Positionnement multidimensionnel de données de dissimilarités floues

Multidimensional scaling of fuzzy dissimilarity data

Mylène Masson Thierry Denoeux

Heuristique et Diagnostic des Systèmes Complexes – UMR CNRS 6599
Université de Technologie de Compiègne
BP 529 - F-60205 Compiègne cedex - France

Résumé

Les techniques de positionnement multidimensionnel permettent de représenter des objets sous forme de points dans un espace à partir de mesures de leur dissimilarité. Dans cet article, on considère que les dissimilarités sont exprimées sous la forme de nombres flous. Chaque objet est alors associé à une région floue dans $\mathbb{R}^p$*. Inspirés des techniques de régression floue, un modèle des moindres carrés, un modèle de possibilité et un modèle de nécessité sont proposés. Un exemple illustre la capacité des modèles à retrouver correctement la structure des objets mais aussi à fournir dans la représentation même des indicateurs sur la précision des données d'entrée.*

Mots Clef

Positionnement multidimensionnel, analyse de données floues, dissimilarités floues.

Abstract

Multidimensional scaling is a technique for representing measurements of dissimilarity among objects as points in a p-dimensional space. In this paper, this method is extended to the case where dissimilarities are expressed as fuzzy numbers. Each object is represented as a fuzzy region in $\mathbb{R}^p$*. Using ideas from fuzzy regression, three models are proposed: a least-squares model, a possibility model and a necessity model. An example illustrates the ability of these models to recover the structure of the objets set and to provide indicators about the precision of dissimilarity measurements.*

Keywords

Multidimensional scaling, fuzzy data analysis, fuzzy dissimilarities.

1 Introduction

Les méthodes de positionnement multidimensionnel sont des méthodes statistiques exploratoires qui permettent l'analyse de données de dissimilarités [1], [2]. Classiquement, on dispose d'une matrice symétrique $n \times n$ de dissimilarités entre objets $\Delta = (\delta_{ij})$, et l'on cherche une configuration de points dans un espace euclidien de dimension p de telle sorte que la matrice $n \times n$ des distances inter-points $D = (d_{ij})$ approche au mieux la matrice d'entrée Δ. Parmi les méthodes les plus courantes, les méthodes dites métriques déterminent une matrice $n \times p$ de coordonnées $X = (x_{il})$ par minimisation itérative d'une fonction d'erreur (*fonction stress*) du type :

$$\sigma(X) = \sum_{i<j} (d_{ij} - \delta_{ij})^2. \tag{1}$$

Ces méthodes sont notamment utilisées pour l'analyse de données de dissimilarités recueillies au cours de tests sensoriels. Lors de ces tests, un ou plusieurs sujets humains sont chargés d'évaluer, sur une échelle donnée, la différence perçue entre chaque paire d'objets. La tâche des évaluateurs est souvent difficile et dans certains cas, il peut être préférable, plutôt que de demander une valeur précise nécessairement entachée d'erreur, d'autoriser une certaine imprécision dans les réponses. Dans cet article, on considère que la réponse d'un sujet peut être fournie sous forme d'un intervalle ou, de façon plus générale, sous forme d'un nombre flou.

Pour traiter ce type de données, nous proposons d'associer à un objet non plus un point précis, mais une région nette ou floue de l'espace $\mathbb{R}^p$, qui tienne compte à la fois de l'incertitude et de l'imprécision contenue dans les données d'entrée. Plusieurs modèles sont proposés : un modèle des moindres carrés, un modèle dit de possibilité et un modèle dit de nécessité. Signalons que cette terminologie n'est due qu'à l'analogie des algorithmes présentés avec ceux développées en régression floue [5] [8] et ne relève pas à proprement parler de la théorie des possibilités.

2 Dissimilarités de type intervalle

On suppose ici que les données disponibles consistent en une matrice $\Delta = ([\delta_{ij}])$ de données de dissimilarités de type intervalle. Chaque intervalle $[\delta_{ij}] = [\delta_{ij}^-, \delta_{ij}^+]$ représente l'ensemble des valeurs plausibles pour la dissimilarité entre l'objet i et l'objet j. A chaque objet i, on associe non plus un point précis dans l'espace, mais une région R_i. Notons d_{ij}^- et d_{ij}^+ respectivement les distances Euclidiennes minimum et maximum entre deux régions définies par :

$$d_{ij}^- = \min_{\mathbf{x}_i \in R_i, \mathbf{x}_j \in R_j} ||\mathbf{x}_i - \mathbf{x}_j|| \quad (2)$$

$$d_{ij}^+ = \max_{\mathbf{x}_i \in R_i, \mathbf{x}_j \in R_j} ||\mathbf{x}_i - \mathbf{x}_j||. \quad (3)$$

Si chaque région R_i est une sphère de centre $\mathbf{c}_i$ et de rayon r_i, alors :

$$d_{ij}^- = \max(0, d_{ij} - r_i - r_j) \quad (4)$$

$$d_{ij}^+ = d_{ij} + r_i + r_j, \quad (5)$$

avec $d_{ij} = ||\mathbf{c}_i - \mathbf{c}_j||$. Il s'agit alors de déterminer les centres et les rayons des différents objets de telle sorte que l'intervalle de distance représente, dans un certain sens, l'intervalle de dissimilarité. Suivant la nature du lien imposé entre les deux intervalles, différents algorithmes peuvent être proposés.

2.1 Modèle des moindres carrés

La solution la plus immédiate consiste à généraliser la fonction stress (1) sous la forme suivante :

$$\sigma'(\mathcal{R}) = \sum_{i<j} (d_{ij}^- - \delta_{ij}^-)^2 + \sum_{i<j} (d_{ij}^+ - \delta_{ij}^+)^2, \quad (6)$$

où $\mathcal{R}$ représente l'ensemble des régions $\{R_1, R_2, ..., R_n\}$, et à minimiser $\sigma'(\mathcal{R})$ par rapport à $\mathcal{R}$ par une méthode de descente de gradient. Notons que pour assurer la condition $r_i \geq 0$, il est possible de reparamétrer le modèle en introduisant de nouvelles variables ρ_i telles que $r_i = \rho_i^2$.

2.2 Modèle de possibilité

Le modèle précédent donne une représentation qui approche « au mieux » (au sens des moindres carrés) les dissimilarités. On cherche ici à proposer une représentation *exacte* des données d'entrée. Supposons que les centres des régions aient été obtenus par minimisation de (1) par exemple. Par analogie avec les modèles de régression sur données de type intervalle, on peut chercher les rayons des sphères les plus petits possibles vérifiant les contraintes :

$$d_{ij}^- \leq \delta_{ij}^- \leq \delta_{ij}^+ \leq d_{ij}^+ \quad \forall i, j. \quad (7)$$

Cette stratégie conduit à formuler le programme linéaire suivant :

$$\min_{\mathbf{r}} \sum_{i=1}^{n} r_i \quad (8)$$

sous les contraintes :

$$r_i + r_j \geq \max(d_{ij} - \delta_{ij}^-, \delta_{ij}^+ - d_{ij}) \quad \forall i, j \quad (9)$$

$$r_i \geq 0 \quad \forall i = 1, n \ , \quad (10)$$

où $\mathbf{r}$ représente le vecteur $(r_1, \ldots, r_n)$. Il est facile d'observer que le programme linéaire précédent a toujours une solution réalisable quelles que soient les données d'entrée. Dans le cas particulier où les dissimilarités sont fournies de façon précise ($\delta_{ij}^- = \delta_{ij}^+ = \delta_{ij}$), la contrainte (9) devient :

$$r_i + r_j \geq |d_{ij} - \delta_{ij}|. \quad (11)$$

Les rayons des sphères sont dans ce cas révélateurs des erreurs d'estimation et de modélisation et servent à l'interprétation de la représentation trouvée. Dans le cas plus général de données de type intervalle, les rayons portent une information supplémentaire relative au caractère imprécis des données.

2.3 Modèle de nécessité

De façon analogue, on peut tenter de rendre maximum le volume des sphères tout en respectant la contrainte :

$$\delta_{ij}^- \leq d_{ij}^- \leq d_{ij}^+ \leq \delta_{ij}^+ \quad \forall i, j. \quad (12)$$

On est alors amené à résoudre le programme linéaire suivant :

$$\max_{\mathbf{r}} \sum_{i=1}^{n} r_i \quad (13)$$

sous les contraintes :

$$r_i + r_j \leq \min(d_{ij} - \delta_{ij}^-, \delta_{ij}^+ - d_{ij}) \quad \forall i, j \quad (14)$$

$$r_i \geq 0 \quad \forall i = 1, n. \quad (15)$$

L'existence d'une solution à ce programme linéaire n'est assurée qu'à la condition que :

$$\delta_{ij}^- \leq d_{ij} \leq \delta_{ij}^+ \quad \forall i, j. \tag{16}$$

L'existence du modèle de nécessité est donc un indicateur de l'adéquation du modèle euclidien aux données d'entrée.

3 Dissimilarités floues

L'approche précédente peut sans difficulté être généralisée aux cas où les données de dissimilarité sont des nombres flous $\widetilde{\delta}_{ij}$, modélisant par exemple des appréciations subjectives exprimées sous forme linguistique. Il est dans ce cas naturel d'associer à chaque objet i une représentation floue, sous la forme d'une sous-ensemble flou $\widetilde{R}_i$ de $\mathbb{R}^p$. Le principe d'extension [4] permet alors de définir la distance floue $\widetilde{d}_{ij}$ entre les régions floues $\widetilde{R}_i$ et $\widetilde{R}_j$ par :

$$\mu_{\widetilde{d}_{ij}}(w) = \sup_{\mathbf{x},\mathbf{y} \in \mathbb{R}^p} \min(\mu_{\widetilde{R}_i}(\mathbf{x}), \mu_{\widetilde{R}_j}(\mathbf{y})), \tag{17}$$

le supremum étant calculé sous la contrainte $||\mathbf{x}-\mathbf{y}|| = w$. En particulier, si $\widetilde{R}_i$ et $\widetilde{R}_j$ sont des nombres flous multidimensionnels [6], chaque coupe de niveau α de $\widetilde{d}_{ij}$ est un intervalle fermé ${}^\alpha\widetilde{d}_{ij} = [{}^\alpha\widetilde{d}_{ij}^-, {}^\alpha\widetilde{d}_{ij}^+]$ dont les bornes sont les distances minimale et maximale entre les α-coupes de $\widetilde{R}_i$ et $\widetilde{R}_j$. Par analogie avec l'approche précédente, on considère que les α-coupes de $\widetilde{R}_i$ sont des sphères concentriques de rayons ${}^\alpha r_i$ et de centre $\mathbf{c}_i$ avec :

$$\alpha > \alpha' \Rightarrow {}^\alpha r_i \leq {}^{\alpha'} r_i \quad \forall \alpha, \alpha' \in [0, 1]. \tag{18}$$

On a dans ce cas :

$$\begin{aligned} {}^\alpha\widetilde{d}_{ij}^- &= \max(0, d_{ij} - {}^\alpha r_i - {}^\alpha r_j) & (19) \\ {}^\alpha\widetilde{d}_{ij}^+ &= d_{ij} + {}^\alpha r_i + {}^\alpha r_j, & (20) \end{aligned}$$

où comme précédemment d_{ij} est la distance euclidienne entre les centres $\mathbf{c}_i$ et $\mathbf{c}_j$. On peut alors étendre de manière naturelle les modèles proposés au paragraphe précédent, en déterminant les régions $\widetilde{R}_i$ de manière à ce que les distances floues $\widetilde{d}_{ij}$ représentent le plus fidèlement possible les dissimilarités floues $\widetilde{\delta}_{ij}$ observées.

3.1 Modèle des moindres carrés

Le critère d'erreur défini par l'équation (6) peut être généralisé dans le cas de dissimilarités floues en choisissant c valeurs comprises entre 0 et 1 :

$$1 = \alpha_1 > \ldots > \alpha_c = 0$$

et en posant :

$$\sigma''(\widetilde{\mathcal{R}}) = \sum_{k=1}^{c} \sum_{i<j} ({}^{\alpha_k}\widetilde{d}_{ij}^- - {}^{\alpha_k}\widetilde{\delta}_{ij}^-)^2 + \sum_{k=1}^{c} \sum_{i<j} ({}^{\alpha_k}\widetilde{d}_{ij}^+ - {}^{\alpha_k}\widetilde{\delta}_{ij}^+)^2, \tag{21}$$

où $\widetilde{\mathcal{R}}$ désigne l'ensemble des régions floues $\widetilde{R}_i$, ${}^0\widetilde{x}$ représentant par convention le support d'un nombre flou quelconque $\widetilde{x}$.

Notons que le critère de stress proposé s'apparente au critère des moindres carrés flous proposé par Diamond [3] [5] et étendu par Ming et al. [7].

Pour assurer la condition ${}^\alpha r_i \geq {}^{\alpha'} r_i$ lorsque $\alpha < \alpha'$, on reparamètre le modèle en introduisant de nouvelles variables ${}^{\alpha_k}\rho_i$ de telle sorte que :

$${}^{\alpha_k} r_i = \sum_{h=1}^{k} {}^{\alpha_h}\rho_i^2. \tag{22}$$

3.2 Modèle de possibilité

Pour chaque niveau α_k considéré, on impose :

$${}^{\alpha_k}\widetilde{d}_{ij}^- \leq {}^{\alpha_k}\widetilde{\delta}_{ij}^- \leq {}^{\alpha_k}\widetilde{\delta}_{ij}^+ \leq {}^{\alpha_k}\widetilde{d}_{ij}^+ \quad \forall i, j, k. \tag{23}$$

Le modèle de possibilité s'obtient donc en résolvant successivement c programmes linéaires pour k variant de 1 à c :

$$\min_{{}^{\alpha_k}\mathbf{r}} \sum_{i=1}^{n} {}^{\alpha_k} r_i \tag{24}$$

sous les contraintes :

$${}^{\alpha_k} r_i + {}^{\alpha_k} r_j \geq \max(d_{ij} - {}^{\alpha_k}\delta_{ij}^-, {}^{\alpha_k}\delta_{ij}^+ - d_{ij}) \quad \forall i, j \tag{25}$$

$${}^{\alpha_k} r_i \geq \begin{cases} 0 & \text{si } k = 1 \\ {}^{\alpha_{k-1}} r_i & \text{si } k > 1 \end{cases} \quad \forall i = 1, n. \tag{26}$$

avec ${}^{\alpha_k}\mathbf{r} = ({}^{\alpha_k} r_1, \ldots, {}^{\alpha_k} r_n)$.

3.3 Modèle de nécessité

De façon analogue, pour chaque niveau α_k considéré, on impose :

$${}^{\alpha_k}\widetilde{\delta}_{ij}^- \leq {}^{\alpha_k}\widetilde{d}_{ij}^- \leq {}^{\alpha_k}\widetilde{d}_{ij}^+ \leq {}^{\alpha_k}\widetilde{\delta}_{ij}^+ \quad \forall i, j, k. \tag{27}$$

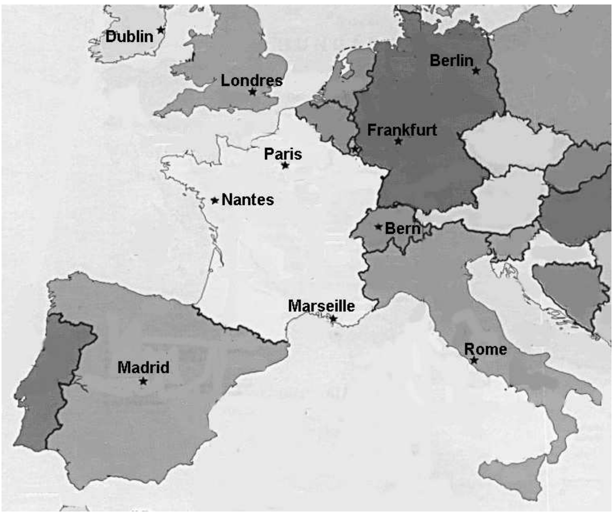

FIG. 1 - *Situation des 10 villes sur la carte d'Europe.*

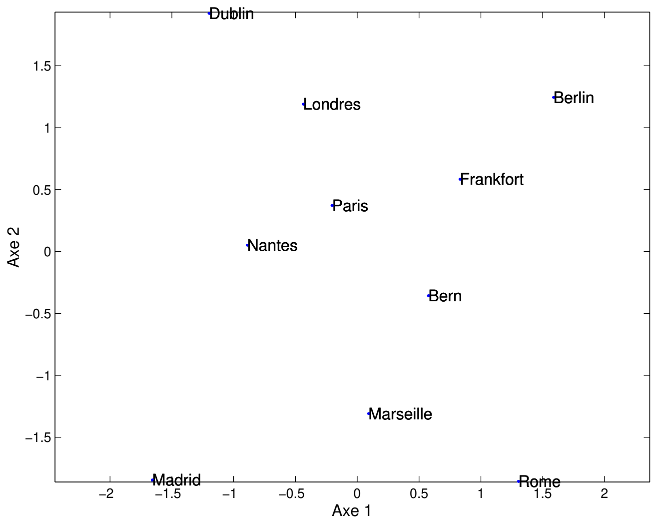

FIG. 2 - *Coupes de niveau 1 des régions calculées à partir des distances exactes.*

Le modèle de possibilité s'obtient en résolvant successivement c programmes linéaires pour k variant de 1 à c :

$$\max_{{}^{\alpha_k}\mathbf{r}} \sum_{i=1}^{n} {}^{\alpha_k}r_i \tag{28}$$

sous les contraintes :

$${}^{\alpha_k}r_i + {}^{\alpha_k}r_j \leq \min(d_{ij} - {}^{\alpha_k}\delta_{ij}^-, {}^{\alpha_k}\delta_{ij}^+ - d_{ij}) \quad \forall i, j \tag{29}$$

$${}^{\alpha_k}r_i \geq \begin{cases} 0 & \text{si } k = 1 \\ {}^{\alpha_{k-1}}r_i & \text{si } k > 1 \end{cases} \quad \forall i = 1, n. \tag{30}$$

Remarquons que ces problèmes n'admettent une solution qu'à la condition d'avoir :

$${}^1\widetilde{\delta}_{ij}^- \leq d_{ij} \leq {}^1\widetilde{\delta}_{ij}^+ \quad \forall i, j. \tag{31}$$

4 Exemple illustratif

Les méthodes proposées sont illustrées à l'aide d'un exemple classique utilisé en positionnemement multidimensionel : la reconstitution d'une carte à partir des distances entre un certain nombre de villes, représentées figure 1. On choisit de représenter les dissimilarités sous forme de nombres flous triangulaires symétriques $\widetilde{\delta}_{ij} = (\delta_{ij}^-, \delta_{ij}, \delta_{ij}^+)$ définis par un support $[\delta_{ij}^-, \delta_{ij}^+]$ et une valeur centrale δ_{ij}.

4.1 Cas des distances exactes

Les δ_{ij} sont ici les distances à vol d'oiseau relevées sur une carte, les bornes des supports étant calculés respectivement par $\delta_{ij}^+ = 0.75\delta_{ij}$ et $\delta_{ij}^- = 1.25\delta_{ij}$. Les figures 2, 3 et 4 représentent les coupes de niveau 1 et 0.5 ainsi que les supports des régions obtenues pour chacun des trois modèles. On peut noter sur la figure 2 que tous les modèles fournissent des rayons nuls, ce qui est cohérent avec le fait que les δ_{ij} sont des distances euclidiennes exactes. Sur les autres représentations, on distingue pour chaque objet trois cercles concentriques : le cercle intérieur correspond au modèle de nécessité, le cercle extérieur au modèle de nécessité, le dernier cercle à celui des moindres carrés. Les cercles les plus importants sont logiquement associés aux villes les plus excentrées, puisque l'imprécision artificiellement introduite augmente avec la distance. La figure 5 montre les distances reconstruites d_{ij}^+ et d_{ij}^- en fonction respectivement des dissimilarités δ_{ij}^+ et δ_{ij}^- pour l'α-coupe de niveau 1 et pour les trois modèles. Cette représentation, appelée diagramme de Shepard, démontre la qualité des distances reconstruites. Les figures 6 et 7 montrent ces mêmes diagrammes pour les supports. Ces représentations confirment l'emboîtement des modèles.

4.2 Cas de distances estimées

On se place maintenant dans le cas où les données d'entrée sont estimées par un sujet humain. L'evaluateur fournit, pour chaque dissimilarité, un intervalle à l'ex-

	Paris	Dublin	Londres	Frankfort	Berlin	Bern	Marseille	Nantes	Rome
Paris	0								
Dublin	[850;1050]	0							
Londres	[250;450]	[450;650]	0						
Frankfort	[500;700]	[1300;1700]	[600;800]	0					
Berlin	[900;1100]	[1700;2300]	[1000;1400]	[450;650]	0				
Bern	[500;700]	[1700;2300]	[850;1050]	[500;700]	[900;1300]	0			
Marseille	[800;1000]	[1800-2400]	[1100;1400]	[1000;1200]	[1600;2000]	[600;800]	0		
Nantes	[250;450]	[900;1100]	[500;700]	[850;1000]	[1500;2100]	[600;800]	[600;800]	0	
Rome	[1400;1800]	[2200;2800]	[1800;2100]	[1000;1200]	[1700;2300]	[800;1000]	[700;900]	[1200;1600]	0
Madrid	[1500;1900]	[1700;2300]	[1700;2000]	[1500;2500]	[2100;2800]	[1400;1800]	[900;1100]	[800;1200]	[1200;1800]

TAB. 1 - *Intervalles de distances estimés par l'évaluateur.*

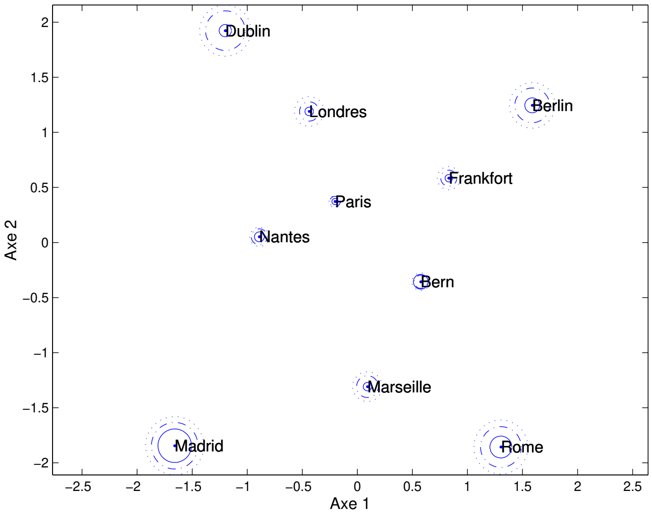

FIG. 3 - *Coupes de niveau 0.5 des régions calculées à partir des distances exactes; trait continu : modèle de nécessité; tirets : moindres carrés; pointillés : modèle de possibilité.*

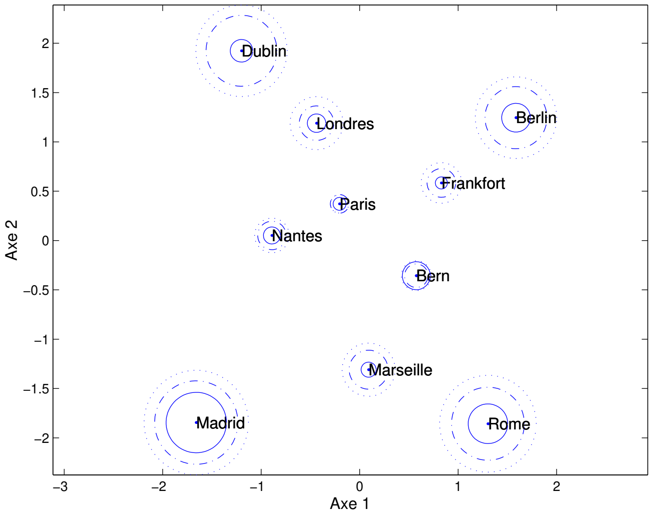

FIG. 4 - *Supports des régions calculés à partir des distances exactes; trait continu : modèle de nécessité; tirets : moindres carrés; pointillés : modèle de possibilité.*

térieur duquel il est certain que la dissimilarité ne se trouve pas. Cet intervalle est considéré comme le support de la dissimilarité floue, le milieu de l'intervalle comme la valeur modale. Les intervalles recueillis sont reportés dans le tableau 1. Les figures 8 et 9 montrent respectivement les α-coupes de niveau 1 et les supports des régions associés à chaque ville. Il faut noter que, dans ce cas, le modèle de nécessité n'existe pas, ce qui met en évidence une certaine incohérence dans l'estimation des dissimilarités. D'autre part, on remarque sur la figure 8 que les rayons du modèle de possibilité ne sont pas nuls, alors qu'ils l'étaient dans le cas des distances exactes. Le modèle de possibilité fournit donc une représentation permettant de juger de la cohérence des données d'entrée avec un modèle euclidien.

5 Discussion

Si l'on dispose de dissimilarités fournies sous forme d'intervalles, comment les traiter ? Si l'on utilise une méthode classique de positionnement (en traitant la matrice moyenne par exemple), on perd nécessairement de l'information. Il est certainement préférable d'utiliser des algorithmes tels que ceux présentés précédemment. Le modèle de nécessité n'a qu'un intérêt purement théorique à notre sens. On constate expérimentalement que les conditions de son existence sont rarement vérifiées. De plus, il pose des problèmes d'interprétation. Le modèle des possibilités est intéressant dans le sens où la représentation fournie est exacte et facilement interprétable : les intervalles de distance encadrent les intervalles fournis par les évaluateurs. Il reste cependant un modèle pessimiste puisqu'on impose que toutes les contraintes soient respectées. Certaines d'entre elles pourraient être relâchées au prix

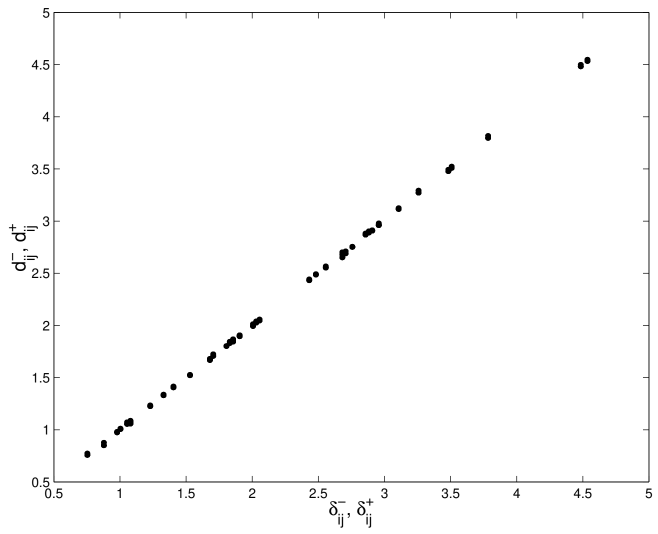

FIG. 5 - *Distances reconstruites en fonction des dissimilarités associées aux α-coupes de niveau 1.*

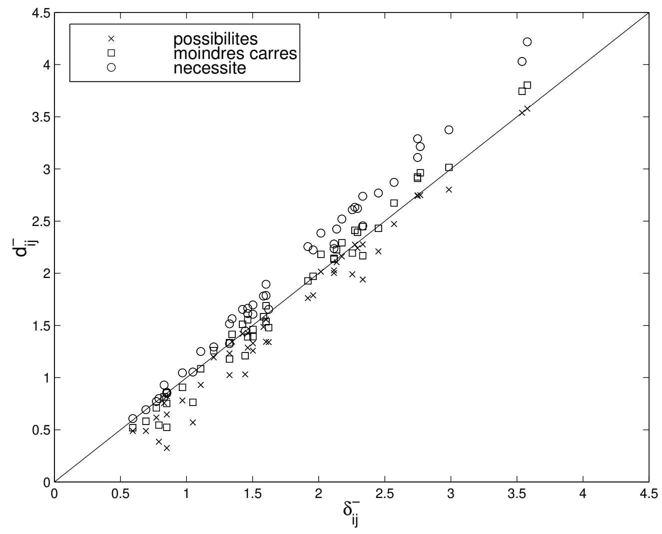

FIG. 6 - *Distances « basses » reconstruites en fonction des dissimilarités associées aux supports.*

de variables supplémentaires. Le modèle des moindres carrés fournit une représentation moyenne qui nous semble la plus satisfaisante. L'apport du flou aux modèles par intervalle peut être discutable. En effet les représentations obtenus sont plus difficilement interprétables, puisque l'on cherche à représenter plus d'informations. Un des intérêt de ces méthodes est cependant de permettre une étude de sensibilité relative aux bornes des intervalles. D'autre part, l'utlisation du flou apparait comme une démarche naturelle si les évaluations sont fournies sous forme linguistique. Le modèle proposé devra à l'usage démontrer son intérêt expérimental.

6 Conclusion

Les méthodes de positionnement multidimensionnel sont des outils de visualisation de données de dissimilarité. Dans cet article, nous avons proposé d'étendre les méthodes classiques pour pouvoir analyser des dissimilarités exprimées sous forme vague et modélisées par des intervalles ou des nombres flous. Dans ce cas, un objet n'est plus représenté comme un point dans l'espace mais comme une région nette ou floue. Après avoir choisi la forme de la région associée à chaque objet, nous avons décrit plusieurs méthodes pour déterminer ces régions. Nos algorithmes s'inspirent des modèles de possibilité et de nécessité introduits en régression floue. Une des applications visée est l'analyse de données d'évaluations subjectives recueillies au cours de tests sensoriels. Notre objectif est de fournir des représentations qui prennent en compte de façon explicite la variablité des réponses fournies par plusieurs évaluateurs.

Références

[1] I. Borg et P. Groenen. *Modern Multidimensional scaling,* Springer, New York 1997.

[2] T. F. Cox et M.A.A. Cox. *Multidimensional Scaling*, Chapman and Hall, London, 1994.

[3] P. Diamond Fuzzy least squares, *Information Science*, 46, 141-157, 1988.

[4] D. Dubois et H. Prade. *Possibility Theory: An approach to computerized processing of uncertainty*, Plenum Press, New-York, 1988.

[5] J. Kacprzik et M. Fedrizzi. *Fuzzy Regression Analysis,* Physica-Verlag, Heidelberg, 1992.

[6] A. Kaufmann et M. M. Gupta. *Introduction to fuzzy arithmetic. Theory and applications*, International Thomson Computer Press, London, 1991.

[7] M. Ming, M. Friedman et A. Kandel General fuzzy least squares, *Fuzzy sets and systems,* 88,107-118,1997.

[8] H. Tanaka, S. Uejima et K. Asai. Fuzzy linear regression models, *IEEE Trans. Systems, Man and Cybernetics,* 12,903-907,1982.

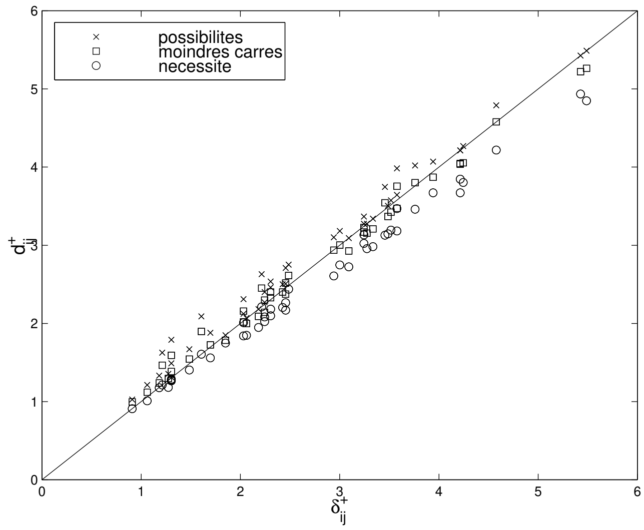

FIG. 7 - *Distances « hautes » reconstruites en fonction des dissimilarités associées aux supports.*

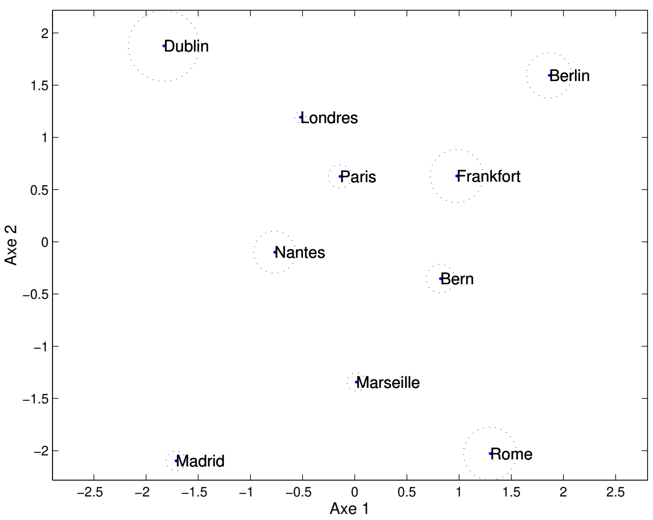

FIG. 8 - *Coupes de niveau 1 calculées à partir des distances estimées pour le modèle de possibilité.*

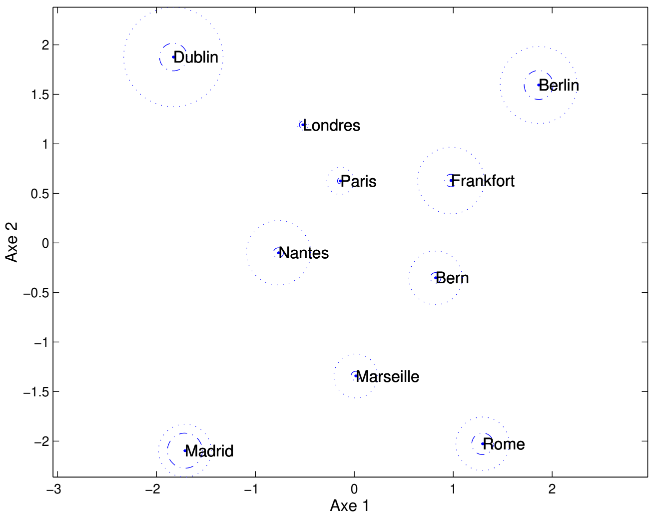

FIG. 9 - *Supports calculés à partir des distances estimées; trait pointillé : modèle de possibilité; tirets : modèle des moindres carrés.*