombres flous, comme par exemple pour visualiser les informations temporelles, les méthodes de défuzzification doivent ne considérer que le nombre à évaluer pour attribuer une valeur (VanLeekwijck et Kerre, 1999).

4.1 Méthodes de défuzzification des nombres flous

Les méthodes de défuzzification présentées ici ne considèrent que le nombre flou à évaluer. VanLeekwijck et Kerre (1999) les séparent en trois classes. Bien que chaque méthode ait ses particularités, les classes proposées suggèrent des utilisations différentes. Le choix de l'utilisation de l'une de ces méthodes dépend donc fortement de l'analyse voulue.

Les méthodes de défuzzification utilisent la notion de support, de cœur, et de hauteur d'une quantité floue. Le support est le domaine pour lequel la valeur de la fonction d'appartenance de la quantité est strictement positive. La hauteur est la valeur maximale de la fonction d'appartenance de la quantité. Le cœur est le domaine pour lequel la valeur de la fonction d'appartenance est égale à la hauteur.

Les méthodes de type maxima et les méthodes dérivées forment la première classe. Elles sélectionnent un élément du cœur de la quantité (nombre) 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 d'abord 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, essentielle pour les contrôleurs flous.

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 dédiées au contrôle flou.

Afin d'explorer visuellement les objets archéologiques selon les représentations de leurs périodes d'activité, nous souhaitons 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.

4.2 Construction du vecteur multidimensionnel d'évaluation de la représentation d'une période d'activité

Afin de construire simplement un vecteur avec des méthodes de chaque classe, nous proposons 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, nous choisissons 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.

En ce qui concerne la seconde classe, nous sélectionnons 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.

Enfin, pour la dernière classe, nous prenons 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.

Nous associons donc à chaque période d'activité un vecteur d'évaluation de dimension 7. C'est par le prisme de ce vecteur que nous explorons les données. Pour cela, nous utilisons l'ensemble des vecteurs en entrée du processus de visualisation de Blanchard et Herbin.

4.3 Visualisation des objets selon les représentations de leurs périodes d'activité

Le processus général de l'exploration est présenté dans la figure 6.

L'image résultat est présentée sur la figure 7. Elle contient 33 pixels en couleurs. Chaque objet est représenté par un pixel construit à partir du nombre flou associé à la période d'activité de l'objet. L'organisation spatiale et l'information couleur des pixels permettent d'observer de façon immédiate des informations de structuration de cet ensemble de périodes. Cette image suggère des regroupements des données par couleurs semblables.

Par ailleurs, 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 8). Cette opération de projection conserve donc la quasi totalité de l'information apportée par les différentes évaluations. La visualisation repose donc sur l'essentiel de l'information temporelle contenue dans *BDFRues*.

De plus, si l'on reprend notre approche sur les exemples des données visualisées dans la figure 5 on obtient la figure 9. Notre approche permet donc de visualiser les activités temporelles de plus de 1000 objets (figure 9c) en considérant leur imprécision et dans une fenêtre permettant d'être insérée dans un système d'information géographique (SIG) pour l'interaction avec les composantes spatiales de l'information archéologique.

Cette interaction est illustrée dans la figure 10 pour les 33 objets de BDFRues. En liant les pixels aux objets archéologiques et en intégrant l'image dans un système d'information géographique, on peut sélectionner les objets proches sur l'image couleurs et voir le résultat de

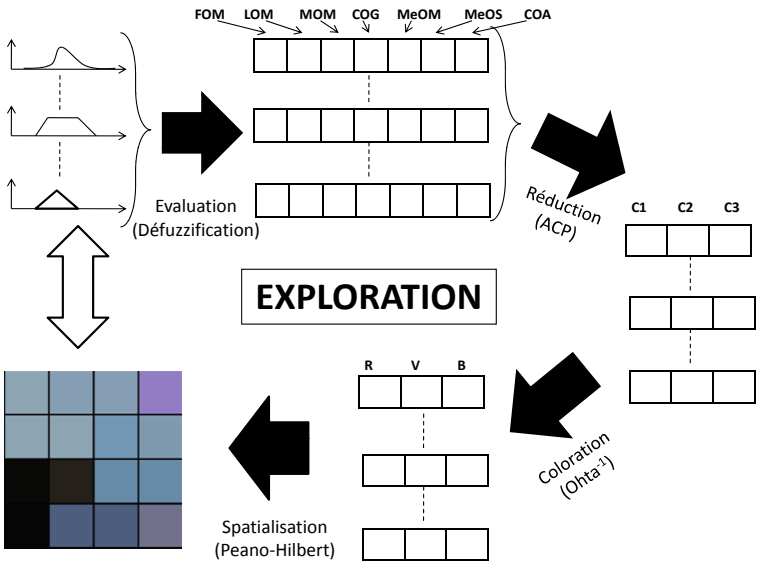


FIG. 6 – *Visualisation des objets selon les représentations floues des périodes d'activité — schéma récapitulatif*

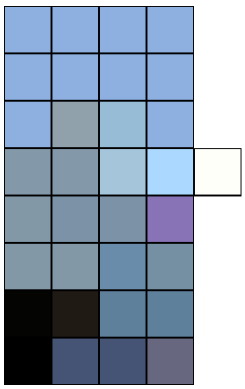


FIG. 7 – *Visualisation des objets de BDFRues par une image couleur selon les représentations de leurs périodes d'activité*

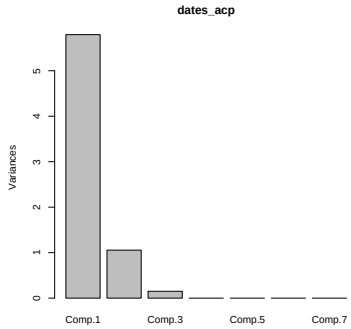
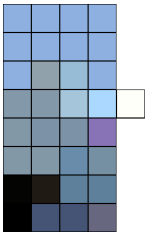
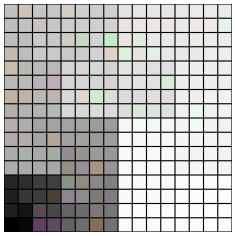


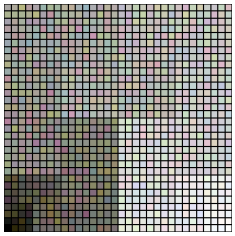
FIG. 8 – *Histogramme de l'ébouli des valeurs propres de l'ACP*



a : 33 nombres flous issues de BDFRues



b : 256 nombres flous générés aléatoirement



c : 1024 nombres flous générés aléatoirement

FIG. 9 – *Visualisation de nombres flous par une image couleur*

cette sélection dans une couche du SIG. Pour conforter la lecture couleur de l'information temporelle floue, nous avons ajouté dans la figure les résultats des sélections selon la visualisation par bande dans une image en niveau de gris. Les informations sélectionnées sont indiquées en rouge.

On peut observer pour les données de BDFRues trois groupes principaux de pixels dans l'image couleur : les pixels sombres (en bas à gauche de l'image), les pixels gris (au centre gauche de l'image) et les bleus clairs (en haut de l'image). Les 3 pixels sombres correspondent aux 3 objets ayant une très courte période d'activité se situant au début de la période de domination romaine à Reims (figure 10a). Les 8 pixels gris correspondent aux 8 objets ayant été en activité durant l'apogée de la cité sous l'empire romain (figure 10b). La sélection des 10 pixels bleus extrait de BDFRues les objets en activité durant l'ensemble de la période romaine (figure 10c).

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 proximité entre les représentations de leurs périodes d'activité. Cet 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.

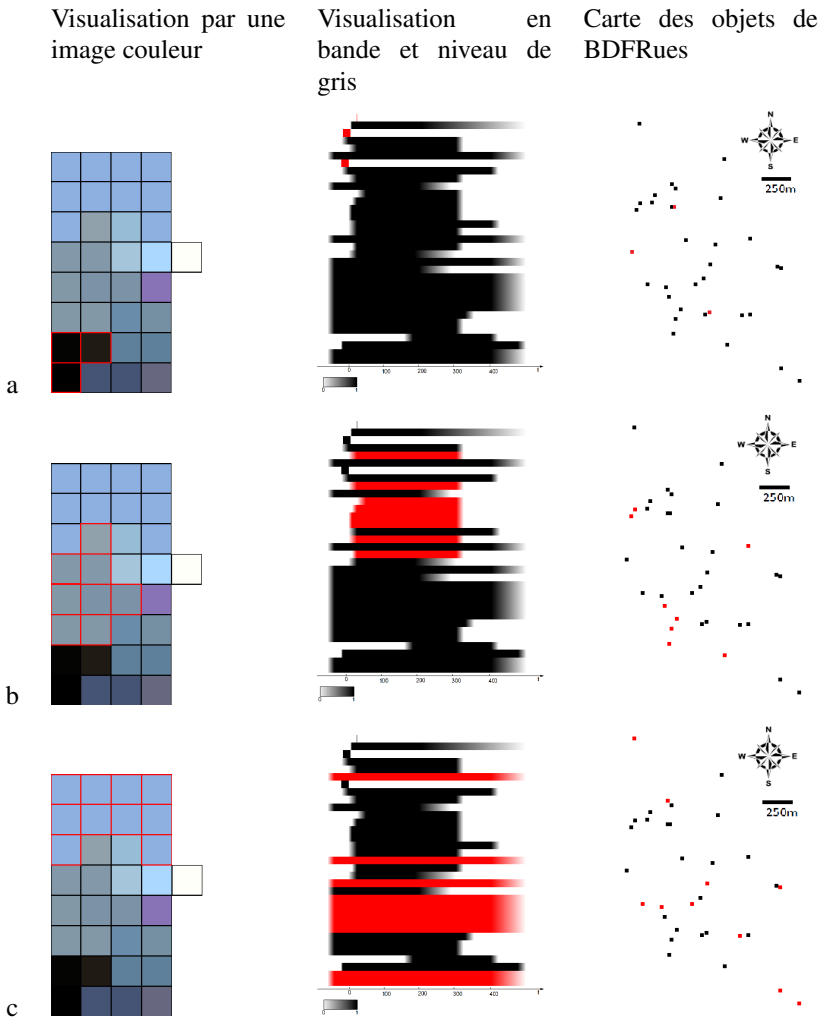


FIG. 10 – *Selection des objets de BDFRues dans l'image issue de notre approche et résultat de la sélection sur la carte et dans un mode usuel de visualisation*

5 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 A_j. L'information archéologique étudiée étant structurée autour de 3 caractéristiques principales, il nous faut exploiter au minimum un indice de dissimilarité pour chacune d'elles. Ainsi, nous utilisons trois indices de dissimilarité : D_{date}

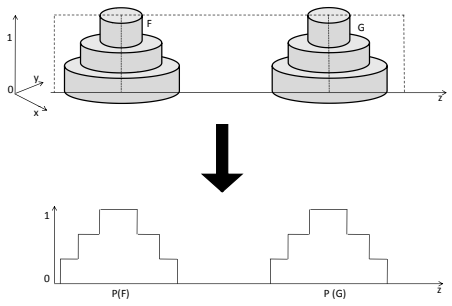


FIG. 11 – *Projections pour le calcul de dissimilarité des localisations*

pour l'aspect temporel, D_{orien} pour l'aspect orientation et D_{loc} pour les localisations. Nous proposons d'exploiter pour les définitions de ces indices, les mêmes que celles utilisées dans de Runz et al. (2008) et qui sont explicitées dans le paragraphe suivant. Ces indices simples à calculer ont montré leur intérêt dans la recherche d'objets représentatifs de la base ce qui suggère qu'ils mettront en valeur les différences entre objets.

5.1 Construction du vecteur multidimensionnel d'évaluation des dissimilarités à un objet sélectionné

Nous proposons d'utiliser une distance classique (Grzegorzewski, 1998) entre nombres et/ou intervalles flous comme mesure de dissimilarité. Soit F et G deux nombres et/ou intervalles flous, soit $F_{\alpha-}$ (resp. $G_{\alpha-}$) et $F_{\alpha+}$ (resp. $G_{\alpha+}$) les bornes inférieure et supérieure de l'α-coupe F_α de F (resp G_α de G), alors la distance entre F et G est obtenue par :

$$D(F,G) = \int_0^1 |F_{\alpha-} - G_{\alpha-}| + |F_{\alpha+} - G_{\alpha+}| d\alpha. \quad (2)$$

Nous utilisons cette mesure pour le calcul de la dissimilarité d'orientations (D_{orien}) et de périodes d'activité entre éléments (D_{date}).

Pour le calcul de la dissimilarité de localisation, en raison du caractère cylindrique de la fonction d'appartenance des ensembles flous spatiaux associés aux données, nous calculons la mesure de dissimilarité D_{loc} à partir de leurs projections floues sur le plan passant par les centres des localisations. Nous obtenons ainsi des nombres et/ou intervalles flous, et calculons comme précédemment la mesure de la dissimilarité D_{loc} des objets en terme de localisation (voir la figure 11).

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 :

$$v_{A_j}(A_i) = (D_{date}(A_j, A_i), D_{orien}(A_j, A_i), D_{loc}(A_j, A_i)). \quad (3)$$

Ce vecteur contient donc trois informations : D_{date}, D_{orien} et D_{loc}.

5.2 Visualisation des dissimilarités

Nous proposons de visualiser la dissimilarité des objets à un objet sélectionné dans la base, en donnant en entrée du processus de visualisation les vecteurs de dimension 3 définis précédemment.

Le processus exploratoire, présenté dans la figure 12, est fortement similaire à celui présenté dans la figure 6 : le calcul des valeurs des indices de dissimilarité entre les objets et l'objet sélectionné A_j fournit un vecteur d'évaluation de chaque objet selon le point de vue de A_j. Ces vecteurs sont fournis en entrée du processus de visualisation (ACP, colorisation par transformée inverse d'Ohta, spatialisation selon Peano-Hilbert).

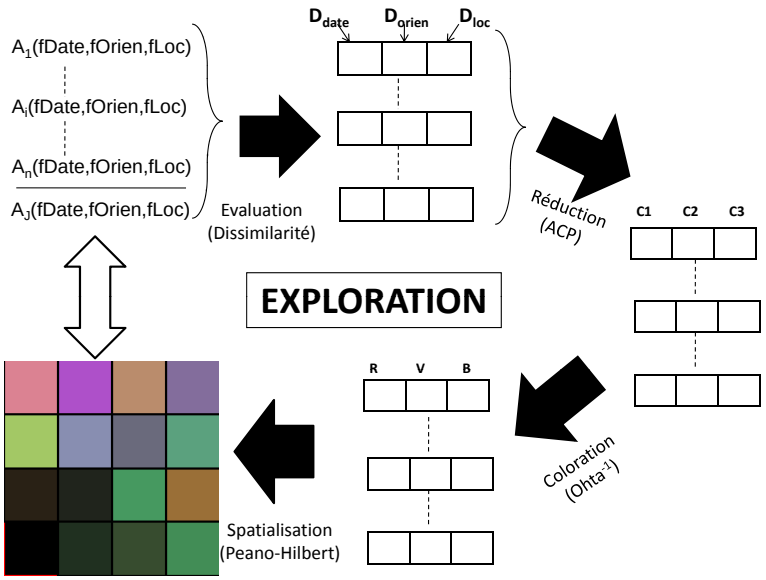


FIG. 12 – *Visualisation des objets selon leurs dissimilarités à un objet sélectionné — schéma récapitulatif*

Le processus de visualisation des dissimilarités des objets à un objet sélectionné donne la figure 13.

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 (entouré de rouge dans l'image) sont dans l'image les plus proches spatialement de celui sélectionné (de contour rouge dans la figure). Il y a donc *a priori* deux objets très similaires, en couleur sombre proche du noir, à celui sélectionné. En effet, de par le fait que leurs dissimilarités vis-à-vis de l'objet sélectionné est faible, cela se traduit par une faible disparité et donc une faible variance. Les composantes

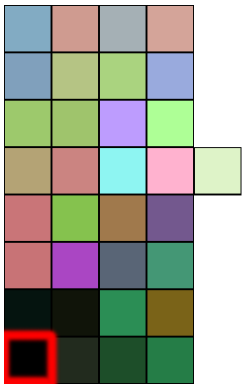


FIG. 13 – *Visualisation de la dissimilarité des objets à un objet sélectionné (contour rouge) par une image couleur*

couleurs des pixels associés aux dits objets auront donc des valeurs faibles ce qui donne une teinte proche du noir.

On peut noter que l'objet sélectionné est l'objet qui figure en bas à droite dans la figure 7. L'impact de la dissimilarité temporelle semble donc évident.

6 Conclusion

Nous avons présenté dans cet article une méthode originale d'exploration visuelle intuitive d'un ensemble d'objets archéologiques dont les composantes spatiales et temporelles sont représentées par des ensembles flous convexes et normalisés. Cette méthode s'est basée sur la construction de vecteurs dont les valeurs furent obtenues soit par plusieurs défuzzifications des représentations 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, nous avons 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), nous avons 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. Notre méthode bien qu'appliquée sur des données archéologiques peut être directement exploitée sur des bases de données stockant des nombres flous.

Afin de pouvoir utiliser des techniques de réduction de dimensionnalité plus récentes et peut-être plus efficaces (Lebart, 2007; Sharma et Paliwal, 2007) nous envisageons de reprendre les travaux d'Ohta et al. afin d'étudier et d'envisager leur adaptation à ces autres techniques.

Nous souhaitons aussi étudier l'apport possible des approches 3D pour compléter la présentation des informations spatiotemporelles imprécises. Par exemple, nous envisageons d'adapter les approches existantes pour les historiques web, à l'instar de Sureau et al. (2008). En effet la combinaison des approches pour le rendu de l'information donnerait selon nous une meilleure vision globale.

Enfin, nous envisageons aussi d'étudier l'impact de l'utilisation de l'ACP pour le rangement de nombres flous et comparer les résultats aux méthodes non statistiques. Dans notre approche nous projetons les données selon un tri sur les 3 premières composantes de l'ACP. Selon nos premières analyses, le rangement obtenu serait fonction de la forme des fonctions d'appartenance et des dates englobantes. Ce rangement différe sensiblement des approches plus classiques et son étude mérite d'être approfondie (stabilité du rangement de l'ACP par exemple).

Références

Auber, D., N. Novelli, et G. Melancon (2007). Visually mining the datacube using a pixel-oriented technique. In *IV '07 : Proceedings of the 11th International Conference Information Visualization*, Washington, DC, USA, pp. 3–10. IEEE Computer Society.

Bellman, R. (1961). *Adaptive control processes : a guide tour*. Princeton University Press.

Blanchard, F., M. Herbin, et L. Lucas (2005). A New Pixel-Oriented Visualization Technique Through Color Image. *Information Visualization 4*(4), 257–265.

Cardoso, J.-F. et P. Comon (1996). Independent Component Analysis, a Survey of Some Algebraic methods. In *ISCAS Conference*, Volume 2, pp. 93–96.

Comon, P. (1994). Independant Component Analysis, A new Concept ? *Signal Processing 36*(2), 287–314.

de Runz, C., F. Blanchard, E. Desjardin, et M. Herbin (2008). Fouilles archéologiques : à la recherche d'éléments représentatifs. In *Atelier Fouilles de Données Complexes - Conférence Extraction et Gestion des Connaissances - EGC'08*, Sophia Antipolis, France, pp. 95–103.

Desjardin, E. et C. de Runz (2009). Gissar : de la saisie de fouilles à l'analyse spatiotemporelle en archéologie. In *Spatial Analysis and GEOmatics*, Paris, France.

Donoho, D. L. (2000). High-Dimensional Data Analysis : The Curses and Blessings of Dimensionality. In *AMS Conference Mathematical Challenges of the 21st Century*.

Grzegorzewski, P. (1998). Metrics and orders in space of fuzzy numbers. *Fuzzy Sets and Systems 97*, 83–94.

Guptill, S. C. (2005). Metadata and data catalogues. In P. A. Longley, M. F. Goodchild, D. J. Maguire, et D. W. Rhind (Eds.), *Geographical Information Systems. Principles, Techniques, Management and Applications*, Volume 2, Chapter 49, pp. 677–692. Wiley. Seconde Edition.

Hyvärinen, A. (1999). Survey on independent component analysis. *Neural Computing Surveys 2*, 94–128.

Jacquemot, T., J. Corbonnois, et S. de Ruffray (2004). Hiérarchisation des processus de la dynamique fluviale par le traitement statistique des données. *Mosella XXIX*(3-4), 363–370.

Jolliffe, I. T. (1986). *Principal Component Analysis*. Springer Verlag.

Keim, D. A. (2000). Designing Pixel-oriented Visualization Techniques : Theory and Applications. *IEEE Transaction on Visualization and Computer Graphics (TVCG) 6*(1), 59–78.

Lebart, L. (2007). Which bootstrap for principal axes methods ? In H.-H. e. a. Bock (Ed.), *Selected Contributions in Data Analysis and Classification*, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 581–588. Springer Berlin Heidelberg.

Moon, B., H. V. Jagadish, C. Faloutsos, et J. H. Saltz (2001). Analysis of the Clustering Properties of the Hilbert Space-Filling Curve. *IEEE Transactions on Knowledge and Data Engineering 13*(1), 124–141.

Nason, G. (1995). Three-dimensional projection pursuit. *Applied Statistics 44*(4), 411–430.

Ohta, Y., T. Kanade, et T. Sakai (1980). Color Information for Region Segmentation. *Computer Graphics and Image Processing 13*, 222–241.

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankya serie A 26*, 329–358.

Sasov, A. (1992). Non-raster isotropic scanning for analytical instruments. *Journal of Microscopy 165*.

Sharma, A. et K. K. Paliwal (2007). Fast principal component analysis using fixed-point algorithm. *Pattern Recognition Letters 28*(10), 1151–1155.

Sureau, F., F. Bouali, et G. Venturini (2008). DataTube2 : exploration interactive de données temporelles en réalité virtuelle. In *Atelier Fouilles de Données Complexes - Conférence Extraction et Gestion des Connaissances - EGC'08*, Sophia Antipolis, France, pp. 13–24.

VanLeekwijck, W. et E. E. Kerre (1999). Defuzzification : criteria and classification. *Fuzzy Sets and Systems 108*, 159–178.

Wang, F. (2009). Factor analysis and principal-components analysis. In R. Kitchin et N. Thrift (Eds.), *International Encyclopedia of Human Geography*, pp. 1 – 7. Oxford : Elsevier.

Wang, X. et E. E. Kerre (2001). Reasonable properties for the ordering of fuzzy quantities (I). *Fuzzy Sets anf Systems 118*, 375–385.

Summary

In this article, we use a specific technique for the visualization of an archaeological dataset in which object components are modeled using normalized convex fuzzy sets. In order to build a color image of data, we use definition of multidimensional vector for each object. The color image gives us a resume of information allowing users to observe and analyze it graphically.