

# NOTES DE STATISTIQUE ET D'INFORMATIQUE

2003/1

LE POSITIONNEMENT MULTIDIMENSIONNEL :  
PRINCIPES ET APPLICATION

R. PALM

Faculté universitaire des  
Sciences agronomiques

Centre de Recherches  
agronomiques

**GEMBLOUX**

(Belgique)

# LE POSITIONNEMENT MULTIDIMENSIONNEL : PRINCIPES ET APPLICATION

R. PALM\*

## RÉSUMÉ

Cette note décrit quelques méthodes de positionnement multidimensionnel pour des tableaux à deux entrées avec un facteur et pour des tableaux à trois entrées avec deux facteurs. Les méthodes sont illustrées par un exemple traité par la procédure MDS du logiciel SAS.

## SUMMARY

This note describes some multidimensional scaling methods for two-way, one-mode tables and for three-way, two-mode tables. The methods are illustrated by an example processed by PROC MDS of SAS software.

## 1. INTRODUCTION

Le *positionnement multidimensionnel*<sup>1</sup> regroupe diverses méthodes permettant d'estimer les coordonnées d'un ensemble d'objets dans un espace de dimension spécifiée, à partir de la connaissance des proximités entre ces objets.

Ces méthodes ont d'abord été développées dans le domaine des sciences humaines, en psychologie notamment, mais sont utilisées également dans les sciences biologiques, parfois sous le nom de *méthodes d'ordination*<sup>2</sup>.

Il existe une grande variété de méthodes d'analyse regroupées sous le terme générique de positionnement multidimensionnel. Ces méthodes varient en fonction de la nature du tableau de départ qui est analysé et des algorithmes de calcul mis en oeuvre. Toutes ont en commun d'aboutir à une représentation graphique des objets dans un espace de dimension préalablement fixée.

---

\* Chargé de cours associé à la Faculté universitaire des Sciences agronomiques de Gembloux.

1. En anglais : *multidimensional scaling*.

2. En anglais : *ordination methods*.

Après un bref examen des types de tableaux de données pouvant faire l'objet du positionnement multidimensionnel (paragraphe 2), nous présentons les méthodes classiques, destinées à traiter des tableaux carrés symétriques contenant des mesures de proximité entre couples d'objets (paragraphe 3). Ensuite nous examinons comment on peut traiter simultanément plusieurs tableaux de ce type (paragraphe 4). Nous discutons alors les problèmes liés à l'interprétation des résultats (paragraphe 5), et nous donnons quelques informations complémentaires (paragraphe 6) avant de conclure (paragraphe 7).

Les différentes méthodes décrites sont illustrées par un exemple traité par la procédure MDS du logiciel SAS [SAS, 1989] et plusieurs figures reprennent directement des extraits des documents de sortie obtenus par ce logiciel. On notera que des différences peuvent exister entre les résultats fournis par les logiciels, les méthodes et les algorithmes de calculs utilisés pouvant varier d'un logiciel à l'autre.

Dans cette note, nous ne passons pas en revue toutes les méthodes de positionnement multidimensionnel proposées dans la littérature. Au contraire, nous nous limitons à des méthodes destinées à l'analyse de tableaux de structures bien particulières. Des informations plus détaillées relatives aux méthodes décrites dans cette note, mais aussi concernant d'autres méthodes, peuvent être trouvées dans les ouvrages consacrés au positionnement multidimensionnel et notamment dans BORG et GROENEN [1997], COX et COX [2001], EVERITT et RABE-HESKETH [1997] et KRUSKAL et WISH [1978].

Enfin, signalons encore que les méthodes examinées dans cette note décrivent les proximités entre les objets par des modèles spatiaux, puisqu'on cherche à représenter les objets par l'estimation de leurs coordonnées dans un espace de dimension fixée. Une autre approche pour l'analyse des données de proximité repose sur la représentation des objets sous la forme d'arbres ou de dendrogrammes. Cet aspect ne sera pas abordé ici. Des informations à ce sujet sont données dans EVERITT et RABE-HESKETH [1997].

## 2. TABLEAUX DES PROXIMITÉS

### 2.1. Similarité, dissimilarité, dominance

Nous avons signalé, dans l'introduction, que le point de départ des méthodes de positionnement multidimensionnel est le tableau des proximités entre objets. Ces proximités peuvent être des indices de dissimilarité, qui indiquent dans quelle mesure deux objets sont différents, ou, au contraire, des indices de similarité, qui indiquent dans quelle mesure deux objets sont semblables.

Les indices de similarité peuvent être transformés en indices de dissimilarité par la relation :

$$\text{dissimilarité} = c - \text{similarité},$$

$c$  étant une constante, en général égale à l'unité ou à 100, car les indices sont le plus souvent compris entre 0 et 1 ou entre 0 et 100.

Certains indices de dissimilarité portent le nom de distances. Une matrice de dissimilarités, dont les éléments sont notés  $\delta_{ij}$  ( $i = 1, \dots, n; j = 1, \dots, n$ ), avec  $\delta_{ii} = 0$  pour tout  $i$ , est appelée matrice de distances ou métrique, si, pour tout triplet  $h, i$ , et  $j$ , on a la relation suivante, connue sous le nom d'inégalité triangulaire :

$$\delta_{ij} + \delta_{ih} \geq \delta_{jh}.$$

De telles matrices jouissent de propriétés particulières. Elles sont notamment toujours symétriques et leurs éléments sont toujours non négatifs. Un exemple typique de matrice de distances est la matrice des distances euclidiennes entre couples d'objets définis par leurs coordonnées,  $y_{ik}$  et  $y_{jk}$ , dans un espace à  $p$  variables :

$$\delta_{ij} = \left[ \sum_{k=1}^p (y_{ik} - y_{jk})^2 \right]^{1/2}.$$

De manière plus générale, une matrice de distances est dite euclidienne si les  $n$  objets peuvent être représentés dans un espace euclidien, tel que la distance entre les objets  $i$  et  $j$  dans cet espace soit égale à  $\delta_{ij}$ .

De nombreux indices de proximité peuvent être calculés lorsque les objets sont définis par des variables quantitatives et/ou binaires. Une discussion et des listes de tels indices peuvent être trouvées, notamment dans CHANDON et PINSON [1981], EVERITT [1993], EVERITT et RABE-HESKETH [1997], LEGENDRE et LEGENDRE [1984]. Un exemple de calcul de proximités à partir de variables binaires sera donné au paragraphe 3.

Les proximités peuvent parfois aussi être observées directement ou résulter d'un jugement subjectif, sans que les objets soient décrits par des variables. C'est, notamment, le cas si on demande à un juge, au cours d'un test de dégustation, d'attribuer une note, par exemple de 0 à 5, qui quantifie la différence de goût entre paires de produits : on obtient alors une matrice symétrique reprenant les  $n(n-1)/2$  notes relatives à toutes les paires de produits, mais on ne dispose pas d'une matrice dont les lignes sont les produits et dont les colonnes sont des variables.

Une autre forme de matrice de départ pour le positionnement multidimensionnel est la matrice de dominance, comme par exemple la matrice de dimensions  $n \times n$  dont l'élément  $\delta_{ij}$  représente la proportion de fois que le produit  $i$  a été préféré au produit  $j$ , à l'issue d'un test de dégustation de  $n$  produits soumis à plusieurs juges.

D'une manière générale, nous considérons par la suite que les couples d'objets sont caractérisés par leur proximité  $\delta_{ij}$ , dont la valeur est d'autant plus grande que les objets sont différents. Cette proximité peut être une distance, une dissimilarité ou le résultat de la transformation d'une similarité en une dissimilarité.

## 2.2. Structure des tableaux de données

Des tableaux de structures différentes peuvent faire l'objet du positionnement multidimensionnel. Les situations les plus courantes sont reprises ci-dessous.

Le premier cas est celui d'un tableau symétrique de dimensions  $n \times n$ , dont l'élément  $\delta_{ij}$  est la proximité entre l'objet  $i$  et l'objet  $j$ . Dans un tel tableau, on retrouve les mêmes objets en lignes et en colonnes. Le point de départ est donc identique au point de départ d'une classification numérique.

Le positionnement multidimensionnel des objets à partir d'un tel tableau sera examiné au paragraphe 3. Nous illustrerons la méthode en considérant la matrice de proximités relatives à 9 stations situées sur la Meuse et la Sambre, les proximités étant définies à partir de la présence ou l'absence de 12 mousses aquatiques lors de relevés réalisés à un moment donné [VANDERPOORTEN, 2000].

Une situation légèrement différente correspond au cas d'un tableau carré, de dimensions  $n \times n$ , qui serait non symétrique. On retrouve les mêmes objets en lignes et en colonnes, mais la proximité  $\delta_{ij}$  n'est pas nécessairement égale à  $\delta_{ji}$ . Un tel tableau pourrait résulter de la comparaison de produits évoquée au paragraphe précédent;  $\delta_{ij}$  correspondrait à la proximité des produits  $i$  et  $j$  lorsque  $i$  est goûté en premier lieu et  $\delta_{ji}$  correspondrait à la proximité des mêmes produits lorsque  $j$  est goûté en premier lieu.

Les tableaux évoqués ci-dessus sont qualifiés de *tableaux à deux entrées et un facteur*<sup>3</sup>, qu'ils soient symétriques ou non.

Lorsque les proximités entre objets sont évaluées par différentes personnes ou à différentes dates par exemple, on dispose d'un tableau de proximités à *trois dimensions avec deux facteurs*<sup>4</sup>. Chaque tranche d'un tel tableau correspond aux données fournies par une personne ou obtenues à une date, et on trouve les mêmes objets en lignes et en colonnes. Un ensemble de matrices carrées contenant les corrélations entre  $n$  variables à différents moments est également un exemple classique de données à trois dimensions et deux facteurs.

Nous examinerons l'analyse d'un tel tableau au paragraphe 4. Nous reprendrons le tableau des proximités des 9 stations évoqué ci-dessus et nous y ajouterons un tableau équivalent construit à partir d'observations réalisées quinze ans plus tard [VANDERPOORTEN, 2000].

Enfin, on rencontre également des tableaux à deux entrées dans lesquels les lignes et les colonnes correspondent à des ensembles d'objets différents. De tels tableaux, à *deux dimensions et à deux facteurs*<sup>5</sup>, sont le plus souvent de forme rectangulaire. Un tableau donnant les valeurs des taux d'apparition de différents symptômes (en colonnes) en fonction de la maladie (en lignes) est un exemple de ce type de données. Nous n'envisagerons pas ce cas dans cette note. Des informations à ce sujet sont données par RABE-HESKETH [1997], notamment.

---

3. En anglais : *two-way, one-mode*.

4. En anglais : *three-way, two-mode*.

5. En anglais : *two-way, two-mode*.

### 3. TABLEAU À DEUX ENTRÉES ET UN FACTEUR

#### 3.1. Positionnement métrique

Le positionnement multidimensionnel a d'abord été développé pour la représentation, dans un espace de dimension  $q$  fixée, d'un tableau de données à deux entrées avec un facteur, c'est-à-dire d'un tableau, de dimensions  $n \times n$ , contenant les proximités entre les paires d'objets.

Positionner un objet  $i$  ( $i = 1, \dots, n$ ) dans un espace de dimension  $q$  revient à déterminer ses coordonnées  $x_{ik}$  ( $k = 1, \dots, q$ ) sur les  $q$  axes qui définissent cet espace. Le but est de représenter les  $n$  objets dans cet espace de manière à ce que les distances entre les points dans l'espace de dimension  $q$  correspondent aussi bien que possible aux proximités observées entre les objets, données dans la matrice initiale.

Soit  $\delta_{ij}$  la proximité observée entre l'objet  $i$  et l'objet  $j$  et soit  $d_{ij}$  la distance entre ces mêmes objets dans l'espace de représentation de dimension  $q$ . Si cette dernière distance est euclidienne, comme dans la plupart des applications, on a :

$$d_{ij} = \left[ \sum_{k=1}^q (x_{ik} - x_{jk})^2 \right]^{1/2},$$

$x_{ik}$  étant la coordonnée de l'objet  $i$  sur l'axe  $k$ .

Une solution relativement naturelle pour positionner les objets consiste à déterminer les coordonnées  $x_{ik}$  des  $n$  objets de manière à minimiser la somme des carrés des écarts :

$$\sum_{i < j} (d_{ij} - \delta_{ij})^2,$$

la somme étant étendue aux  $n(n-1)/2$  paires d'objets.

Une telle approche suppose implicitement que la relation entre  $\delta_{ij}$  et  $d_{ij}$  est du type :

$$\delta_{ij} = d_{ij} + e_{ij},$$

$e_{ij}$  représentant l'erreur liée au fait que les proximités ne correspondent pas exactement à une configuration dans un espace de dimension  $q$ .

Ce modèle est particulièrement rigide et une solution un peu plus compliquée, mais plus souple, revient à considérer que la relation entre  $\delta_{ij}$  et  $d_{ij}$  est du type :

$$\delta_{ij} = \beta d_{ij} + e_{ij}$$

ou 
$$\delta_{ij} = \alpha + \beta d_{ij} + e_{ij},$$

ou, de manière plus générale encore :

$$\delta_{ij} = f(d_{ij}) + e_{ij},$$

$f(d_{ij})$  étant une fonction quelconque choisie par l'utilisateur.

Les quantités :

$$\hat{d}_{ij} = f(d_{ij})$$

sont les *distances transformées* ou *ajustées*<sup>6</sup>.

L'ajustement du modèle revient à trouver, d'une part, les valeurs optimales des coordonnées des objets dans l'espace de dimension  $q$ , qui déterminent les valeurs des distances  $d_{ij}$  et, d'autre part, les paramètres de la fonction  $f(d_{ij})$ , par exemple la valeur  $\beta$  ou les valeurs  $\alpha$  et  $\beta$  dans le cas des modèles ci-dessus, qui déterminent les distances ajustées  $\hat{d}_{ij}$ .

Deux approches sont utilisées pour l'ajustement du modèle. Certains logiciels utilisent le critère du maximum de vraisemblance et d'autres, comme le logiciel SAS, utilisent un critère des moindres carrés non linéaires. Dans ce cas, par analogie avec les modèles de régression, on minimise la somme des carrés des erreurs :

$$\sum_{i < j} [\delta_{ij} - f(d_{ij})]^2 = \sum_{i < j} e_{ij}^2,$$

ou, ce qui revient au même, le *critère d'ajustement*<sup>7</sup> suivant :

$$\sqrt{\frac{\sum_{i < j} e_{ij}^2}{\sum_{i < j} \delta_{ij}^2}}.$$

Nous verrons, au paragraphe 5.1, qu'il existe des variantes à ce critère d'ajustement.

La présentation du modèle ci-dessus correspond au modèle de régression proposé dans la procédure MDS du logiciel SAS. D'autres présentations sont possibles. Ainsi, DE SOETE et CARROL [1998], EVERITT et RABE-HESKETH [1997] et KRUSKAL et WISH [1978] écrivent le modèle de la manière suivante :

$$d_{ij} = f(\delta_{ij}) + e_{ij},$$

alors que pour BERG et GROENEN [1997] et YOUNG [1985], le modèle est :

$$f(\delta_{ij}) = d_{ij} + e_{ij}.$$

Les résidus sont alors les écarts entre les distances  $d_{ij}$  et les proximités transformées  $f(\delta_{ij})$ , comme c'est d'ailleurs le cas pour le logiciel SAS pour le positionnement non métrique (paragraphe 3.2).

A titre d'illustration, nous reprenons les données de VANDERPOORTEN [2000] relatives à la présence ou à l'absence de 12 bryophytes aquatiques dans 9 stations situées sur la Meuse et sur un de ses affluents, la Sambre. Ces données sont reprises en annexe. Les stations de la Meuse sont, de l'amont vers l'aval : Anseremme, La Plante, Andenne, Ivoz-Ramet, l'île de Monsin et Visé. Les deux premières stations se situent en amont de la confluence de la Sambre et de la Meuse. Les stations de la Sambre sont Charleroi, Floreffe et Namur. Par la suite,

6. En anglais : *fitted distances, target distances, optimally scaled distances, disparities*.

7. En anglais : *badness of fit criterion, stress, stress-1*.

Iteration	Type	Badness-of-Fit Criterion	Change in Criterion	Convergence Measure
0	Initial	0.0945	.	0.8132
1	Lev-Mar	0.0501	0.0444	0.5901
2	Gau-New	0.0405	0.009553	0.0477
3	Gau-New	0.0405	0.0000452	0.003653

Convergence criterion is satisfied.

Figure 1. Positionnement métrique : ajustement du modèle.

dans les différentes figures, ces stations seront représentées par des abréviations du nom, précédées de la lettre M ou S, selon qu'il s'agit d'une station de la Meuse ou de la Sambre. Les observations ont été réalisées en 1972 et 1997. Dans ce paragraphe, nous ne prenons en considération que les données de 1972; les autres observations seront utilisées au paragraphe 4.

A partir des données de présence ou absence des 12 bryophytes, nous avons calculé la matrice des coefficients de similitude de SOKAL et MICHENER, qui est la proportion de co-présences et co-absences. Ainsi, par exemple, pour les stations d'Ivoz-Ramet et de Anseremme, 4 plantes sont absentes des deux relevés et 3 plantes sont présentes dans les deux relevés, les 5 autres plantes ne sont présentes que dans un des deux relevés. Le coefficient de similitude est par conséquent égal à  $7/12$ , soit 0,58. Le choix du coefficient de SOKAL et MICHENER se justifie dans cet exemple, car on peut estimer que la co-absence d'une plante doit être prise en considération, au même titre que la co-présence pour la mesure de la ressemblance des stations.

Les coefficients de similitude ont été transformés en coefficients de dissimilarité, en prenant le complément à l'unité et en exprimant le résultat en pour-cent. Pour Anseremme et Ivoz-Ramet, la proximité est, par exemple, égale à 42. Nous ne reproduisons pas ici toutes les proximités entre couples de stations. Celles-ci figureront dans les documents de sorties examinés par la suite.

La matrice des proximités, de dimensions  $9 \times 9$  a été analysée avec le logiciel SAS, en utilisant le modèle suivant :

$$\delta_{ij} = \alpha + \beta d_{ij} + e_{ij}.$$

La figure 1 donne les informations relatives à la procédure d'ajustement. Les différentes lignes correspondent aux itérations successives. Pour chacune de celles-ci, la figure précise la nature de l'itération réalisée, la valeur du critère de qualité d'ajustement, la variation de ce critère par rapport à l'itération précédente et une

Obs	_TYPE_	Station	Dim1	Dim2
1	CRITERION		0.0405	.
2	CONFIG	M_ ANS	2.0774	-0.63050
3	CONFIG	M_ PLANT	1.6542	-0.32087
4	CONFIG	M_ AND	1.6242	0.65101
5	CONFIG	M_ IVOZ	-0.2603	0.29204
6	CONFIG	M_ MONS	-0.4428	0.94507
7	CONFIG	M_ VISE	-0.7065	-0.00988
8	CONFIG	S_ NAM	-1.1504	-0.27522
9	CONFIG	S_ CHAR	-1.6292	-0.41918
10	CONFIG	S_ FLOR	-1.1665	-0.23246
11	INTERCEPT		-1.1405	.
12	SLOPE		17.9183	.

Figure 2. Positionnement métrique : coordonnées des stations dans le plan.

mesure de convergence du processus d'itération. Des informations plus détaillées à ce sujet sont données dans SAS [1989]. Le critère de qualité d'ajustement défini ci-dessus est égal à 0,04 pour la dernière itération. La faible valeur de ce coefficient montre que la représentation géométrique dans l'espace de dimension 2, qui est la dimension par défaut, respecte très bien les proximités entre les stations. Nous discuterons, au paragraphe 5.2, le choix de cette dimension.

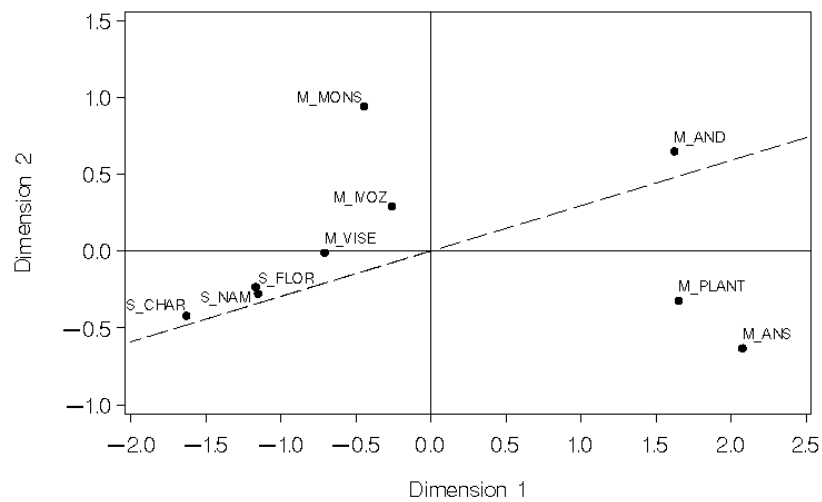


Figure 3. Positionnement métrique : représentation des stations dans le plan.

La figure 2 reprend, notamment, les coordonnées des stations dans l'espace de dimension 2 ainsi que les valeurs estimées  $\hat{\alpha}$  et  $\hat{\beta}$  du modèle. Ces coordonnées permettent de réaliser la représentation graphique des stations (figure 3).

Pour l'axe 1, les stations de la Sambre se situent à gauche de la figure, alors que les trois stations de la Meuse les plus en amont se situent à droite du graphique, les stations de la Meuse les plus en aval occupant une position intermédiaire. Cet axe est lié à la richesse en bryophytes. En effet, la corrélation entre les coordonnées des stations sur cet axe et le nombre d'espèces présentes est de 0,987. Dans les stations de la Sambre, on observe au maximum une espèce de bryophyte; dans les stations de l'aval de la Meuse, on observe entre 2 et 4 espèces et dans les stations de l'amont le nombre d'espèces est de 7 ou 8. Ce nombre d'espèces est lui-même à mettre en relation avec le degré de pollution des eaux à l'époque des observations. Les eaux de la Sambre, subissant une forte pollution liée aux industries de la région de Charleroi, étaient nettement plus polluées que les eaux de la Meuse. Le débit important de la Meuse diluait cette pollution, ce qui explique que la station d'Andenne, bien que située en aval de la confluence se trouve, sur la figure 3, à proximité des stations d'Anseremme et de La Plante. Le bassin industriel liégeois était, par contre, une nouvelle source importante de pollution, responsable de l'appauvrissement de la bryoflore aquatique [VANDERPOORTERN, 2000].

Nous verrons, au paragraphe 5.3, que la direction correspondant à la richesse en bryophytes peut être déterminée de façon plus précise et correspond à la droite représentée en traits discontinus dans la figure 3.

La figure 4 donne, pour chaque couple de stations, les proximités  $\delta_{ij}$ , les distances euclidiennes  $d_{ij}$  entre stations dans l'espace de dimension 2, la transformation linéaire de ces distances  $f(d_{ij})$  et les résidus  $e_{ij}$ . Les proximités entre couples, qui sont donc les dissimilarités dont nous avons explicité le calcul ci-dessus, permettraient de reconstituer la matrice, de dimensions  $9 \times 9$ , qui sert de point de départ à la procédure MDS du logiciel SAS.

Nous allons examiner plus en détail le couple de stations Ivoz-Ramet et Anseremme, pris à titre d'exemple. On peut vérifier que la distance entre ces deux stations dans l'espace de dimension 2 (figure 3) est bien égale à :

$$\sqrt{[2,0774 - (-0,2603)]^2 + [-0,63050 - 0,29204]^2} = 2,5131,$$

les nombres intervenant dans l'expression ci-dessus étant les coordonnées des deux stations reprises dans la figure 2. La distance transformée s'obtient par la relation suivante :

$$-1,1405 + (17,9183)(2,5131) = 43,89,$$

et le résidu est égal à :

$$42 - 43,89 = -1,89,$$

la proximité des deux stations étant égale à 42.

La somme des carrés des proximités reprises dans la colonne DATA de la figure 4 et la somme des carrés des résidus sont respectivement égales à 49.425

Obs	_ROW_	_COL_	DATA	DISTANCE	TRANDIST	RESIDUAL
1	S_FLOR	S_NAM	0	0.04569	-0.3218	0.32181
2	M_PLANT	M_ANS	8	0.52438	8.2554	-0.25537
3	M_MONS	M_IVOZ	8	0.67807	11.0094	-3.00935
4	M_VISE	M_IVOZ	8	0.53874	8.5127	-0.51270
5	S_NAM	M_VISE	8	0.51720	8.1268	-0.12681
6	S_CHAR	S_NAM	8	0.49999	7.8184	0.18162
7	S_FLOR	M_VISE	8	0.51108	8.0171	-0.01705
8	S_FLOR	S_CHAR	8	0.49896	7.8000	0.20005
9	M_AND	M_PLANT	17	0.97235	16.2822	0.71777
10	M_VISE	M_MONS	17	0.99068	16.6107	0.38930
11	S_NAM	M_IVOZ	17	1.05552	17.7726	-0.77259
12	S_CHAR	M_VISE	17	1.00947	16.9474	0.05261
13	S_FLOR	M_IVOZ	17	1.04708	17.6214	-0.62139
14	M_AND	M_ANS	25	1.35928	23.2153	1.78467
15	S_NAM	M_MONS	25	1.41060	24.1349	0.86508
16	S_CHAR	M_IVOZ	25	1.54268	26.5017	-1.50167
17	S_FLOR	M_MONS	25	1.38214	23.6250	1.37497
18	M_IVOZ	M_PLANT	33	2.01019	34.8785	-1.87853
19	M_IVOZ	M_AND	33	1.91840	33.2338	-0.23378
20	S_CHAR	M_MONS	33	1.80797	31.2551	1.74493
21	M_IVOZ	M_ANS	42	2.51312	43.8903	-1.89028
22	M_MONS	M_PLANT	42	2.44952	42.7506	-0.75056
23	M_MONS	M_AND	42	2.08788	36.2706	5.72940
24	M_VISE	M_PLANT	42	2.38105	41.5238	0.47620
25	M_VISE	M_AND	42	2.42259	42.2680	-0.26803
26	M_MONS	M_ANS	50	2.97220	52.1162	-2.11616
27	M_VISE	M_ANS	50	2.85220	49.9659	0.03407
28	S_NAM	M_PLANT	50	2.80498	49.1199	0.88010
29	S_NAM	M_AND	50	2.92517	51.2734	-1.27336
30	S_FLOR	M_PLANT	50	2.82211	49.4267	0.57328
31	S_FLOR	M_AND	50	2.92726	51.3109	-1.31094
32	S_NAM	M_ANS	58	3.24731	57.0456	0.95439
33	S_CHAR	M_PLANT	58	3.28490	57.7191	0.28090
34	S_CHAR	M_AND	58	3.42496	60.2288	-2.22881
35	S_FLOR	M_ANS	58	3.26825	57.4209	0.57909
36	S_CHAR	M_ANS	67	3.71265	65.3837	1.61635

Figure 4. Positionnement métrique : proximités  $\delta_{ij}$  (DATA), distances  $d_{ij}$  (DISTANCE), distances transformées  $f(d_{ij})$  (TRANDIST) et résidus  $e_{ij}$  (RESIDUAL).

Tableau 1. Régression monotone croissante de  $y$  sur  $x$  : valeurs ajustées obtenues aux étapes successives.

$i$	$x_i$	$y_i$	Etapas				
			$\hat{y}_i^{(0)}$	$\hat{y}_i^{(1)}$	$\hat{y}_i^{(2)}$	$\hat{y}_i^{(3)}$	$\hat{y}_i^{(4)}$
1	10	0,66	0,66	0,66	0,66	0,66	0,66
2	16	0,76	0,76	0,74	0,74	0,74	0,74
3	18	0,72	0,72	0,74	0,74	0,74	0,74
4	19	0,82	0,82	0,82	0,82	0,82	0,82
5	27	1,25	1,25	1,25	1,25	1,18	1,16
6	29	1,35	1,35	1,35	1,15	1,18	1,16
7	32	0,95	0,95	0,95	1,15	1,18	1,16
8	35	1,10	1,10	1,10	1,10	1,10	1,16
9	37	1,42	1,42	1,42	1,42	1,42	1,42
10	43	1,51	1,51	1,51	1,51	1,51	1,51

et 81,03. A partir de ces deux valeurs, on retrouve la valeur du critère de qualité de l'ajustement donnée dans la figure 1 :

$$\sqrt{81,03/49.425} = 0,0405.$$

### 3.2. Positionnement non métrique

Alors que dans le positionnement métrique on considère un modèle du type :

$$\delta_{ij} = f(d_{ij}) + e_{ij},$$

la nature de la fonction  $f(d_{ij})$  étant déterminée *a priori*, le positionnement non métrique fait appel à la notion de régression monotone croissante dont nous allons d'abord présenter le principe sur base d'un exemple artificiel.

Le tableau 1 donne 10 couples d'observations  $(x_i, y_i)$ , classés par ordre croissant des valeurs  $x_i$ . Ces couples sont les données de départ et l'objectif est de déterminer les valeurs  $\hat{y}_i$  qui respectent les contraintes suivantes :

- a) les valeurs  $\hat{y}_i$  doivent être non décroissantes;
- b) la somme des carrés des écarts entre les valeurs  $y_i$  et  $\hat{y}_i$  doit être minimum.

La détermination des  $\hat{y}_i$  se fait par étapes successives, en définissant des valeurs ajustées temporaires qui peuvent faire l'objet de modifications ultérieures de manière à satisfaire les contraintes. Pour l'exemple, ces valeurs sont reprises dans le tableau 1.

Une première série de valeurs ajustées  $\hat{y}_i^{(0)}$  est obtenue en posant  $\hat{y}_i^{(0)} = y_i$  pour toutes les lignes du tableau. On recherche ensuite la première valeur ajustée

qui ne vérifie pas la condition de non-décroissance. Pour l'exemple, il s'agit de la valeur correspondant à la troisième ligne. On procède alors à une modification des valeurs ajustées, de manière à assurer la non-décroissance :

$$\hat{y}_2^{(1)} \leq \hat{y}_3^{(1)} ,$$

tout en rendant minimum l'expression :

$$\left( y_2 - \hat{y}_2^{(1)} \right)^2 + \left( y_3 - \hat{y}_3^{(1)} \right)^2 .$$

La solution est :

$$\hat{y}_2^{(1)} = \hat{y}_3^{(1)} = (y_2 + y_3)/2 = 0,74 .$$

Après cette modification, les valeurs ajustées  $\hat{y}_i^{(1)}$  sont non décroissantes jusqu'à la ligne 6. On peut donc définir :

$$\hat{y}_6^{(2)} = \hat{y}_7^{(2)} = (y_6 + y_7)/2 = 1,15 .$$

Cette correction a cependant comme effet de ne plus assurer la non-décroissance pour les lignes 5 et 6. Une révision de la solution conduit aux résultats suivants :

$$\hat{y}_5^{(3)} = \hat{y}_6^{(3)} = \hat{y}_7^{(3)} = (y_5 + y_6 + y_7)/3 = 1,18 .$$

Une décroissance s'observe encore à la huitième ligne. L'élimination de celle-ci conduit aux valeurs ajustées suivantes :

$$\hat{y}_5^{(4)} = \hat{y}_6^{(4)} = \hat{y}_7^{(4)} = \hat{y}_8^{(4)} = (y_5 + y_6 + y_7 + y_8)/4 = 1,16 .$$

A la suite de cette dernière modification, on obtient les valeurs ajustées qui sont bien toutes non décroissantes. Les valeurs  $\hat{y}_i^{(4)}$  sont donc les valeurs ajustées finales :

$$\hat{y}_i = \hat{y}_i^{(4)} .$$

Ces valeurs constituent une transformation monotone croissante des valeurs  $x$ . Cela signifie que si on porte sur un graphique les  $\hat{y}_i$  en fonction des  $x_i$  et qu'on relie les points correspondant aux  $x_i$  successifs, les segments de droite ne sont jamais décroissants.

Les figures 5 à 7 donnent les résultats du positionnement non métrique obtenus par le logiciel SAS pour les stations de la Meuse et de la Sambre. Le modèle retenu par ce logiciel est le suivant :

$$t(\delta_{ij}) = d_{ij} + e_{ij} ,$$

le critère d'ajustement à minimiser étant :

$$\sum_{i < j} [t(\delta_{ij}) - d_{ij}]^2 / \sum_{i < j} [t(\delta_{ij})]^2 .$$

Iteration	Type	Badness-of-Fit Criterion	Change in Criterion	Convergence Measures	
				Monotone	Gradient
0	Initial	0.0947	.	.	.
1	Monotone	0.0906	0.004172	0.0288	0.8217
2	Lev-Mar	0.0445	0.0461	.	.
3	Monotone	0.0399	0.004594	0.0148	0.4249
4	Gau-New	0.0397	0.000204	.	.
5	Monotone	0.0380	0.001615	0.0113	0.3922
6	Gau-New	0.0368	0.001208	.	.
7	Monotone	0.0363	0.000514	0.006255	0.2983
8	Gau-New	0.0347	0.001659	.	0.0114
9	Gau-New	0.0347	2.2331E-6	.	0.001564

Convergence criteria are satisfied.

Figure 5. Positionnement non métrique : ajustement du modèle.

Obs	_TYPE_	Station	Dim1	Dim2
1	CRITERION		0.03466	.
2	CONFIG	M_ANS	2.05162	-0.65766
3	CONFIG	M_PLANT	1.67003	-0.28264
4	CONFIG	M_AND	1.64618	0.63409
5	CONFIG	M_IVOZ	-0.26416	0.26143
6	CONFIG	M_MONS	-0.44926	0.94083
7	CONFIG	M_VISE	-0.70501	-0.00142
8	CONFIG	S_NAM	-1.21089	-0.15555
9	CONFIG	S_CHAR	-1.52741	-0.58406
10	CONFIG	S_FLOR	-1.21110	-0.15501

Figure 6. Positionnement non métrique : coordonnées des stations dans le plan.

Dans ces expressions, la notation  $t(\delta_{ij})$  représente la transformation monotone des  $\delta_{ij}$ . En relation avec la présentation de la régression monotone ci-dessus, on voit que les distances  $d_{ij}$  entre les stations dans le plan jouent le rôle des  $y_i$ , que les proximités  $\delta_{ij}$  jouent le rôle des  $x_i$  et, enfin, que les  $t(\delta_{ij})$  jouent le rôle des  $\hat{y}_i$ .

Par rapport au positionnement métrique, on constate que le critère de qualité de l'ajustement est légèrement plus faible pour le positionnement non métrique

que pour le positionnement métrique (0,035 contre 0,040 : voir figures 1 et 5). Cela s'explique par le caractère moins rigide de la transformation monotone, par rapport à la transformation linéaire. On notera cependant que les deux valeurs ne sont pas tout à fait comparables, les formules de calcul étant différentes : dans la formule ci-dessus, ce sont les proximités qui sont transformées, alors que dans la formule du paragraphe 3.1, la transformation porte sur les distances.

La comparaison des coordonnées des stations dans les deux configurations (figures 2 et 6) montre que la position des stations dans l'espace de dimension 2 est pratiquement identique pour les deux méthodes, la divergence la plus forte s'observant pour la station de Charleroi.

L'exemple traité est un cas particulier à deux égards. Tout d'abord, on note la présence d'un grand nombre de proximités identiques : les 36 couples de stations ne présentent que 9 valeurs différentes de proximité. Cela est évidemment lié à la nature des données de départ. Pour des données de présence/absence relatives à 12 espèces de bryophytes, les proximités ne peuvent être que des multiples de 100/12, soit après arrondis, 0, 8, 17, 25, 33, etc. La seconde particularité de l'exemple est que, pour une valeur donnée de la proximité, les distances sont inférieures à toutes les distances relatives à des proximités plus grandes et elles sont supérieures à toutes les distances relatives à des proximités plus petites. Par exemple, pour  $\delta_{ij} = 25$ , les distances sont comprises entre 1,33464 et 1,52009 et on a : pour tous les couples de stations :

$$d_{ij} < 1,33464 \quad \text{lorsque} \quad \delta_{ij} < 25$$

et 
$$d_{ij} > 1,52009 \quad \text{lorsque} \quad \delta_{ij} > 25.$$

Comme on a imposé que :

$$t(\delta_{ij}) = t(\delta_{kh}) \quad \text{si} \quad \delta_{ij} = \delta_{kh},$$

il en résulte que les  $t(\delta_{ij})$  changent de valeur chaque fois que les  $\delta_{ij}$  changent de valeur.

L'approche utilisée dans cet exemple en cas d'égalité de proximités est connue sous le nom de *seconde approche*<sup>8</sup> des *ex-aequo*. Elle est plus restrictive que la *première approche*<sup>9</sup> qui n'impose pas l'égalité des valeurs transformées pour des proximités identiques. Appliquée à l'exemple ci-dessus, cette première approche conduit à une valeur du critère d'ajustement pratiquement égale à zéro.

On remarque également que, dans la figure 7, la valeur  $t(\delta_{ij})$  correspondant à une valeur  $\delta_{ij}$  n'est pas exactement la moyenne arithmétique des distances de tous les couples caractérisés par la même proximité. Ainsi, par exemple, trois couples de stations ont une proximité égale à 33. La moyenne des distances pour ces trois couples est :

$$(2,00926 + 1,94635 + 1,86754)/3 = 1,94105,$$

---

8. En anglais : *secondary approach*.

9. En anglais : *primary approach*.

Obs	_ROW_	_COL_	DATA	TRANDATA	DISTANCE	RESIDUAL
1	S_FLOR	S_NAM	0	0.00085	0.00058	0.00027
2	M_PLANT	M_ANS	8	0.55920	0.53502	0.02418
3	M_MONS	M_IVOZ	8	0.55920	0.70416	-0.14496
4	M_VISE	M_IVOZ	8	0.55920	0.51327	0.04593
5	S_NAM	M_VISE	8	0.55920	0.52883	0.03037
6	S_CHAR	S_NAM	8	0.55920	0.53273	0.02646
7	S_FLOR	M_VISE	8	0.55920	0.52888	0.03032
8	S_FLOR	S_CHAR	8	0.55920	0.53304	0.02616
9	M_AND	M_PLANT	17	0.99648	0.91704	0.07944
10	M_VISE	M_MONS	17	0.99648	0.97634	0.02013
11	S_NAM	M_IVOZ	17	0.99648	1.03449	-0.03801
12	S_CHAR	M_VISE	17	0.99648	1.00787	-0.01139
13	S_FLOR	M_IVOZ	17	0.99648	1.03447	-0.03799
14	M_AND	M_ANS	25	1.38844	1.35388	0.03457
15	S_NAM	M_MONS	25	1.38844	1.33496	0.05348
16	S_CHAR	M_IVOZ	25	1.38844	1.52009	-0.13165
17	S_FLOR	M_MONS	25	1.38844	1.33464	0.05380
18	M_IVOZ	M_PLANT	33	1.94130	2.00926	-0.06795
19	M_IVOZ	M_AND	33	1.94130	1.94635	-0.00504
20	S_CHAR	M_MONS	33	1.94130	1.86754	0.07377
21	M_IVOZ	M_ANS	42	2.37571	2.49149	-0.11579
22	M_MONS	M_PLANT	42	2.37571	2.44709	-0.07138
23	M_MONS	M_AND	42	2.37571	2.11777	0.25794
24	M_VISE	M_PLANT	42	2.37571	2.39163	-0.01593
25	M_VISE	M_AND	42	2.37571	2.43556	-0.05986
26	M_MONS	M_ANS	50	2.91361	2.96808	-0.05447
27	M_VISE	M_ANS	50	2.91361	2.83366	0.07995
28	S_NAM	M_PLANT	50	2.91361	2.88372	0.02989
29	S_NAM	M_AND	50	2.91361	2.96418	-0.05057
30	S_FLOR	M_PLANT	50	2.91361	2.88395	0.02965
31	S_FLOR	M_AND	50	2.91361	2.96424	-0.05063
32	S_NAM	M_ANS	58	3.31089	3.30091	0.00998
33	S_CHAR	M_PLANT	58	3.31089	3.21162	0.09927
34	S_CHAR	M_AND	58	3.31089	3.39935	-0.08846
35	S_FLOR	M_ANS	58	3.31089	3.30120	0.00969
36	S_CHAR	M_ANS	67	3.60849	3.57979	0.02870

Figure 7. Positionnement non métrique: proximités  $\delta_{ij}$  (DATA), proximités transformées  $t(\delta_{ij})$  (TRANDATA), distances  $d_{ij}$  (DISTANCE) et résidus  $e_{ij}$  (RESIDUAL).

soit une valeur très légèrement différente de la distance transformée, qui est égale à 1,94130. Cette discordance est liée au caractère itératif du procédé d'ajustement qui alterne le calcul de la régression monotone et le calcul par moindres carrés non linéaires d'une nouvelle configuration, comme on peut le vérifier dans la figure 5. En augmentant le nombre d'itérations, on constaterait d'ailleurs que les  $t(\delta_{ij})$  se rapprochent des moyennes des distances.

#### 4. TABLEAU À TROIS ENTRÉES AVEC DEUX FACTEURS

On considère maintenant le cas où on dispose de  $m$  tableaux de proximités, chacun de ceux-ci étant un tableau carré symétrique, de dimensions  $n \times n$ . Comme signalé au paragraphe 2.2, une telle situation se présente, par exemple, lorsque  $n$  objets sont évalués par  $m$  personnes différentes, un tableau correspondant aux proximités des  $n$  objets évalués par une personne. Une autre situation classique résulte de l'évaluation de la proximité entre une série d'objets à des dates différentes. Ainsi, pour illustrer la méthode d'analyse, nous reprendrons les données relatives aux proximités observées entre les neuf stations sur la base de la bryoflore aquatique observée à deux dates différentes. Les données utilisées dans les paragraphes précédents concernent l'année 1972. Nous y ajouterons les données de l'année 1997, également disponibles dans VANDERPOORTEN [2000].

On suppose que les proximités entre objets sont définies par les différentes personnes ou au cours des années successives sur la base des  $q$  mêmes facteurs, qui définissent un *espace commun*<sup>10</sup>, mais que les poids attribués à ces facteurs diffèrent d'une personne ou d'une date à l'autre ou, de manière plus générale, d'un tableau à l'autre. Le modèle est appelé *modèle euclidien pondéré*<sup>11</sup>.

L'analyse donne alors lieu, d'une part, à une configuration moyenne des objets dans l'espace commun de dimension  $q$ , et, d'autre part, à un ensemble de coefficients permettant d'obtenir les configurations individuelles à partir de la configuration moyenne.

Soit  $x_{ik}$  la coordonnée sur l'axe  $k$  ( $k = 1, \dots, q$ ) de l'objet  $i$  ( $i = 1, \dots, n$ ) dans la configuration moyenne et soit  $x_{ikl}$  la coordonnée sur l'axe  $k$  de l'objet  $i$  pour la personne ou la date  $l$  ( $l = 1, \dots, m$ ). On a :

$$x_{ikl} = \sqrt{w_{kl}} x_{ik} .$$

$w_{kl}$  étant le poids attribué à la dimension  $k$  par la personne ou pour la date  $l$ . La racine carrée de ce poids est aussi appelée *coefficient de dimension*<sup>12</sup>.

La distance entre deux objets  $i$  et  $j$ , pour un tableau  $l$ , s'écrit alors :

$$d_{ijl} = \left[ \sum_{k=1}^q (\sqrt{w_{kl}} x_{ik} - \sqrt{w_{kl}} x_{jk})^2 \right]^{1/2} = \left[ \sum_{k=1}^q w_{kl} (x_{ik} - x_{jk})^2 \right]^{1/2} .$$

10. En anglais : *common space*.

11. En anglais : *weighted Euclidean model*.

12. En anglais : *dimension coefficient*.

Dans le cas du positionnement métrique, on utilise le critère d'ajustement suivant :

$$\sum_{l=1}^m \sum_{i < j} [\delta_{ijl} - f_l(d_{ijl})]^2 / \sum_{l=1}^m \sum_{i < j} \delta_{ijl}^2.$$

Dans cette expression,  $\delta_{ijl}$  est la proximité entre les objets  $i$  et  $j$  dans le tableau  $l$ ;  $f_l(d_{ijl})$  est une fonction dont la nature est définie par l'utilisateur (par exemple une fonction linéaire), mais dont les paramètres sont éventuellement différents d'un tableau à l'autre.

L'optimisation vise donc à déterminer :

- les coordonnées  $x_{ik}$  des  $n$  objets dans l'espace commun,
- les poids  $w_{kl}$  des facteurs pour chaque tableau,
- les coefficients relatifs à la transformation des distances  $d_{ijl}$  pour chaque tableau.

Pour le positionnement non métrique, le critère s'écrit :

$$\sum_{l=1}^m \sum_{i < j} [t(\delta_{ijl}) - d_{ijl}]^2 / \sum_{l=1}^m \sum_{i < j} [t(\delta_{ijl})]^2,$$

$t(\delta_{ijl})$  représentant la transformation monotone des  $\delta_{ijl}$ .

Des variantes des critères d'ajustement définis ci-dessus peuvent être utilisées, comme nous le verrons au paragraphe 5.1.

A titre d'illustration, les matrices de proximités entre stations, définies à partir des relevés réalisés en 1972 et en 1997 et repris en annexe, ont été soumises à l'analyse non métrique.

La figure 8 donne la valeur du critère de qualité de l'ajustement, les coordonnées des stations dans l'espace commun et les coefficients par lesquels il faut multiplier les coordonnées des stations pour obtenir la position des stations pour chacune des années. La figure 9 donne la représentation graphique des neuf stations dans l'espace commun.

On peut tout d'abord remarquer que la valeur du critère de qualité de la configuration, égale à 0,125 (figure 8), est sensiblement plus élevée que la valeur obtenue pour l'année 1972, qui est égale à 0,035 (figure 5). L'analyse de la seule année 1997, non reprise ici, donne une valeur de 0,013. Le fait que le critère présente une valeur supérieure pour l'analyse simultanée des deux années s'explique par le caractère plus contraignant de cette analyse : le positionnement des stations pour une année donnée est obtenu en multipliant les coordonnées des points moyens sur un axe par une constante. Ainsi, pour Anseremme, par exemple, les coordonnées dans l'espace commun sont égales à 1,59943 et 1,14684 (figure 8). Pour les années 1972 et 1997, elles valent respectivement :

$$(1, 59943)(1, 24227) = 1, 98692 \quad \text{et} \quad (1, 14684)(0, 67584) = 0, 77508$$

Obs	annee	_TYPE_	Station	Dim1	Dim2
1		CRITERION		0.12453	.
2		CONFIG	M_ANS	1.59943	1.14684
3		CONFIG	M_PLANT	1.12512	1.39785
4		CONFIG	M_AND	1.34798	0.47679
5		CONFIG	M_IVOZ	-0.18505	0.27868
6		CONFIG	M_MONS	-0.52759	0.66481
7		CONFIG	M_VISE	-0.96263	-0.12748
8		CONFIG	S_NAM	-0.58844	-1.11210
9		CONFIG	S_CHAR	-1.16655	-1.27075
10		CONFIG	S_FLOR	-0.64227	-1.45464
11	72	DIAGCOEF		1.24227	0.67584
12	97	DIAGCOEF		0.96823	1.03080

Figure 8. Positionnement non métrique: coordonnées des stations dans l'espace commun.

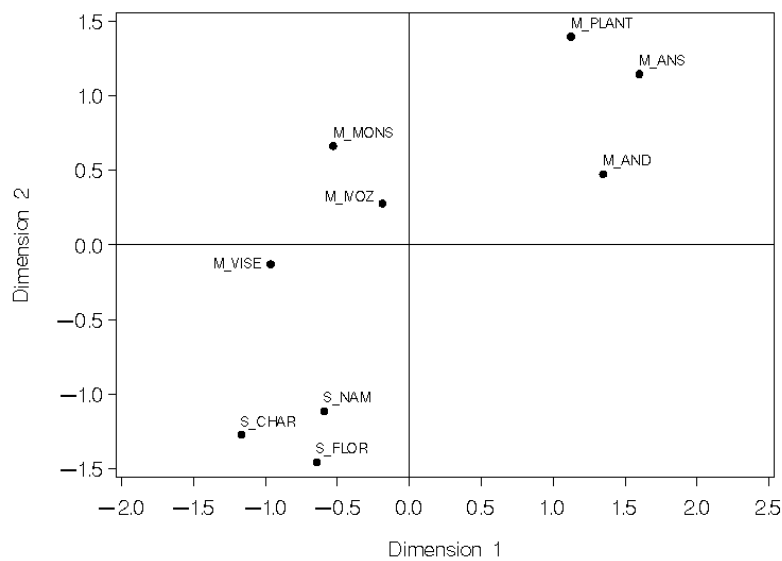


Figure 9. Positionnement non métrique: représentation des stations dans l'espace commun.

et  $(1,59943)(0,96823) = 1,54862$  et  $(1,14684)(1,03080) = 1,18216$ .

Les quatre constantes qui interviennent dans les relations ci-dessus sont reprises dans la figure 8. Elles peuvent être représentées dans l'espace de

dimension 2 par les deux points de coordonnées :

$$(1, 24227 ; 0, 67584) \quad \text{et} \quad (0, 96823 ; 1, 03080) .$$

Cet espace est appelé *espace des poids* ou encore *espace des sujets*<sup>13</sup>.

Ces constantes ont une interprétation géométrique, puisqu'elles indiquent dans quelle mesure chacune des dimensions doit être étirée ou au contraire comprimée pour que l'espace commun devienne l'espace propre à une année. On voit ainsi que, pour obtenir une représentation des stations en 1972, il faut dilater la figure 9 selon la dimension 1 (coefficient égal à 1,24) et la comprimer selon la dimension 2 (coefficient égal à 0,68). Pour l'année 1997, on a la situation inverse, puisqu'il faut comprimer l'axe 1 et étirer l'axe 2.

La comparaison de la représentation des stations dans l'espace commun aux deux années (figure 9) et de la représentation pour l'année 1972 uniquement (figure 3) fait apparaître, à première vue, d'importantes différences qui s'expliquent en grande partie par les procédures de calcul des coordonnées. Des informations à ce sujet seront données au paragraphe 5.3 et la comparaison des deux représentations sera développée au paragraphe 5.5.

## 5. INTERPRÉTATION DES RÉSULTATS

### 5.1. Qualité de la représentation

Dans les paragraphes 3 et 4, le critère de qualité d'ajustement retenu est la racine carrée de la somme des carrés des résidus, cette somme étant divisée par la somme des carrés des  $\delta_{ij}$  ou des  $t(\delta_{ij})$ , selon qu'on a affaire au positionnement métrique ou non métrique. Cette forme de standardisation conduit au paramètre souvent appelé *stress-1*, par référence à la formule numérotée 1 par KRUSKAL [1964]. Une autre standardisation consiste à remplacer la somme des carrés des proximités, ou d'une transformation monotone de ces proximités, par la somme des carrés des écarts par rapport à la moyenne de ces proximités ou de leur transformation. On obtient alors le paramètre *stress-2*, dont les valeurs sont plus grandes que celles du paramètre *stress-1*.

La valeur du critère de qualité de l'ajustement dépend de plusieurs facteurs, ce qui rend son interprétation délicate. Elle dépend, en effet, du nombre d'objets pris en compte, de la présence ou non d'*ex-aequo*, de la méthode de positionnement utilisée et de la dimension de l'espace retenu. Ces différents facteurs sont discutés par KRUSKAL et WISH [1978]. Ces auteurs signalent, notamment, que l'effet indirect du nombre d'objets  $n$  et de la dimension de l'espace de représentation  $q$  est peu important si  $n > 4q$ . Ils signalent également que l'utilisation du positionnement métrique au lieu du positionnement non métrique a le plus souvent peu d'effet sur la configuration, mais augmente toujours, et parfois de manière importante, la valeur du paramètre de qualité de l'ajustement, comme cela a d'ailleurs été observé pour l'exemple au paragraphe 3.2.

---

13. En anglais : *weight space* ou *subject space*.

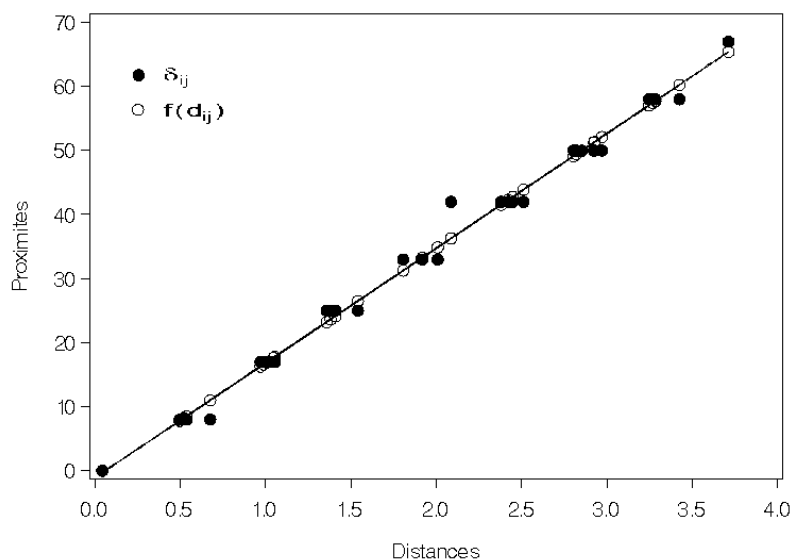


Figure 10. Positionnement métrique: diagramme de dispersion des proximités  $\delta_{ij}$  et des distances transformées  $f(d_{ij})$  en fonction des distances  $d_{ij}$ .

Ce critère de qualité de l'ajustement est un critère global. Pour une étude plus détaillée, on peut examiner individuellement les résidus, à l'aide de diagrammes de dispersion notamment, parfois dénommés diagrammes de SHEPARD [EVERITT et RABE-HESKETH, 1997].

Dans le cas du positionnement métrique, on peut représenter les proximités  $\delta_{ij}$  en fonction des distances  $d_{ij}$  et, sur le même graphe, représenter la fonction  $f(d_{ij})$  en fonction des distances. Sur un tel graphique, les écarts entre les  $\delta_{ij}$  et les valeurs de  $f(d_{ij})$  correspondent aux résidus et l'allure générale du nuage de points permet de vérifier si la fonction  $f(d_{ij})$  est adéquate. Si par exemple, une relation linéaire a été retenue, mais que le diagramme montre l'existence d'une relation non linéaire, la valeur du critère de qualité de l'ajustement sera inutilement élevée et il peut se justifier de réanalyser les données en utilisant une fonction plus adéquate.

La figure 10 donne, pour l'exemple traité au paragraphe 3.1, la représentation graphique décrite ci-dessus. On vérifie que la dispersion des points autour de la droite est, dans l'ensemble, assez faible, et que la transformation linéaire des distances ne doit pas être remise en cause. L'adéquation du modèle est d'ailleurs confirmée par le coefficient de corrélation des  $\delta_{ij}$  et  $f(d_{ij})$ , qui est égal à 0,9968.

Pour le positionnement non métrique, on peut porter en abscisse les proximités  $\delta_{ij}$  et en ordonnée, d'une part, les distances  $d_{ij}$  et, d'autre part, les proximités transformées  $t(\delta_{ij})$ . La figure 11, qui concerne l'exemple du paragraphe 3.2, est une illustration de ce type de graphique. Les écarts entre les  $\delta_{ij}$  et les  $t(\delta_{ij})$  correspondent aux résidus et les segments de droite reliant les  $t(\delta_{ij})$  correspondant

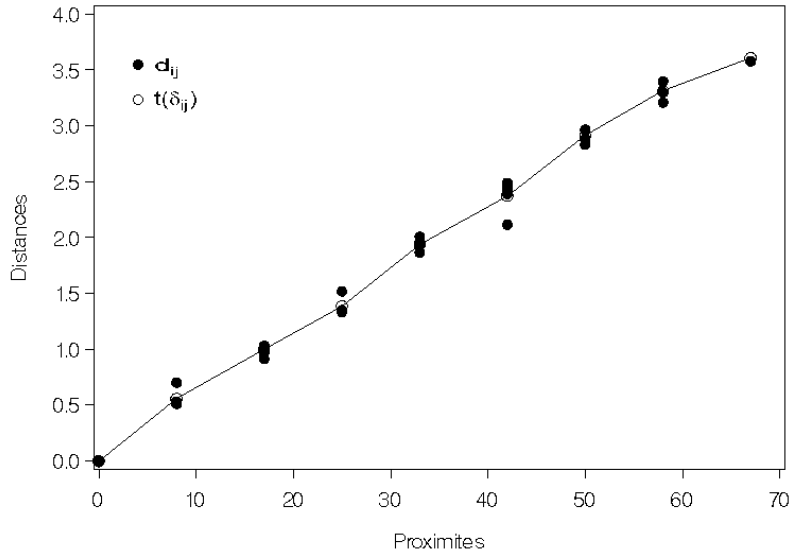


Figure 11. Positionnement non métrique : diagramme de dispersion des distances  $d_{ij}$  et des proximités transformées  $t(\delta_{ij})$  en fonction des proximités  $\delta_{ij}$ .

aux  $\delta_{ij}$  croissants ne sont jamais de pente négative, comme nous l'avons signalé au paragraphe 3.2.

Les graphiques décrits ci-dessus permettent de donner une interprétation géométrique aux critères de qualité de l'ajustement, qui sont des rapports de moyennes quadratiques. Pour le coefficient *stress-1*, le numérateur est la moyenne quadratique des distances entre les points du diagramme de dispersion et la ligne constituée par la jonction des valeurs ajustées et le dénominateur est la moyenne quadratique des distances entre les points et l'axe des abscisses, toutes ces distances étant mesurées parallèlement à l'axe des ordonnées.

Pour le coefficient *stress-2*, l'interprétation du numérateur est identique à celle donnée ci-dessus, mais le dénominateur est la moyenne quadratique des distances mesurées verticalement entre les points du graphique et une droite horizontale d'équation  $d = \bar{d}$  ou  $\delta = \bar{\delta}$ , selon qu'on a affaire au positionnement métrique ou non métrique,  $\bar{d}$  et  $\bar{\delta}$  représentant la moyenne des distances  $d_{ij}$  ou des proximités  $\delta_{ij}$ . Dans le cas du positionnement métrique, lorsque la fonction  $f(d_{ij})$  est la transformation linéaire, le critère est lié au coefficient de détermination  $r^2$  de la régression simple des proximités en fonction des distances :

$$stress-2 = \sqrt{1 - r^2}.$$

Ainsi, pour l'exemple du paragraphe 3.1, le critère *stress-2* est égal à :

$$\sqrt{1 - 0,9968^2} = 0,08,$$

0,9968 étant la corrélation des proximités et des distances.

## 5.2. Dimension de l'espace de représentation

Le choix de la dimension  $q$  de l'espace permettant une représentation satisfaisante des proximités est fort semblable au choix du nombre de composantes à retenir dans les analyses factorielles. La valeur  $q$  résulte souvent d'un compromis entre qualité et simplicité de la représentation.

Le plus souvent, les représentations dans des espaces de dimensions croissantes sont réalisées et la décroissance du critère de qualité de la représentation en fonction de la dimension de l'espace est examinée de manière à choisir un compromis : lorsque l'augmentation de  $q$  ne provoque qu'une faible réduction du critère, on considère qu'il n'y a pas lieu d'augmenter la dimension. Les possibilités d'interprétation de l'espace retenu entrent également en ligne de compte.

Des informations plus détaillées concernant le choix de  $q$  sont données par KRUSKAL et WISH [1978].

## 5.3. Recherche de directions interprétables

Le positionnement des  $n$  objets dans un espace de dimension  $q$  fixée constitue une configuration. Celle-ci est décrite par les coordonnées des points sur un ensemble de  $q$  axes. Elle pourrait cependant être décrite par une infinité de matrices de coordonnées, chacune de ces matrices correspondant à un système particulier d'axes : on dit que la configuration est invariante par rapport à des transformations rigides telles que translation, rotation, réflexion du système d'axes, car ces transformations ne modifient pas les distances entre les objets. Parmi cette infinité de descriptions possibles d'une même configuration, les logiciels en proposent une seule, en fonction de critères bien définis.

Pour illustrer ce point, nous considérons un exemple classique, décrit dans la plupart des ouvrages traitant du positionnement multidimensionnel. Il s'agit de localiser, dans un espace de dimension deux, une série de villes d'une région donnée à partir des distances par route entre ces villes. Le graphique qui en découle reproduit en général assez fidèlement la carte géographique de la région, mais ce graphique n'est pas orienté : le nord ne se trouve pas automatiquement en haut et l'est à droite de la feuille et les deux axes ne correspondent généralement pas aux directions nord-sud et est-ouest, qui sont des directions privilégiées pour les cartes géographiques.

Pour les tableaux à deux entrées avec un facteur, comme dans l'exemple traité au paragraphe 3, les axes retenus sont tels que :

- la moyenne des projections sur chaque axe est nulle;
- la distance moyenne quadratique des objets à l'origine est égale à l'unité,
- le système d'axes a subi une rotation de manière à obtenir une orientation selon les axes principaux; cela signifie que les axes successifs sont déterminés de manière à maximiser la variance des projections sous la contrainte d'orthogonalité par rapport aux axes précédents.

Une telle standardisation du système de coordonnées a l'avantage de faciliter la comparaison des résultats de différentes analyses. C'est ainsi que nous avons pu vérifier facilement la très grande similitude des configurations obtenues par le positionnement métrique et non métrique (paragraphe 3.2). Elle n'implique cependant pas que les axes obtenus correspondent aux directions les plus intéressantes du point de vue de l'interprétation de la configuration. A cette fin, d'autres axes peuvent être plus utiles.

La recherche de tels axes peut se faire par la régression multiple, à condition de disposer d'une ou de plusieurs variables, observées pour chaque objet et susceptibles d'expliquer la position des objets dans la configuration. Nous allons illustrer la méthode en reprenant l'exemple du paragraphe 3.1. Nous avons déjà signalé que l'axe 1 dans la figure 1 est un axe de richesse en espèces de bryophytes. Si on exprime le nombre d'espèces de bryophytes dans le site  $i$ ,  $y_i$ , en fonction des coordonnées du site  $i$  sur les axes 1 et 2, notées  $x_{1i}$  et  $x_{2i}$ , on obtient l'équation de régression multiple suivante :

$$y = 3,67 + 2,11 x_1 + 0,63 x_2.$$

Cette relation peut être représentée dans le plan  $x_1$  et  $x_2$  par une droite dont les coefficients directeurs sont :

$$c_1 = 2,11/\sqrt{2,11^2 + 0,63^2} = 0,958$$

et

$$c_2 = 0,63/\sqrt{2,11^2 + 0,63^2} = 0,286.$$

Les coefficients directeurs étant les cosinus des angles définis par la droite et les axes  $x_1$  et  $x_2$ , cette droite forme donc un angle de  $17^\circ$  avec l'axe 1 et un angle de  $73^\circ$  avec l'axe 2; elle a été représentée dans la figure 3.

Le coefficient de détermination multiple associé à l'équation de régression calculée ci-dessus est égal à 0,987. Cette valeur est aussi égale au carré de la corrélation entre les coordonnées des stations sur l'axe  $y$  données par l'équation ci-dessus et les nombres d'espèces présentes dans les stations. La direction représentée par l'axe  $y$  est donc bien une direction interprétable: l'axe  $y$  est un axe qui représente un gradient de richesse en espèces.

Pour qu'un des axes fournis par le logiciel soit directement interprétable, il faudrait que l'axe  $y$  calculé ci-dessus coïncide avec l'axe  $x_1$  ou l'axe  $x_2$  et que les coordonnées des stations sur cet axe soient fortement corrélées avec les observations  $y_i$ . En d'autres termes, il faudrait que la régression multiple de  $y$  en fonction de  $x_1$  et de  $x_2$  soit telle que le coefficient de détermination multiple soit suffisamment élevé et que le coefficient de régression partielle d'une des deux variables soit très faible de manière à ce que l'axe  $y$  forme un angle très faible avec un des axes initiaux. Dans le cas de l'exemple traité ci-dessus, le coefficient de régression partielle pour la variable  $x_2$  est non significatif au niveau de signification 0,05, la probabilité associée au test de signification étant de 0,057. On peut par conséquent considérer que le nombre d'espèces  $y_i$  est avant tout fonction de  $x_{1i}$  et interpréter directement l'axe 1 comme un axe exprimant la richesse en espèces, comme nous l'avons signalé au paragraphe 3.1.

La procédure qui vient d'être décrite pour une variable externe peut être répétée pour d'autres variables, pour autant que celles-ci soient disponibles. On pourrait ainsi placer un deuxième axe, voire un troisième ou un quatrième, dans le plan défini par  $x_1$  et  $x_2$ . Ces axes ne seraient évidemment pas forcément orthogonaux. Des exemples concrets sont donnés par KRUSKAL et WISH [1978], notamment.

Pour les tableaux à trois entrées avec deux facteurs, la situation est un peu différente. En effet, à un facteur d'échelle près, les axes sont uniques, dans le sens où l'étirement ou la contraction des axes qui permet de passer de la configuration moyenne à une configuration individuelle ne peut se faire que selon les axes obtenus par le logiciel. Ceux-ci jouent donc un rôle particulier, mais cela ne signifie pas qu'ils ont nécessairement une interprétation concrète. Ainsi, aucune signification particulière n'a pu être donnée aux axes de la figure 10. Une éventuelle rotation des axes peut être envisagée pour faciliter l'interprétation de cet espace commun mais, dans ce cas, la représentation des poids ou sujets ne peut plus être interprétée directement dans ce nouveau système de coordonnées.

Une autre conséquence du mode de détermination des axes dans le cas des tableaux à trois entrées est que les coordonnées sur les différents axes ne sont plus non corrélés. Ainsi l'examen de la figure 9 montre clairement qu'il existe une corrélation positive élevée entre les abscisses et les ordonnées des stations. Le calcul de la corrélation entre  $x_{1i}$  et  $x_{2i}$  confirme cette impression, le coefficient de corrélation obtenu étant égal à 0,76.

#### 5.4. Autres aides à l'interprétation

La réalisation d'une classification numérique des objets sur la base de la matrice de proximités et le report des groupes obtenus sur le graphique donnant la configuration peut aider l'utilisateur dans son effort d'interprétation. L'identification de caractéristiques communes aux objets d'un même groupe peut, en effet, suggérer une interprétation de ces groupes et faire apparaître d'éventuelles directions interprétables.

Il peut également être utile de compléter la représentation graphique de la configuration par des informations concernant les proximités entre objets. Cela peut se faire en reliant, dans la configuration, les objets dont la proximité est inférieure à un seuil donné. Plusieurs seuils peuvent d'ailleurs être retenus, ceux-ci étant symbolisés par des traits de couleurs ou de types différents.

Certaines formes de contradictions entre les proximités  $\delta_{ij}$  et les distances  $d_{ij}$  dans la configuration peuvent ainsi être mises en évidence. Par exemple deux points relativement proches sur le graphique peuvent être caractérisés par une valeur  $\delta_{ij}$  importante. Une telle situation peut se présenter si l'espace retenu est de dimension trop petite. La prise en compte d'un axe supplémentaire écarterait alors ces deux points. Des discordances peuvent aussi s'expliquer par le fait que le positionnement multidimensionnel reproduit mieux les grandes valeurs de proximités (la structure globale) que les petites valeurs (la structure locale). Pour obtenir une meilleure représentation de cette structure locale, il peut d'ailleurs

se justifier de réaliser des analyses séparées pour des groupes d'objets assez proches sur le graphique ou sur des groupes d'objets identifiés par la classification numérique, à condition que ces groupes ne soient pas d'effectifs trop réduits. De telles analyses complémentaires peuvent parfois révéler des structures qui sont cachées lorsque tous les objets sont pris en compte simultanément.

### 5.5. Comparaison de deux configurations

Nous avons signalé, au paragraphe 5.2, que les configurations sont invariantes par rapport aux translations, rotations et réflexions et que, de plus, le facteur d'échelle est arbitraire pour la plupart des méthodes de positionnement multidimensionnel. Il en résulte que deux configurations, fondamentalement très proches, peuvent apparaître fort différentes dans les représentations graphiques.

Pour faciliter la comparaison, on peut modifier ces représentations sans modifier les configurations. Soit  $\mathbf{X}_0$  la matrice des coordonnées des  $n$  objets dans l'espace de dimension  $q$  pour la première configuration et soit  $\mathbf{Y}_0$  la matrice correspondante pour la deuxième configuration. On suppose que, pour chacune des deux représentations, le nuage de points est centré sur l'origine, comme c'est généralement le cas. Dans la négative, il suffit de soustraire de chaque colonne de la matrice des coordonnées, la moyenne de la colonne en question.

Si la distance moyenne des points au centre de gravité n'est pas identique pour les deux représentations graphiques, une standardisation est nécessaire. Celle-ci peut être réalisée en multipliant les matrices  $\mathbf{X}_0$  et  $\mathbf{Y}_0$  par une constante :

$$\mathbf{X} = k_1 \mathbf{X}_0 \quad \text{et} \quad \mathbf{Y} = k_2 \mathbf{Y}_0$$

avec 
$$k_1 = \sqrt{\text{tr}(\mathbf{X}'_0 \mathbf{X}_0) / nq} \quad \text{et} \quad k_2 = \sqrt{\text{tr}(\mathbf{Y}'_0 \mathbf{Y}_0) / nq}.$$

Dans les relations ci-dessus,  $nq$  représente la somme des carrés des  $q$  coordonnées pour les  $n$  objets après transformation, alors que la trace de la matrice  $\mathbf{X}'_0 \mathbf{X}_0$  ou  $\mathbf{Y}'_0 \mathbf{Y}_0$  représente la somme des carrés des  $q$  coordonnées pour les  $n$  objets avant transformation. Les nouvelles coordonnées reprises dans les matrices  $\mathbf{X}$  et  $\mathbf{Y}$  sont donc telles que la distance moyenne quadratique des objets à l'origine est égale à l'unité.

Après cette standardisation, on applique à la deuxième configuration une rotation orthogonale :

$$\mathbf{Z} = \mathbf{Y} \mathbf{A}',$$

la matrice  $\mathbf{A}'$ , de dimensions  $q \times q$ , étant telle que la somme des carrés des distances entre les objets dans les deux représentations soit minimum. Autrement dit, la matrice  $\mathbf{A}'$  doit rendre minimum la quantité :

$$\sum_{i=1}^n \sum_{k=1}^q (x_{ik} - z_{ik})^2.$$

On peut montrer que cette matrice s'obtient à partir de la décomposition par valeurs singulières de  $\mathbf{Y}' \mathbf{X}$ . En effet, celle-ci permet d'écrire [HEALY, 1992] :

$$\mathbf{Y}' \mathbf{X} = \mathbf{T} = \mathbf{U} \mathbf{\Lambda} \mathbf{V}',$$

$U$  étant la matrice des vecteurs propres de  $\mathbf{T}\mathbf{T}'$ ,  $V$  la matrice des vecteurs propres de  $\mathbf{T}'\mathbf{T}$  et  $\Lambda$  la matrice diagonale dont les éléments sont les racines carrées des valeurs propres non nulles de  $\mathbf{T}'\mathbf{T}$  ou de  $\mathbf{T}\mathbf{T}'$ . La matrice  $A$  est alors égale à  $VU'$  et la somme des carrés qui a été minimisée se calcule par la relation :

$$\sum_{i=1}^n \sum_{k=1}^q (x_{ik} - z_{ik})^2 = 2 (nq - \text{tr}(\Lambda)) .$$

La méthode décrite ci-dessus s'appelle *analyse procrustéenne*<sup>14</sup>. Elle peut être généralisée au cas de plus de deux configurations. Des informations à ce sujet sont données par EVERITT et RABE-HESKETH [1997], notamment.

Considérons, à titre d'illustration, la comparaison de la représentation des stations obtenue à partir des relevés de la bryoflore aquatique en 1972 (figure 3) et de la représentation des stations dans l'espace commun, obtenue lors de l'analyse simultanée des deux années d'observation (figure 9).

La matrice  $X$  contient les coordonnées de la configuration reprises dans la figure 2 et la matrice  $Y$  est constituée des coordonnées de la configuration données à la figure 8. Ces matrices sont déjà standardisées, les sommes des colonnes étant nulles et les sommes des carrés des éléments de chacune des matrices étant égales à 18.

L'application des relations ci-dessus a conduit au résultat suivant :

$$A' = \begin{bmatrix} 0,7712 & -0,6366 \\ 0,6366 & 0,7712 \end{bmatrix} .$$

La valeur 0,7469 correspond au cosinus de l'angle  $\alpha$  que forment les nouveaux axes avec les axes antérieurs. Le nouveau système de coordonnées pour la représentation de l'espace commun lors de la prise en compte des deux années subit donc une rotation de :

$$\alpha = \arccos(0,7712) = 40^\circ$$

dans le sens trigonométrique par rapport au système de coordonnées initiales, la position des points étant inchangée. La figure 9 subit par conséquent une rotation de  $40^\circ$  dans le sens horlogique avant d'être superposée à la figure 3.

La figure 12 donne la représentation simultanée des deux nuages de points. On remarque que la concordance entre les deux ensembles de points est excellente selon l'axe 1, la somme des carrés des écarts entre les coordonnées des deux ensembles sur l'axe 1 étant très faible :

$$\sum_{i=1}^n (x_{i1} - z_{i1})^2 = 0,49 .$$

Pour le deuxième axe, la concordance est un peu moins bonne puisque :

$$\sum_{i=1}^n (x_{i2} - z_{i2})^2 = 2,61 .$$

---

14. En anglais : *Procrustes analysis*.

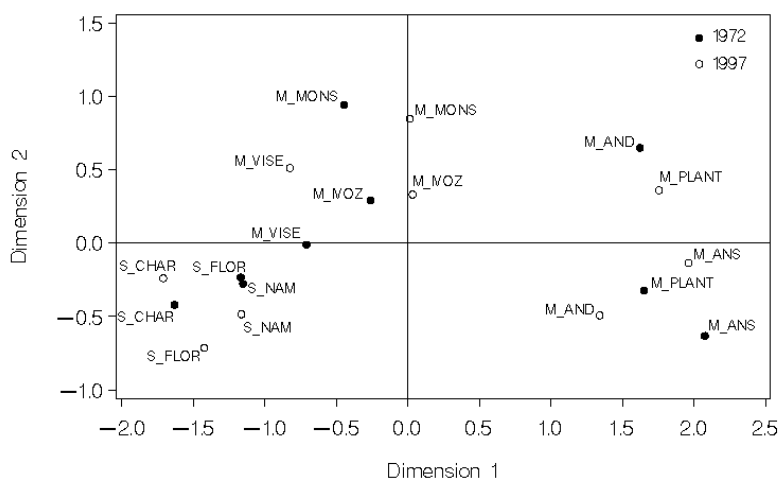


Figure 12. Comparaison de la configuration relative à l'année 1972 et de la configuration dans l'espace commun relatif aux années 1972 et 1997.

Pour l'espace de dimension 2, on a donc une somme de carrés d'écarts totale égale à :

$$0,49 + 2,61 = 3,09.$$

La distance moyenne quadratique entre les points des deux configurations est donc égale à :

$$\sqrt{3,09/18} = 0,41,$$

et on conclut que les deux représentations sont relativement proches, alors que les figures 3 et 9 sont, à première vue, fort différentes.

Un examen plus détaillé des écarts entre les positions des différentes stations montrerait que près de 45 % de la somme des carrés des écarts totale ci-dessus est liée à la station d'Andenne.

## 6. QUELQUES INFORMATIONS COMPLÉMENTAIRES

### 6.1. Algorithmes de minimisation et minima locaux

Les algorithmes permettant de déterminer les coordonnées des objets dans l'espace de dimension  $q$  ainsi que les distances ajustées sont de type itératif. Ils modifient une configuration initiale par itérations successives de manière à minimiser la somme des carrés des résidus. La configuration initiale peut être déterminée par le logiciel à partir de la matrice des proximités, ou bien être

déterminée par un générateur de nombres aléatoires ou encore être donnée par l'utilisateur.

Ces algorithmes sont susceptibles de ne pas converger vers le minimum ou de conduire à des minima locaux. Le problème est analogue à celui rencontré en régression non linéaire. Les logiciels prévoient des règles d'arrêt du processus itératif, d'une part, en fixant le nombre maximum d'itérations et, d'autre part, en fixant des critères de convergence. Si le nombre maximum d'itérations est atteint au cours d'un processus d'optimisation, cela signifie que les critères de convergence n'ont pas été vérifiés et donc que le minimum n'a pas été trouvé. Il peut, dans ce cas, être utile de relancer la procédure de calcul avec un plus grand nombre d'itérations ou avec une meilleure solution initiale et de vérifier si les critères de convergence n'ont pas été fixés de manière exagérément sévère.

Pour s'assurer que l'optimum trouvé n'est pas un minimum local, il peut être utile de vérifier la stabilité de la solution en modifiant la configuration initiale.

On peut noter que le graphique donnant l'évolution de la valeur du critère de qualité de l'ajustement en fonction de la dimension de l'espace  $q$  est toujours décroissant et présente globalement une concavité tournée vers le haut. Le non-respect de l'une ou l'autre de ces deux caractéristiques peut être un indice de la présence de convergence incomplète ou de minimum local [KRUSKAL et WISH, 1978].

## 6.2. Positionnement multidimensionnel et analyse en composantes principales

Considérons une matrice  $\mathbf{X}$ , de dimensions  $n \times p$ , reprenant des observations quantitatives réalisées pour  $p$  variables sur  $n$  objets. Après une analyse en composantes principales de cette matrice, on peut représenter les  $n$  objets dans l'espace des  $p$  axes principaux et, dans cet espace de dimension  $p$ , les distances  $d_{ij}$  entre les objets sont identiques aux distances euclidiennes  $\delta_{ij}$  entre les objets qu'on peut calculer à partir des variables initiales. Par contre, si on ne retient que les  $q$  premières composantes principales, les distances  $d_{ij}$  entre les objets dans ce sous-espace ne sont plus égales à  $\delta_{ij}$ , les écarts étant, en moyenne, d'autant plus faibles que les axes qui ont été négligés sont de faible importance.

Concrètement, pour représenter les objets dans l'espace de dimension  $q$ , il suffit de calculer les  $q$  premiers vecteurs propres normés à la valeur propre de la matrice  $\mathbf{X} \mathbf{X}'$ ,  $\mathbf{X}$  étant la matrice des variables initiales centrées.

On peut par ailleurs montrer que les éléments de la matrice  $\mathbf{X} \mathbf{X}'$ , notés  $y_{ij}$ , peuvent être obtenus à partir de la matrice des distances  $\delta_{ij}$ , par la relation suivante [EVERITT et RABE-HEKETH, 1997]:

$$y_{ij} = -\frac{1}{2} [\delta_{ij}^2 - \delta_{i.}^2 - \delta_{.j}^2 + \delta_{..}^2]$$

avec

$$\delta_{i.}^2 = \frac{1}{2} \sum_{j=1}^n \delta_{ij}^2, \quad \delta_{.j}^2 = \frac{1}{2} \sum_{i=1}^n \delta_{ij}^2 \quad \text{et} \quad \delta_{..}^2 = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \delta_{ij}^2.$$

Tableau 2. Coordonnées des neuf stations dans le plan obtenues par l'analyse en composantes principales.

Stations	Avant standardisation		Après standardisation	
	axe 1	axe 2	axe 1	axe 2
Anseremme	37,20	- 4,32	2,16	- 0,25
La Plante	29,36	- 3,14	1,71	- 0,18
Andenne	27,16	0,92	1,58	0,05
Ivoz-Ramez	- 3,89	4,40	- 0,23	0,26
Monsin	- 7,56	17,53	- 0,44	1,02
Visé	- 12,42	0,06	- 0,72	0,00
Namur	- 20,47	- 4,10	- 1,19	- 0,24
Charleroi	- 28,90	- 7,27	- 1,68	- 0,42
Floreffe	- 20,47	- 4,10	- 1,19	- 0,24

Cela signifie que la représentation des objets dans l'espace de dimension  $q$  peut être réalisée par analyse en composantes principales à partir de la connaissance des seules distances entre objets. On notera toutefois que la solution obtenue n'est pas identique à la solution fournie par les méthodes de positionnement multidimensionnel car les critères d'optimisation retenus par les deux méthodes sont différents.

Lorsque la matrice des proximités n'est pas une matrice de distances euclidiennes, on peut toujours calculer les  $y_{ij}$  à partir de  $\delta_{ij}$ . La matrice des  $y_{ij}$  n'est cependant plus nécessairement une matrice semi-définie positive, c'est-à-dire que ses valeurs propres ne sont pas toutes positives ou nulles. Si les  $q$  premières valeurs propres sont positives, les vecteurs propres associés à ces valeurs propres correspondent, à une constante près, aux coordonnées des objets dans un espace de dimension  $q$ .

Cette configuration obtenue par l'analyse en composantes principales peut constituer un point de départ du processus itératif de calcul pour le positionnement multidimensionnel. C'est d'ailleurs l'approche utilisée, par défaut, par la procédure MDS de SAS.

A titre d'illustration, les coordonnées des neuf stations dans l'espace de dimension 2 ont été calculées par l'analyse en composantes principales, à partir des proximités  $\delta_{ij}$  données dans la figure 4. Les résultats obtenus sont repris dans la première partie du tableau 2.

Les sommes des carrés des coordonnées sur les deux axes sont respectivement égales à 4.884,2 et 442,7. Ces valeurs sont aussi les deux premières valeurs propres de la matrice  $\mathbf{X} \mathbf{X}'$  reconstruite à partir des distances  $\delta_{ij}$ . Si on souhaite standardiser les coordonnées de manière à obtenir une distance moyenne quadratique égale à l'unité, il faut multiplier les coordonnées obtenues par l'analyse en composantes principales par la constante  $k$  :

$$k = \sqrt{18/(4.884,2 + 442,7)} = 0,05813.$$

Les résultats sont donnés dans la partie droite du tableau 2. Les valeurs sont égales, aux erreurs d'arrondis près, à la configuration initiale utilisée par SAS et sont relativement proches de la configuration finale, donnée à la figure 2.

## 7. CONCLUSIONS

Dans cette note, nous avons décrit quelques méthodes de positionnement multidimensionnel et nous les avons illustrées par un exemple simple traité avec la procédure MDS du logiciel SAS.

Ces méthodes, qui ont comme objectif de décrire les positions relatives des différents objets dans un espace défini par deux ou plusieurs axes, présentent des points communs avec d'autres méthodes multivariées.

Ainsi, nous avons signalé que le positionnement multidimensionnel s'applique notamment aux matrices symétriques de proximités entre objets et que, dans ce cas, le point de départ est identique au point de départ d'une classification numérique (paragraphe 2.2). Outre le fait que les deux méthodes ne nécessitent pas un tableau de données dont les lignes sont les objets et dont les colonnes sont les variables, elles visent toutes deux à analyser les proximités, par la représentation des objets dans un espace de dimension réduite dans le cas du positionnement multidimensionnel, et par le regroupement d'objets en classes dans le cas de la classification numérique. Cette différence dans l'approche rend les deux méthodes plus complémentaires que concurrentes (paragraphe 5.4).

La représentation d'objets dans un espace de dimension réduite peut également être obtenue par d'autres méthodes multivariées, notamment par l'analyse en composantes principales et par l'analyse des correspondances. La relation entre l'analyse en composantes principales et le positionnement multidimensionnel a été discuté au paragraphe 6.2. Appliqué à une matrice de distances euclidiennes, le positionnement métrique conduit à des résultats analogues à ceux obtenus par l'analyse en composantes principales. Appliqué à un tableau de distances au sens du chi-carré entre lignes (ou entre colonnes), le positionnement métrique donne une configuration analogue à la représentation des lignes (ou des colonnes) obtenue par l'analyse factorielle des correspondances. Du point de vue de la représentation des objets, l'analyse en composantes principales et l'analyse des correspondances peuvent par conséquent être considérées comme des cas particuliers de positionnement multidimensionnel.

Des informations complémentaires relatives aux liens existant entre différentes méthodes multivariées et le positionnement multidimensionnel sont données par EVERITT et RABE-HESKETH [1997].

Le positionnement multidimensionnel constitue donc une technique très souple dans la mesure où elle permet d'analyser des matrices de proximités variées, tant au point de vue de la nature des proximités que de la structure des matrices, comme nous l'avons signalé au paragraphe 2.

## BIBLIOGRAPHIE

- BORG I., GROENEN P. [1997]. *Modern multidimensional scaling. Theory and applications*. New York, Springer, 471 p.
- CHANDON J.L., PINSON S. [1981]. *Analyse typologique. Théorie et applications*. Paris, Masson, 254 p.
- COX T.F., COX M.A.A. [2001]. *Multidimensional scaling*. New York, Chapman-Hall, 328 p.
- DE SOETE G., CARROLL J.D. [1998]. Multidimensional scaling. *In* : ARMITAGE P., COLTON T. (edit.). *Encyclopedia of statistics* (vol. 4). New York, Wiley, 2716-2725.
- EVERITT B. [1993]. *Cluster analysis*. New York, Wiley, 170 p.
- EVERITT B., RABE-HESKETH S.C. [1997]. *The analysis of proximity data*. London, Arnold, 178 p.
- HEALY M.J.R. *Matrices for statistics*. Oxford, Clarendon, 89 p.
- KRUSKAL J.B. [1964]. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29, 1-27.
- KRUSKAL J.B., WISH M. [1978]. *Multidimensional scaling*. Beverly Hills, Sage, 95 p.
- LEGENDRE L., LEGENDRE P. [1984]. *Ecologie numérique : la structure des données écologiques*. Paris, Masson, 335 p.
- SAS INSTITUTE INC [1989]. *SAS/STAT Users' guide, version 6*. Fourth edition (2 volumes). Cary NC: SAS Institution Inc. 943 + 946 p.
- VANDERPOORTEN A. [2000]. *Hydrochemical determinism and molecular systematics in the genus Amblystegium (Musci). Application to the biomonitoring of surface waters* (Thèse de doctorat). Gembloux, Faculté universitaire des Sciences agronomiques. 99 p. + annexes.
- YOUNG F.W. [1985]. Multidimensional scaling. *In* : KOTZ S., JOHNSON N.L. (edit.). *Encyclopedia of statistical sciences* (vol. 5). New York, Wiley, 649-659.

Annexe. Présence et absence de 12 bryophytes aquatiques (a, . . . , l)  
dans 9 stations : relevés réalisés en 1972 et en 1997 [VANDERPOORTEN, 2000].

Relevés réalisés en 1972

Stations	a	b	c	d	e	f	g	h	i	j	k	l
M-ANS	0	0	1	1	1	1	0	1	1	0	1	1
M-PLANT	0	0	1	0	1	1	0	1	1	0	1	1
M-AND	0	0	1	0	1	1	0	0	1	1	1	1
M-IVOZ	0	0	1	0	1	1	0	0	0	0	0	0
M-MONS	0	0	1	0	1	1	1	0	0	0	0	0
M-UISE	0	0	1	0	0	1	0	0	0	0	0	0
S-NAM	0	0	1	0	0	0	0	0	0	0	0	0
S-CHAR	0	0	0	0	0	0	0	0	0	0	0	0
S-FLOR	0	0	1	0	0	0	0	0	0	0	0	0

Relevés réalisés en 1997

Stations	a	b	c	d	e	f	g	h	i	j	k	l
M-ANS	1	1	1	1	1	1	1	1	1	1	1	1
M-PLANT	1	1	1	1	1	1	1	1	1	1	1	1
M-AND	1	0	1	1	1	1	1	1	1	1	1	1
M-IVOZ	1	1	1	1	1	1	0	1	0	0	0	0
M-MONS	1	1	1	1	1	1	0	1	0	0	0	0
M-UISE	0	0	1	0	1	1	1	0	0	0	0	0
S-NAM	0	0	1	1	0	0	0	1	0	0	0	1
S-CHAR	1	0	1	0	0	0	0	0	0	0	0	1
S-FLOR	0	0	1	1	0	0	0	0	0	0	0	1

La collection

### **NOTES DE STATISTIQUE ET D'INFORMATIQUE**

réunit divers travaux (documents didactiques, notes techniques, rapports de recherche, publications, etc.) émanant des services de statistique et d'informatique de la Faculté universitaire des Sciences agronomiques et du Centre de Recherches agronomiques de Gembloux (Belgique).

La liste des notes disponibles peut être obtenue sur simple demande à l'adresse ci-dessous :

*Faculté universitaire des Sciences agronomiques  
Unité de Statistique et Informatique  
Avenue de la Faculté d'Agronomie, 8  
B-5030 GEMBLoux (Belgique)  
E-mail : statinfo@fsagx.ac.be*

Plusieurs notes sont directement accessibles à l'adresse Web suivante, section Publications :

*<http://www.fsagx.ac.be/si/>*

En relation avec certaines notes, des programmes spécifiques sont également disponibles à la même adresse, section Macros.

Quelques titres récents sont cités ci-après :

- CLAUSTRIAUX J.J., IEMMA A.F. [1999]. A propos des qualificatifs complet, orthogonal et équilibré en analyse de la variance. *Notes Stat. Inform.* (Gembloux) 99/2, 14 p.
- IEMMA A.F., CLAUSTRIAUX J.J. [1999]. Etude des hypothèses de l'analyse de la variance à deux critères de classification : approche par l'exemple. *Notes Stat. Inform.* (Gembloux) 99/3, 14 p.
- PALM R. [1999]. L'analyse discriminante décisionnelle : principes et application. *Notes Stat. Inform.* (Gembloux) 99/4, 41 p.
- PALM R. [1999]. Indices d'aptitude des procédés de production. *Notes Stat. Inform.* (Gembloux) 99/5, 26 p.
- PALM R. [2000]. L'analyse de la variance multivariée et l'analyse canonique discriminante : principes et applications. *Notes Stat. Inform.* (Gembloux) 2000/1, 40 p.
- PALM R., IEMMA A.F. [2002]. Conditions d'application et transformations de variables en régression linéaire. *Notes Stat. Inform.* (Gembloux) 2002/1, 34 p.
- BROSTAUX Y. [2002]. Introduction à l'environnement de programmation statistique R. *Notes Stat. Inform.* (Gembloux) 2002/2, 22 p.