

Fouilles archéologiques : à la recherche d'éléments représentatifs

Cyril De Runz*, Frédéric Blanchard*
Eric Desjardin*, Michel Herbin*,**

*CReSTIC-SIC, IUT de Reims Châlons Charleville,
Rue des Crayères, BP 1035, Reims Cedex 2, France
{cyril.de-runz, frederic.blanchard, eric.desjardin}@univ-reims.fr,
<http://crestic.univ-reims.fr>

**Antenne CReSTIC-Châlons, chaussée du port,
BP 541, 51012 Châlons-en-Champagne Cedex, France
michel.herbin@univ-reims.fr

Résumé. Définir les éléments les plus représentatifs au sein d'un SIG archéologique est une question d'actualité. En effet, déterminer l'élément qui représente le mieux un ensemble de fouilles archéologiques est important pour la valorisation de ces fouilles. Nous avons développé au sein du CReSTIC-SIC une méthode statistique de sélection de l'élément le plus représentatif d'un échantillon. La notion -quantitative- de représentativité est basée sur la transformation par rangs des dissimilarités entre éléments. Nous avons appliqué cette méthode statistique à la recherche de l'élément le plus représentatif au sein de données issues de fouilles, dont les caractéristiques sont représentées par des ensembles flous convexes et normalisés, sur les rues de la ville de Reims à l'époque romaine. Dans ce cadre, nous avons donc utilisé une métrique classique entre ensembles flous en tant que dissimilarité pour déterminer le tronçon de rue le plus représentatif de Reims à cette époque.

1 Introduction

Les Systèmes d'Information Géographique (SIG) permettent de stocker et de visualiser à la fois les objets spatiaux et les informations associées. 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 architecturalement représentatif des boulevards parisiens, le résultat serait vraisemblablement un boulevard de type haussmanien. En archéologie, dans l'optique de valorisation des éléments découverts durant des fouilles, il s'avère intéressant de déterminer celui qui représente le mieux les éléments découverts en terme de localisation, de période d'activité, de forme. Nous disposons dans le cadre du projet SIGRem (de Runz et al. (2007a)) d'une Base de Données Géographiques (BDG) intitulée *BDRues* sur les tronçons de rues de Durocortorum (Reims à l'époque Romaine). Nous

cherchons dans ce cadre le tronçon de rue le plus représentatif spatio-temporellement. Au sein du groupe SIC (Signal Image et Connaissance) du CReSTIC (Centre de Recherche en Sciences et Technologies de l'Information et de la Communication), nous avons défini une statistique (Blanchard (2005)) pour déterminer l'élément le plus représentatif d'un échantillon de données. Nous proposons dans cette communication d'utiliser cette statistique sur *BDRues* afin de déterminer les tronçons de rues les plus représentatifs en fonction de leurs périodes d'activité, de leurs localisations et de leurs orientations.

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 connaissent depuis quelques années un regain d'intérêt (David et Ngaraja (2003)) et 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 (Galambos (1975); Barnett (1976)). Ces statistiques amplement utilisées dans des domaines tels que le traitement d'images (Lukac et al. (2006); Vautrot et al. (2006)) ne le sont que peu dans l'exploitation des SIG archéologiques pour la valorisation. Nous proposons ici d'utiliser le Vecteur de Meilleur Rang Moyen (VMRM) (de Runz et al. (2007d)), qui extrait l'élément de représentativité maximale d'un ensemble de données.

La définition quantitative de représentativité, utilisée par le VMRM nécessite de disposer d'un indice de dissimilarité entre éléments. Cette dissimilarité est liée à la description des données. Or comme nos données archéologiques sont incertaines dans leurs localisations, orientations, et datations, nous avons précédemment (de Runz et al. (2007b,c)) modélisé, dans *BDRues*, ces différentes caractéristiques par des ensembles flous convexes et normalisés. C'est pourquoi, nous proposons d'utiliser une métrique classique (Grzegorzewski (1998)) entre ensembles flous convexes et normalisés comme dissimilarité afin de déterminer le VMRM dans *BDRues*.

Après avoir présenté le principe théorique de Vecteur de Meilleur Rang Moyen, son utilisation dans le contexte des rues de Durocortorum est proposée. Une discussion et des perspectives sont enfin exposées avant de conclure.

2 Le vecteur de meilleur rang moyen

Considérons un échantillon de données multidimensionnelles $S = \{x_1, x_2, \dots, x_n\}$ dans un espace E de dimension $p \in \mathbb{N}$. 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 notre échantillon S . 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.

2.1 Statistiques de rangs marginales

On notera X_1, X_2, \dots, X_n les variables (vecteurs) aléatoires dont les éléments de l'échantillon S sont les observations. Les statistiques d'ordres associées sont les $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ triés par ordre croissant (et on notera $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ les observations ordonnées associées). Par définition, les statistiques d'ordres sont donc intrinsèquement

liées à la façon dont sont triées les variables aléatoires. Dans le cas multidimensionnel, le tri n'est pas trivial, on se reportera à Barnett (1976) pour une étude des différentes techniques pour trier des vecteurs.

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 , nous classons 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 plus similaires 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 S la plus similaire 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.

2.2 Rang moyen d'une donnée

On calcule ensuite, pour chaque donnée x_j de S , la moyenne des rangs $Rg_{x_i}(x_j)$ qu'elle a obtenus 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 nous 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 plus similaire* à l'ensemble des autres. On appelle cette notion la *représentativité d'une donnée dans son échantillon*.

2.3 Statistique de meilleur rang moyen

Nous terminons notre processus par la recherche dans l'échantillon, de la donnée ayant le plus petit rang moyen, c'est à dire la donnée de l'échantillon la plus représentative dudit échantillon.

Finalement, nous avons donc, au cours de ces étapes, défini une statistique exprimée comme une fonctionnelle des statistiques de rangs marginales et notée *VMRM* (Vecteur de Meilleur Rang Moyen) :

$$VMRM : (X_1, X_2, \dots, X_n) \mapsto \underset{X_i, i=1..n}{argmin} (\overline{Rg}(X_j))$$

Cette statistique associe à un échantillon l'élément qui le représente le mieux. Cette notion de meilleur représentant d'un échantillon rejoint, d'un point de vue sémantique, la notion de représentant de classes en classification automatique des données. De plus, au même titre que la médiane, notre statistique est un estimateur robuste de position de l'échantillon. En effet, la donnée de meilleur rang moyen est un élément typique et représentatif de l'échantillon.

Dans la partie suivante, nous abordons l'utilisation du VMRM dans *BDRues* afin de déterminer les tronçons de rues trouvés de Durocorturum les plus représentatifs en terme de localisation, de période d'activité, d'orientation et des trois cumulées.

Fouilles archéologiques : à la recherche d'éléments représentatifs

3 Tronçons de rues romaines les plus représentatifs

Trouver l'objet représentant le mieux un ensemble d'objets issus de fouilles archéologiques peut avoir un intérêt fort pour la valorisation du travail de fouilles en archéologie. Cet objet représente l'objet le plus classique trouvé au cours des fouilles en fonction des critères choisis. Dans cette partie, nous présenterons d'abord la BDG dédiée aux éléments de rues romaines à Reims, *BDRues*, qui nous servira de base de travail. Ensuite, notre travail se penchera sur la ou les mesures de dissimilarité choisies. Enfin nous étudierons les résultats issus de ce travail.

3.1 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 Dragicevic et Marceau (2000), s'approchent conceptuellement de notre cadre de travail. Dans la base de données sur les rues de Durocortorum, 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...). 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 incertaine et imprécise. Le géoréférencement est lui aussi sujet à de l'imprécision et/ou de l'incertitude liées 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 soit respectivement par des nombres flous, des intervalles flous et ensembles flous spatiaux (2D). On peut ainsi prendre en compte cette incertitude (voir Figure 1).

Afin de pouvoir obtenir les vecteurs de meilleurs rangs moyens en terme de localisation, orientation, période d'activité et des trois conjuguées, nous devons définir les mesures de dissimilarités associées.

3.2 Détermination de la dissimilarité

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 :

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

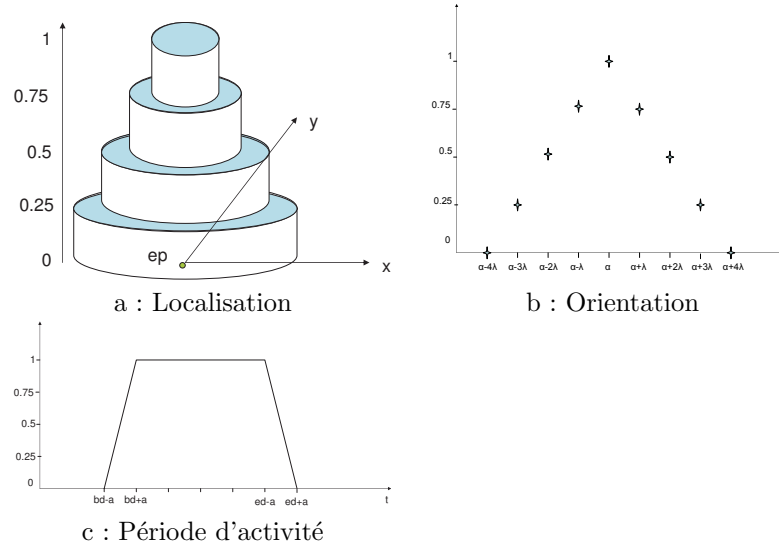


Fig. 1 – Modèles flous pour la localisation, l'orientation et les périodes d'activité des rues romaines

Nous utiliserons 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 (voir Figure 2).

À l'instar des mesures de dissimilarités pour le calcul des VMRM en terme de localisation, d'orientation ou de période d'activité, nous avons besoin de prendre une mesure de dissimilarité pour l'ensemble des caractéristiques. Afin de l'obtenir, nous normalisons la dissimilarité liant un objet à un autre par la dissimilarité maximale du premier objet à l'échantillon. Nous effectuons 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 éléments X et Y de $BDRues$ alors :

$$D_{global}(X, Y) = \frac{1}{3} \times \sum_{i \in \{loc, orien, date\}} \frac{D_i(F, G)}{\max_{J \in BDRues} D_i(F, J)}$$

Nous utilisons cette dissimilarité afin d'extraire le VMRM global.

3.3 Résultats

En calculant le VMRM pour la localisation, pour l'orientation, pour la période d'activité et les trois conjuguées nous obtenons les plans de la Figure 3.

Fouilles archéologiques : à la recherche d'éléments représentatifs

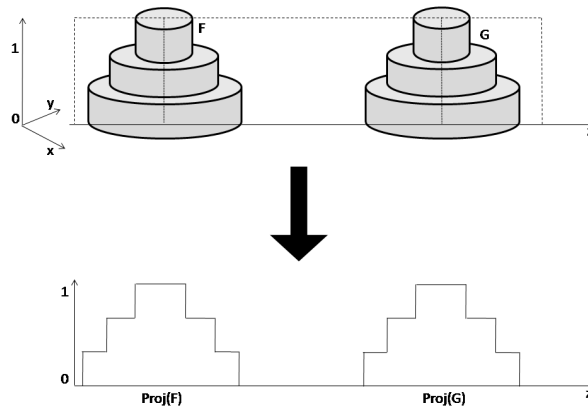


FIG. 2 – Projections pour le calcul de dissimilarité des localisations

On peut s'apercevoir que les tronçons représentatifs sont différents (excepté pour le VMRM global et le VMRM date) en fonction de la recherche effectuée, et en cela reflète bien l'influence de chacune des caractéristiques des données archéologiques. Ainsi, nous observons 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 entraîné le décalage au centre du VMRM Global par rapport au VMRM Localisation. Si nous ne regardons 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 4) qui se situe dans l'âge d'or de la période romaine de Reims (début du gallo-romain - fin du Bas-Empire).

4 Discussion et conclusion

L'extraction et la visualisation d'éléments représentatifs au sein d'un SIG sont importantes pour la valorisation des données et plus encore dans le cadre de données de fouilles archéologiques. Les données de fouilles sont incertaines, nous avons donc choisi de les représenter par des ensembles flous. Nous avons dans des précédents travaux défini une statistique de rangs permettant d'extraire d'un échantillon de données, l'élément de meilleur rang moyen. Nous l'avons utilisée dans le cadre des données de fouilles sur les rues de Reims à l'époque romaine pour en extraire les tronçons trouvés de rues romaines les plus représentatifs en terme de date, de localisation et/ou d'orientation. Ce travail constitue la première étape d'extraction d'éléments caractéristiques d'un échantillon de données issues de fouilles archéologiques.

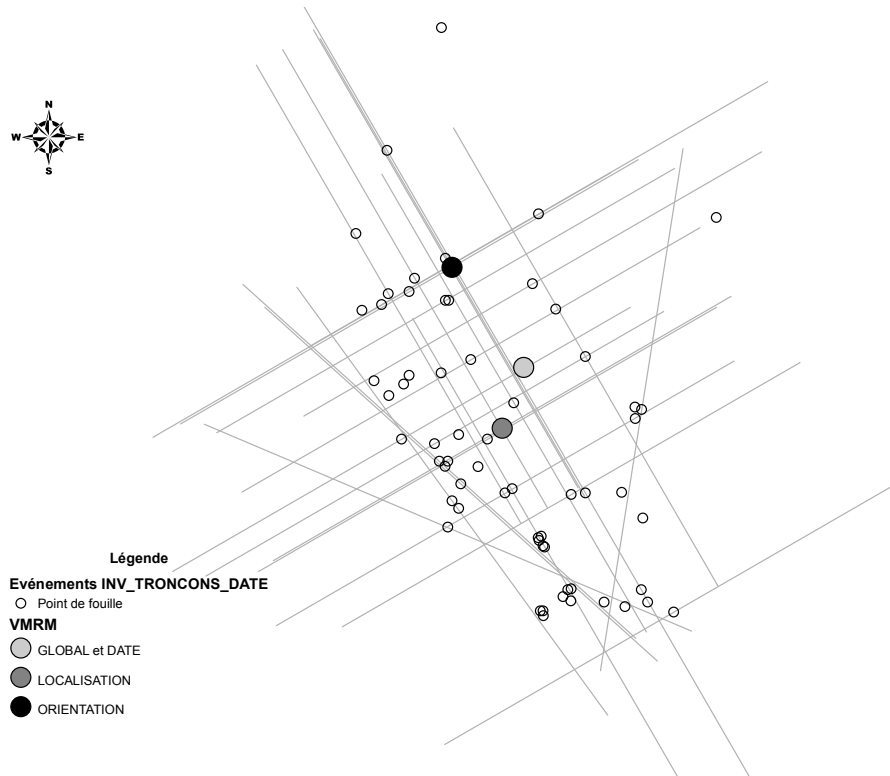


FIG. 3 – Différents VMRM pour BDRues

En effet, l'élément le plus représentatif d'un jeu de données est l'élément le plus similaire aux autres, mais pas forcément le plus emblématique. Ainsi, bien que, pour le nombre de visiteurs, la Tour Eiffel soit l'élément le plus emblématique des monuments de Paris, il ne sera vraisemblablement pas le plus représentatif. Dans de futurs travaux, nous nous attacherons donc à la détermination de techniques permettant d'extraire d'autres éléments typiques en fonction de leurs propriétés.

Remerciements

Nous tenons à remercier le Service Régional d'Archéologie de Champagne-Ardenne et le centre rémois de l'Institut National de Recherche en Archéologie Préventive pour nous avoir permis d'accéder à leurs données. Nous tenons de même à souligner la contribution de Dominique Pargny, ingénieur d'études au laboratoire GEGENA, et de Frédéric Piantoni, Maître de Conférences au laboratoire HABITER, au projet SIGRem, porté par l'Université de Reims Champagne-Ardenne, sur lequel se base ce travail.

Fouilles archéologiques : à la recherche d'éléments représentatifs

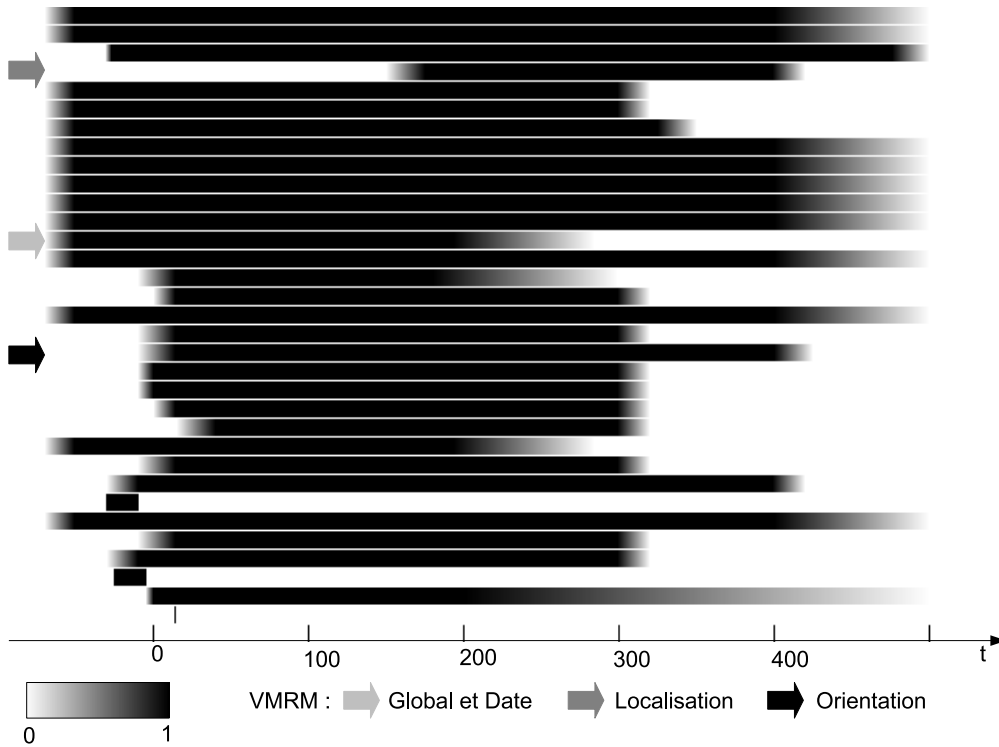


FIG. 4 – Représentation des périodes d'activités des objets de BDRues

Références

- Barnett, V. (1976). The ordering of multivariate data. *Journal of the Royal Statistical Society, Series A (General)* 139(3), 318–355.
- Blanchard, F. (2005). *Visualisation et classification de données multidimensionnelles. Application aux images multicomposantes*. Thèse de doctorat, Université de Reims Champagne-Ardenne, France.
- David, H. A. et H. N. Nagaraja (2003). *Order Statistics* (Third ed.). Wiley.
- de Runz, C., E. Desjardin, M. Herbin, D. Pargny, F. Piantoni, et F. Berthelot (2007a). Simulation de cartes et prédictivité à partir de données archéologiques traitées par sig. *Culture et Recherche* (111), 35–35.
- de Runz, C., E. Desjardin, F. Piantoni, et M. Herbin (2007b). Management of multimodal data using the fuzzy hough transform : Application to archaeological simulation. In *First International Conference on Research Challenges in Information Science*, Maroc, Ouarzazate, pp. 351–356. Colette Rolland, Oscar Pastor and Jean-Louis Cavarero.

- de Runz, C., E. Desjardin, F. Piantoni, et M. Herbin (2007c). Using fuzzy logic to manage uncertain multi-modal data in an archaeological gis. In *International Symposium on Spatial Data Quality*, Pays-Bas, Enschede.
- de Runz, C., M. Herbin, F. Blanchard, L. Hussenet, V. Vrabie, et P. Vautrot (2007d). Le vecteur de meilleur rang moyen : une statistique pour l'analyse de données multidimensionnelles - application au filtrage d'images couleurs. In *GRETSI*, France, Troyes.
- Dragicevic, S. et D. Marceau (2000). An application of fuzzy logic reasoning for gis temporal modeling of dynamic processes. *Fuzzy Sets and Systems* 113, 69–80.
- Galambos, J. (1975). Order statistics of samples from multivariate distributions. *Journal of the American Statistical Association* 70(351), 674–680.
- Grzegorzewski, P. (1998). Metrics and orders in space of fuzzy numbers. *Fuzzy Sets and Systems* 97, 83–94.
- Lukac, R., B. Smolka, K. Plataniotis, et A. Venetsanopoulos (2006). Vector sigma filters for noise detection and removal in color images. *J. Vis. Commun. Image R.* 17, 1–26.
- Vautrot, P., L. Hussenet, et M. Herbin (2006). A robust filtering method using owa filters : Application to color images. In *3rd European Conference on Colour in Graphics, Imaging and Vision*, University of Leeds, UK.

Summary

In Geographical Information Systems devoted to archeology, the selection of the most representative element of the Geographical Database is important to increase the value of city excavations. We developed, in the CReSTIC-SIC, a statistical method to select the most representative element of a dataset. This method is based on ranking statistics, but also on a dissimilarity measure between data to determine ranks. We apply it on the extraction of the most representative elements of excavation data. In the GIS about the Roman streets in Reims, the features of the information are modeled by convex and normalized fuzzy sets; hence we use a classical metric between convex and normalized fuzzy sets to measure the dissimilarity. This allows us to extract the most representative Roman street in Reims.