

Exploration d'un ensemble de quantités floues

Exploration of a set of fuzzy quantities

C. de Runz¹

F. Blanchard¹

E. Desjardin¹

M. Herbin²

¹ CReSTIC-SIC

² CReSTIC Antenne Châlons

CReSTIC-SIC, IUT de Reims Châlons Charleville, rue des Crayères, BP 1035, 51687 Reims cedex 2, France

(cyril.de-runz, frederic.blanchard, eric.desjardin)@univ-reims.fr

Antenne CReSTIC-Châlons, chaussée du port, BP 541, 51012 Châlons-en-Champagne cedex, France

michel.herbin@univ-reims.fr

Résumé :

Nous proposons dans cette communication une méthode d'exploration visuelle de quantités floues. Afin de pouvoir visualiser ces quantités nous proposons tout d'abord de les décrire par des vecteurs dont les valeurs sont des évaluations quantitatives obtenues avec différentes approches classiques de défuzzification. L'étape de visualisation consiste alors à calculer une couleur pour chaque vecteur (à partir des trois premières composantes principales d'une ACP sur l'ensemble des vecteurs) puis à organiser spatialement ces pixels dans une image couleur. Nous avons appliqué cette méthode sur des données archéologiques.

Mots-clés :

Fouille de données visuelle, Exploration de données floues, Défuzzification

Abstract:

In this paper we expose a visual data mining method on fuzzy quantities. To obtain the visualization, we first describe those quantities by vectors for which the values are the results of classical defuzzification processes. The visualization step consists in assigning a color to each vector and in organizing those pixels in a picture. The color is obtained by using the first three components of a principal component analysis on the set of vectors. We have applied this method on archaeological data.

Keywords:

Visual data mining, Fuzzy data exploratory, Defuzzification

1 Introduction

L'exploration d'un ensemble de données nécessite l'utilisation d'outils de visualisation. Ces visualisations à visée exploratoire sont préalables à l'analyse et à l'interprétation des données. Les travaux de Keim [12] présentent un résumé des diverses techniques de visualisation utilisées pour explorer un ensemble

de données. Dans ce papier, nous présentons une technique orientée-pixel consistant à représenter un ensemble par une image où chaque pixel correspond à une et une seule donnée. Dans ce travail, les couleurs des pixels sont déterminées « objectivement ». 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 [2]. Dans cet article, nous proposons d'adapter cette technique à la visualisation des ensembles de quantités floues, l'objectif demeure l'exploration d'un ensemble de données en l'occurrence d'un ensemble de quantités floues.

Pour visualiser des données floues nous proposons d'abord de les évaluer, car synthétiser l'information contenue dans les quantités floues est un préalable à leur analyse. Pour la comparaison, par exemple, la plupart des méthodes défuzzifient les quantités à comparer en leur associant une valeur réelle représentative [11, 10]. Parmi les différentes méthodes de défuzzification, on distingue principalement trois classes de méthodes [18] : celles qui sont basées sur les maxima de la fonction d'appartenance, celles qui convertissent la fonction d'appartenance en une fonction de densité de probabilité (comme par exemple la méthode dite du centre de gravité) et celles

basées sur l'aire sous la courbe représentative de la fonction d'appartenance. Dans un contexte d'exploration préliminaire d'un ensemble de données floues, nous ne privilégions aucune des méthodes de défuzzification. Dans ce travail, chaque donnée floue est donc transformée en un vecteur dont les composantes sont les différentes évaluations obtenues avec un panel des principales méthodes de défuzzification. La visualisation d'un ensemble de données floues consiste alors à représenter un ensemble de vecteurs où chaque vecteur est une évaluation multidimensionnelle de la quantité floue.

Notre méthode de visualisation affecte un triplet (Rouge, Vert, Bleu) à chaque donnée multidimensionnelle. Dans cet article, cela reviendra à attribuer une couleur à chaque quantité floue via son vecteur de défuzzification. Pour cela, nous effectuons une Analyse en Composantes Principales sur l'ensemble des vecteurs afin de les réduire aux 3 premières composantes. Ensuite, comme la transformée d'Ohta *et al.* [15] propose une transformation linéaire qui approxime les trois premières composantes à partir des triplets (R, V, B) d'une image couleur, nous utiliserons la transformée inverse afin d'obtenir les triplets (R, V, B) à affecter aux quantités floues à partir des trois premières composantes. Enfin, afin de pouvoir explorer de manière intuitive l'ensemble des données, nous organisons celles-ci par la construction d'une image couleur via une courbe de Peano-Hilbert.

Nous avons appliqué, dans le cadre du projet SIGRem (voir à ce propos [16]), notre méthode sur une base de données archéologiques *BDRues*. Les objets contenus dans cette base représentent les tronçons de rues datant de l'époque romaine découverts durant des fouilles préventives à Reims. Les périodes d'activité de ces objets sont modélisées par des ensembles flous convexes et normalisés (se référer à [5], [6] et [7]). L'objectif de l'application de notre méthode sur cette base est de fournir un outil intuitif d'interrogation temporelle des données.

Cet article commencera par présenter les

méthodes d'évaluation des quantités floues proposées dans [18] afin de construire les vecteurs multidimensionnels constitués des évaluations de ces quantités. Après quoi, nous présenterons les différentes étapes de notre méthode de visualisation : affectation des couleurs, construction de l'image. Ensuite, nous exposerons les résultats de l'application de notre approche sur les données archéologiques dont nous disposons, nous étudierons particulièrement le rendu de l'ACP. Enfin, après une discussion, nous conclurons ce travail, et exposerons les différentes perspectives envisagées.

2 Vecteur d'évaluations d'une quantité floue

L'objectif de cette section est de construire pour chaque donnée un vecteur d'évaluations de celle-ci. Ce vecteur sera alors une donnée multidimensionnelle représentant la quantité floue sur laquelle il a été construit. Dans cette section, nous exposerons la problématique de l'évaluation des données, les différentes classes de méthodes permettant de la faire et le choix fait pour la construction des vecteurs.

2.1 Évaluation de quantité floue

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. Elles sont de trois types (voir [19]) : soit elles ne considèrent que la quantité à évaluer à l'instar de [1], soit elles la considèrent par rapport à une autre quantité comme dans [4], soit elles la considèrent par rapport à l'ensemble des quantités [13].

L'évaluation de quantités floues est nécessaire à l'étude des systèmes les mettant en jeu. Elle donne une valeur à chaque entité floue ou à l'ensemble des entités. Afin d'explorer de tels en-

sembles, nous proposons de les visualiser en attribuant à chaque donnée une couleur. Ce sont donc les évaluations ne prenant en compte que la donnée qui nous intéressent.

2.2 Méthodes de défuzzification

Les méthodes de défuzzification présentées ici ne considèrent que la quantité floue à évaluer. Dans [18], Van Leekwijck et Kerre distinguent trois classes de méthodes. Bien que chaque méthode ait ses avantages et ses inconvénients, les classes proposées invitent à des utilisations différentes. Le choix de l'utilisation d'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 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 probabilité 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.

Dans la troisième classe, les méthodes utilisent les aires sous les fonctions d'appartenance pour

évaluer les quantités floues. Comme pour les méthodes de la seconde classe, elles sont principalement utilisables dans le cadre du contrôle flou.

Nous souhaitons, afin d'explorer un ensemble de quantités floues, définir pour chacune de ces quantités un vecteur d'évaluation la représentant dans le processus de visualisation.

2.3 Construction du vecteur multidimensionnel d'évaluation de quantité floue

Afin de construire un vecteur simple de conception et constitué de méthodes de chaque classe, nous proposons de n'utiliser que des méthodes de défuzzification présentes dans [18] et qui ne prennent pas en considération de paramètre autre que l'ensemble à considérer.

Pour la première classe, nous avons choisi les méthodes suivantes : le "first of maximum" (FOM) qui retourne le plus petit élément du cœur d'une quantité floue ; le "last of maximum" (LOM) qui renvoie le plus grand élément du cœur d'une quantité floue ; le "middle of maximum" (MOM) qui permet de récupérer l'élément médian du cœur d'une quantité floue.

Pour la seconde classe, nous avons sélectionné les méthodes suivantes : le "center of gravity" (COG) qui donne le centre de gravité de la fonction d'appartenance d'une quantité floue ; le "mean of maxima" (MeOM) qui calcule la moyenne du cœur d'une quantité floue ; le "mean of support" (MeOS) par lequel on obtient la moyenne du support d'une quantité floue.

Pour la dernière classe, nous avons pris le "center of area" (COA) car celui-ci nous 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 quantité un vecteur d'évaluation de dimension 7. C'est par le prisme de ce vecteur que nous souhaitons explorer les données. Pour cela, nous utilisons notre

méthode de visualisation, qui est le sujet de la section suivante, sur ces vecteurs.

3 Visualisation de données multidimensionnelles par une image couleur

A cette étape du processus, nous disposons donc d'un tableau de données numériques où les quantités floues sont décrites par 7 variables. Autrement dit, ces quantités floues sont représentées par des vecteurs de \mathbb{R}^7 . C'est sur ce tableau de données que nous allons appliquer une méthode de visualisation orientée-pixel que nous avons déjà utilisée avec succès sur des bases de données classiques, des données simulées ainsi que sur des images de fluorescence X. Cette méthode est décrite avec précision dans [2] et nous allons en présenter ici les grandes lignes. Le principe général consiste à associer un pixel couleur à chaque donnée puis à organiser spatialement ces pixels sous forme d'une image couleur.

3.1 Réduction des vecteurs d'évaluation

La première tâche est d'abord d'effectuer une analyse en composantes principales de ce tableau. Le premier intérêt est de réduire la dimensionnalité des données et d'éviter ainsi les problèmes inhérents aux grandes dimensions [8, 3]. Le second objectif est de « préparer » les données avant de les transformer en pixels couleurs. On réalise donc cette ACP en conservant les trois premières composantes. La dimensionnalité est donc réduite à 3 et chaque quantité floue est alors décrite par un vecteur de \mathbb{R}^3 .

3.2 Affectation d'une couleur à chaque vecteur

Dans l'étape qui suit, ces trois composantes sont transformées linéairement pour former trois composantes chromatiques définissant ainsi les pixels couleurs associés. Cette transformation correspond à l'inverse de celle proposée par Ohta, Kanade et Sakai dans [15]. Dans ces tra-

vaux, les auteurs construisent une transformation linéaire qui permet d'obtenir, à partir d'une image couleur « naturelle », une bonne approximation des résultats obtenus à l'aide d'une ACP sur les composantes chromatiques de cette image. Les couleurs générées à partir des composantes principales à l'aide de notre transformation sont donc en quelque sorte construites pour permettre à l'oeil humain de discriminer au mieux les informations apportées par les données. De cette manière, les couleurs ne sont pas affectées selon une échelle de couleurs arbitraire mais calculées à partir de la description des données. Cette construction est donc guidée par les données elles-mêmes.

Si on désigne par C_1 , C_2 et C_3 les composantes principales et R , V , B les composantes couleurs, nous avons la relation suivante :

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

3.3 Construction de l'image de l'ensemble de données

À ce stade nous avons donc associé une couleur à chaque quantité floue. La dernière phase permet enfin d'organiser spatialement ces pixels sous forme d'une image. Après avoir trié les pixels en utilisant comme clé les scores sur les axes principaux, nous les avons placés dans une image en utilisant une courbe de remplissage de Peano-Hilbert.

La courbe de Peano-Hilbert constitue le moyen le plus classique pour effectuer cette construction [14] (voir sur la Figure 1 un exemple de construction d'une telle courbe). Contrairement à un remplissage ligne par ligne (par exemple), les courbes de Peano ont des propriétés intéressantes parmi lesquelles [14] : deux points, qui sont proches sur la courbe, se retrouvent proches dans l'image (propriété de regroupement).

Avec ces deux étapes de tri des données puis de remplissage de l'image par une courbe de

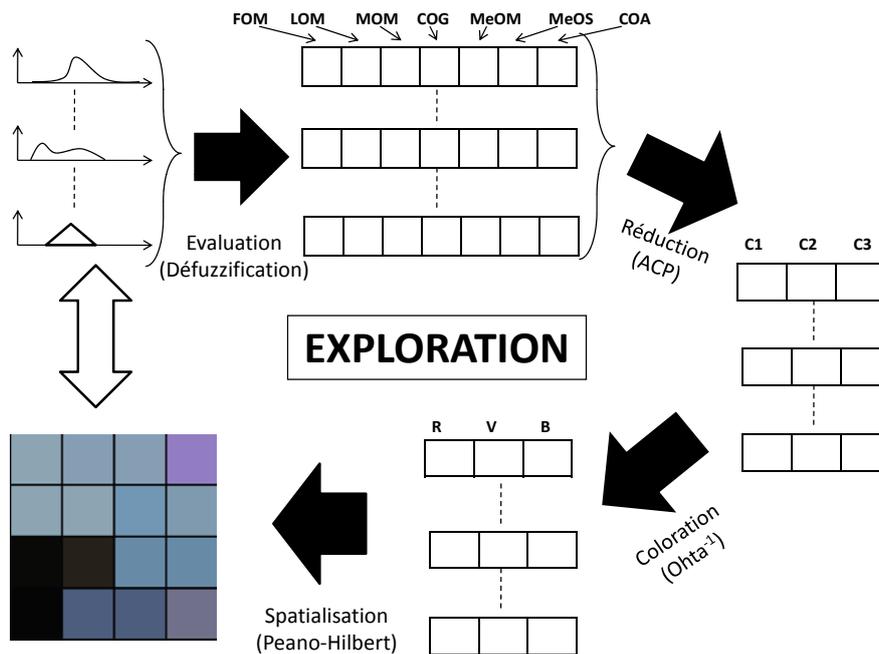


Figure 2 – Exploration d'ensemble de quantités floues - schéma récapitulatif

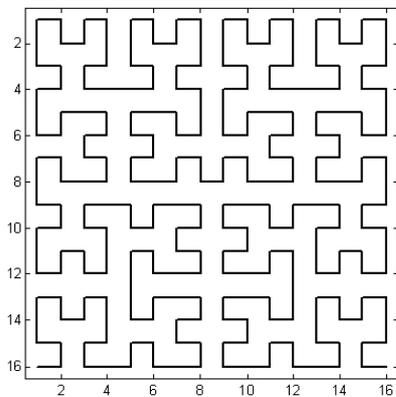


Figure 1 – Remplissage par une courbe de Peano-Hilbert

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.

A l'issue de ce processus de visualisation,

résumé dans la figure 2, nous obtenons une image couleur représentant les quantités floues, relative aux défuzzifieurs retenus. La couleur permet de discriminer l'information de structuration et la spatialisation permet d'observer de manière immédiate cette information. L'image obtenue permet de révéler de façon immédiate et synthétique, des informations de structuration de l'ensemble observé. Nous allons maintenant utiliser notre méthode dans le cadre du projet SIGRem.

4 Application dans le cadre du 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 Rèmes¹, le projet *SIGRem* [16], 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 Système d'Information Géographique (SIG) pluridisciplinaire [17]. Il relève d'une ambition scientifique puisant ses outils conceptuels dans la recherche fondamentale, ses méthodes opérationnelles dans les technologies informatiques en matière d'analyse spatiale et son application pratique dans la mise en valeur des données archéologiques recueillies durant les trente dernières années.

Dans cette partie, nous présenterons en premier la base de données *BDRues*, partie intégrante du projet *SIGRem*. Cette base est dédiée aux éléments de rues romaines à Reims, *BDRues*. Les objets, qui y sont stockés, ont chacun une période d'activité représentée par un ensemble flou convexe et normalisé. Pour chacun de ces ensembles, nous allons déterminer son vecteur d'évaluations. Nous étudierons alors la visualisation obtenue grâce à notre méthode.

4.1 A propos de *BDRues*

Les données archéologiques sont des données spatio-temporelles, ce qui diffère des cas classiques des données géographiques. Quelques études, telles que [9], s'approchent conceptuellement de notre cadre de travail. Dans la base de données sur les rues de Durocortorum², les tronçons de rues sont caractérisés notamment par 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...). De plus, la codification linguistique de périodes temporelles n'a pas toujours la même représentation. Par exemple l'estimation du début du Bas Empire varie selon les experts entre 193 et 284 après J.C. Elle est donc largement imprécise.

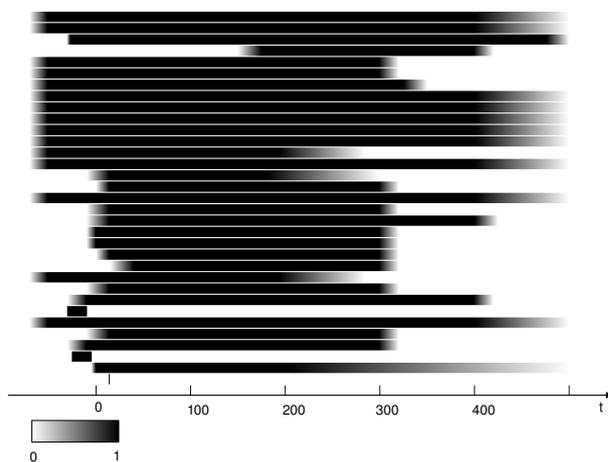


Figure 3 – Périodes floues d'activité des objets de *BDRues* (Chaque ensemble flou -chaque période- est représenté par une "bande". Le niveau de gris correspond au degré d'appartenance et l'abscisse au temps).

Nous représentons les périodes d'activité par des ensembles flous convexes et normalisés (généralement des intervalles flous). On peut ainsi prendre en compte cette imprécision. Une représentation visuelle de ces ensembles est proposée dans la Figure 3.

Nous déterminons ensuite pour chacun de ces ensembles son vecteur d'évaluations. Ces vecteurs sont de dimension 7 ce qui est difficile à visualiser et exploiter de manière intuitive. Notre méthode a pour objectif de le permettre.

4.2 Visualisation des périodes d'activité floues

L'image résultante est présentée sur la figure 4. Elle contient 33 pixels en couleurs. Chaque pixel représente une période d'activité floue. L'organisation spatiale et la couleur des pixels permettent d'observer de façon immédiate

¹Cité des Rèmes : Reims et ses environs à l'époque romaine

²Durocortorum : Reims à l'époque romaine

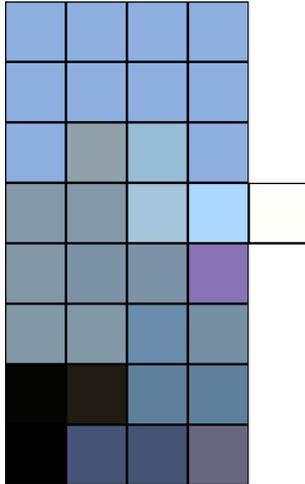


Figure 4 – Visualisation des périodes floues d’activité des objets de *BDRues* par une image couleur (chaque pixel -chaque carré- couleur représente une période floue).

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 périodes d’activité de profil proche. En effet, les périodes d’activité dont les supports sont les plus larges sont représentées par des pixels de couleur bleue claire (haut de l’image de visualisation), tandis que celles dont les cœurs sont de cardinalité moyenne sont colorisées dans les gris (milieu de l’image).

On observe par ailleurs que les composantes principales calculées sur l’ensemble des vecteurs décrivant les périodes d’activité floues issus de *BDRues* permet d’expliquer plus de 99% de la variance totale (voir figure 5). 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 *BDRues*.

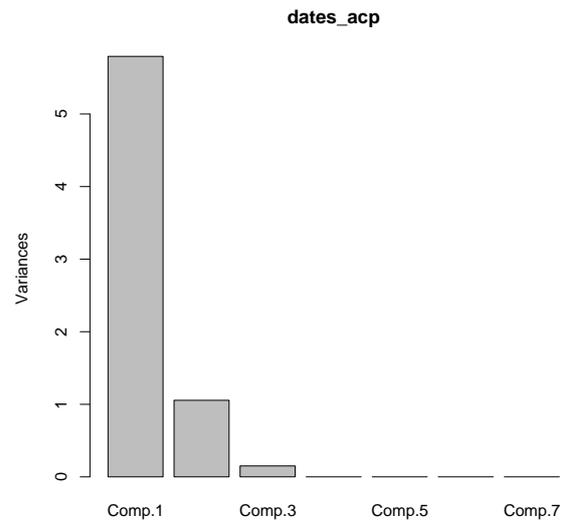


Figure 5 – Éboulis des valeurs propres de l’ACP

5 Conclusion et perspectives

Nous avons présenté dans ce papier une méthode originale d’exploration visuelle d’un ensemble de quantités floues. Cette méthode se base sur la construction de vecteurs dont les valeurs sont obtenues par plusieurs défuzzifications des quantités. L’étape de visualisation consiste à affecter à chaque quantité une couleur pour obtenir des pixels que l’on organise spatialement dans une image. Dans ce but, on réduit ces vecteurs par une ACP à des vecteurs de dimension 3. Par la transformée inverse de celle d’Ohta *et al.* on calcule les couleurs des pixels représentant les quantités. L’image est alors construite en utilisant une courbe de Peano-Hilbert. L’image résultante fournit une carte synthétique de l’ensemble étudié.

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 de fouille qui présente d’autant plus d’intérêt que le nombre de données augmente (il offre la possibilité de visualiser plusieurs millions de données).

Nous avons étudié l'application de cette méthode sur des quantités floues particulières : des ensembles convexes et normalisés représentant les périodes d'activité des objets de *BDRues*. Dans cette application on a pu remarqué que les trois principales composantes de l'ACP sur les vecteurs contenaient la quasi totalité de l'information obtenue par les différentes méthodes de défuzzification. L'image résultante de la procédure fait apparaître visuellement des regroupements dans l'ensemble des quantités floues. Cette exploration est donc intuitive et simple d'utilisation.

La particularité des ensembles flous de l'application font que deux des quantifieurs (MOM et MeOM) fournissent exactement les mêmes résultats. Par contre, cela ne peut expliquer complètement que l'essentiel de l'information qu'ils fournissent puisse être contenu dans uniquement 3 composantes. Dans des études futures, nous approfondirons les rapports entre les quantifieurs sur des jeux de données plus variés et plus importants. En effet, au vu de certains indices préliminaires, nous souhaiterions vérifier de manière statistique les relations entre quantifieurs.

Références

- [1] J.M. Adamo. Fuzzy Decision trees. *Fuzzy Sets and Systems*, 4 : 207-219 , 1985.
- [2] F. Blanchard, L. Lucas, M. Herbin. A New Pixel-Oriented Visualization Technique through Color Image. *Information Visualization*, 4(4) : 257-265, 2005.
- [3] F. Camastra. Data dimensionality estimation methods : a survey. *Pattern Recognition*, 36 : 2945-2954, 2003.
- [4] C. de Runz, E. Desjardin, M. Herbin, F. Piantoni. A new Method for the Comparison of two fuzzy numbers extending Fuzzy Max Order. *Actes d'IPMU, 2006*, pp. 127-133.
- [5] C. de Runz, E. Desjardin, F. Piantoni, M. Herbin. Management of multi-modal data using the Fuzzy Hough Transform : Application to archaeological simulation. *Actes de Research Challenge in Information Science, 2007*, pp. 351-356.
- [6] C. de Runz, E. Desjardin, F. Piantoni, M. Herbin. Using fuzzy logic to manage uncertain multi-modal data in an archaeological GIS. *Actes d'ISSDQ, 2007*, <http://www.itc.nl/ISSDQ2007/proceedings/>.
- [7] C. de Runz, E. Desjardin, F. Piantoni, M. Herbin. Toward handling uncertainty of excavation data into a GIS. *Actes de CAA, 2008* à paraître.
- [8] D.L. Donoho. High-Dimensional Data Analysis : The Curses and Blessings of Dimensionality. *Actes de AMS Conference Mathematical Challenges of the 21st Century, 2000*.
- [9] S. Dragicevic, D. J. Marceau. An application of fuzzy logic reasoning for GIS temporal modeling of dynamic processes. *Fuzzy Sets and Systems*, 113 : 69-80, 2000.
- [10] G. Facchinetti. Ranking functions induced by weighted average of fuzzy numbers *Fuzzy Optimization and Decision Making*, 1(3) : 313-327, 2002.
- [11] G. Facchinetti, N. Pacchiarotti. Evaluation of fuzzy quantities *Fuzzy Sets and Systems*, 157(7) : 892-903, 2006.
- [12] D.A. Keim. Designing Pixel-Oriented Visualization Techniques : Theory and Applications. *IEEE Trans. Visualization and Computer Graphics*, 6(1) : 1-20, 2000.
- [13] E.E. Kerre. The use of fuzzy set theory in electrocardiological diagnostics. *Dans Approximate Reasoning in Decision-Analysis*. North-Holland Publishing Company. pp. 277-282, 1982.
- [14] B. Moon, H.V. Jagadish, C. Faloutsos, 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.
- [15] Y. Ohta, T. Kanade, T. Sakai. Color Information for Region Segmentation. *Computer Graphics and Image Processing*, 13 : 222-241, 1980.
- [16] D. Pargny, F. Piantoni. SIGRem : un Système d'Information Géographique pour l'archéologie en Champagne-Ardenne. *Actes du colloque Archéologie en Champagne-Ardenne, 2005*.
- [17] F. Piantoni. Le SIGRem. Problématique et méthodologie. *Actes du séminaire de recherche du CIRAR, 2005*.
- [18] W. Van Leekwijck, E.E. Kerre. Defuzzification : criteria and classification *Fuzzy Sets and Systems*, 108 : 159-178, 1999.
- [19] X. Wang, E.E. Kerre.. Reasonable properties for the ordering of fuzzy quantities (I). *Fuzzy Sets and Systems*, 118 : 375-385, 2001.