UNE INTRODUCTION AU POSITIONNEMENT MULTIDIMENSIONNEL

Dominique Desbois
INRA-SAE2 Nancy et SCEES - 251, rue de Vaugirard, 75732 Paris Cedex 15.
Courriel :dominique.desbois@agriculture.gouv.fr Fax : +33 1 49 55 85 00

Résumé :
*Le **positionnement multidimensionnel**, est une méthode de représentation graphique d'un ensemble de similarités ou de dissimilarités mais aussi une procédure de construction d'échelles communes à un ensemble d'attributs subjectifs. Cette note a pour but d'aider les utilisateurs dans la mise en oeuvre du positionnement multidimensionnel au moyen des procédures SPSS PROXIMITIES et ALSCAL. Cette mise en oeuvre concerne l'analyse des tableaux de dissimilarité construits à partir de tableaux multidimensionnels de données individuelles. Le listage des résultats obtenus à partir d'exemples vient illustrer l'exposé théorique consacré aux méthodes.*

Abstract :
Multidimensional positioning is a graphical method used to chart a set of similarities or dissimilarities but also a procedure used to build common scales from a set of subjective attributes. The aim of this note is to help the users in the implementation of multidimensional scaling by means of SPSS procedures, PROXIMITIES and ALSCAL. This implementation relates to the analysis of the dissimilarity tables built starting from multidimensional tables of individual data. Starting from examples, the listing of outputs obtained comes to illustrate the theoretical part devoted to the methods.

Mots clés :
Positionnement multidimensionnel, analyse des proximités, tableaux de dissimilarités, analyse factorielle sur tableaux de distance , logiciel statistique SPSS.

1. L'analyse des préférences pour l'étude de la rationalité des choix

Sur quels critères choisit-on un aliment : sa couleur, sa texture, son odeur, son prix ? Quels sont les éléments déterminant le choix des électeurs : l'appartenance politique du candidat, les positions exprimées, son charisme individuel ? Quelles caractéristiques interpersonnelles rentrent en jeu dans le comportement de sélection mutuelle aboutissant à la constitution de groupes d'individus au sein d'un collectif ? En d'autres termes, peut-on expliquer ces comportements individuels d'achat, de vote, d'adhésion à un groupe en identifiant les déterminants des choix effectués afin d'en expliciter la rationalité ?

L'analyse des préférences exprimées par un groupe de sujets relativement à un ensemble d'objets a pour objectif de mettre en évidence les caractéristiques des objets corrélées avec le choix des sujets en réponse au stimulus que constitue la situation d'achat, de vote ou d'adhésion. Comment faire pour révéler la structure qui se cache derrière les disparités de comportement observées dans l'expression de ces préférences ?

En analyse des données, il existe deux façons de répondre à ces questions. La première est la voie proposée par l'ensemble des techniques factorielles qui cherchent à construire des échelles objectives correspondant aux choix implicites des individus en analysant directement les tableaux multidimensionnels de données individuelles par des méthodes telles que l'analyse en composantes principales ou l'analyse des correspondances.

La seconde est d'analyser les **proximités** entre individus, ressemblances ou dissemblances résultant de l'observation des comportements, qu'elles soient relevées directement à l'issue de l'expérimentation ou qu'elles soient calculées sur la base des tableaux multidimensionnels de données individuelles au moyen d'indices de **similarité** (plus le nombre est grand plus les objets sont semblables) ou de **dissimilarité** (plus le nombre est grand plus les objets sont dissemblables).

D'un usage moins répandu que l'analyse factorielle chez les statisticiens francophones, le positionnement multidimensionnel constitue pourtant l'une des techniques fondamentales utilisées en analyse des préférences sur l'ensemble des domaines d'application de la psychométrie, en particulier dans la conception et le marketing des produits agro-alimentaires.
À partir de la matrice des similarités ou dissimilarités interindividuelles ainsi obtenue, le positionnement multidimensionnel permet d'obtenir une représentation géométrique s'ajustant au mieux selon un critère donné à l'ensemble des proximités observées et d'en proposer une interprétation révélée par la structure du nuage des points représentant les stimuli projetés dans un **espace euclidien**. Si l'ajustement de cette représentation euclidienne aux dissimilarités observées se fait d'après une procédure qui respecte les écarts entre dissimilarités, on parle de **modèle métrique**. Si le mode d'ajustement respecte simplement l'ordre entre les dissimilarités, on parle alors de **modèle non métrique**.

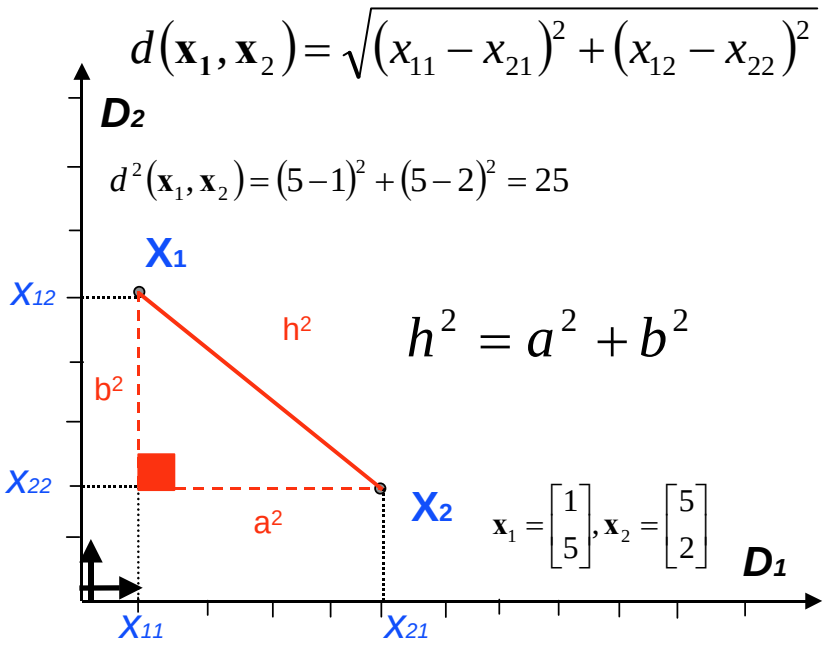


Figure 1 : géométrie euclidienne en dimension 2.

Figure2 : une vérification graphique du théorème de Pythagore.

Classiquement, **similarités** ou **dissimilarités** sont obtenues en demandant directement aux **sujets** d'estimer ou de classer ressemblances ou dissemblances entre paires d'**objets**. En marketing, on peut vouloir analyser simultanément les préférences d'un panel de consommateurs pour un ensemble déterminé de produits. Les produits agro-alimentaires sont alors considérés en tant qu'objets ou stimuli de l'expérimentation, et les panélistes sont les sujets ou experts participants à l'expérimentation (pour un exemple dans le domaine oenologique, cf. [Pagès 2003]). De même dans le domaine financier ou politique, on peut vouloir synthétiser les jugements émis par différents experts sur les risques présentés un ensemble d'opérateurs financiers ou de pays : il y aura ainsi plusieurs matrices de similarités ou de dissimilarités à examiner.
Les similarités ou dissimilarités peuvent également être issues de mesures objectives telles que la distance à vol d'oiseau entre deux villes ou de dénombrements factuels effectués selon une nomenclature ou une typologie fixée a priori au sein d'une population donnée ou d'un groupe d'espèces pour certaines unités spatiales ou territoriales. Enfin, on peut calculer similarités ou dissimilarités au sein de la population étudiée sur la base d'indices statistiques basés sur des données multivariées issues de mesures, d'enquêtes ou de recensements. Par exemple, les disparités entre régions ou pays d'Europe peuvent être calculées chaque année pour un tableau de bord donné (batterie fixe d'indicateurs statistiques). Les pays ou régions figurent alors en tant qu'objets ou stimuli dont les proximités ou disparités peuvent alors être représentées par une image euclidienne bi ou tridimensionnelle, susceptible de varier au cours du temps. Dans un tel contexte d'application, les années d'observation peuvent également tenir lieu d'individus ou de sujets.
Comme pour bien d'autres méthodes exploratoires, la **nature des données** conduit à privilégier l'utilisation de certaines variantes parmi l'ensemble des modèles de positionnement multidimensionnel disponibles. De ce point de vue, on distingue essentiellement trois types d'échelles de mesure : ordinal, intervalle ou ratio, les exemples de dissimilarités basées sur des données nominales étant rarement traités dans le cadre méthodologique du positionnement multidimensionnel. Ces trois niveaux de mesure définissent en fait deux types de dissimilarités et d'analyse. Le niveau ordinal conduit à des dissimilarités qualitatives et des analyses non métriques. Les échelles à intervalle ou de ratio aboutissent à des dissimilarités quantitatives et à des analyses métriques ou non métriques.
Le format des données doit également être pris en compte dans la spécification de la procédure de traitement des données. Les données peuvent être de format symétrique ou asymétrique pour des matrices de similarités ou de dissimilarités (lignes et colonnes de la matrice se réfère au même ensemble *I* d'objets), ou bien de format rectangulaire (les dissimilarités sont définies pour les objets de l'ensemble *I* des lignes relativement aux critères de l'ensemble *J* des colonnes).

Portrait de Pythagore de Samos, « Images of Mathematicians on Postage Stamps » Jeff Miller (Webring « Mathematics on Stamps »).

2. Introduction aux concepts du positionnement multidimensionnel

Les techniques de positionnement multidimensionnel trouvent leur origine dans les études psychométriques [Richardson, 1938] visant à comprendre comment les individus tissent des associations entre objets pour effectuer des regroupements, des classifications. Cependant, des études antérieures en biologie comparée, classant les espèces sur la base des réactions sériques interspécifiques [Boyden, 1933], utilisent sans la nommer une construction géométrique s'apparentant au positionnement multidimensionnel. Depuis lors, le positionnement multidimensionnel est devenu une technique de représentation géométrique largement utilisée dans des domaines aussi divers que le marketing, la sociologie électorale, ou plus récemment l'analyse sensorielle.

L'idée fondamentale du **positionnement multidimensionnel** est de représenter chaque objet ou stimulus dans un **espace euclidien**, habituellement bi ou tridimensionnel, de telle sorte que deux objets semblables soient représentés par deux points proches l'un de l'autre, et un couple dissemblable par des points éloignés.

Le problème mathématique fondamental que pose la réalisation concrète de cette idée est de trouver une représentation euclidienne d'un ensemble de points à partir de leurs distances respectives. Ce problème a été résolu en 1938 par Young et Householder au moyen d'une méthode très proche de l'analyse factorielle, faisant appel à la décomposition spectrale d'une matrice symétrique (cf. infra partie 4, encadré).

Le choix du mode d'ajustement des proximités observées $\delta_{ii'} = \delta(i,i')$ aux distances issues de la représentation euclidienne $d_{ii'} = d(i,i')$ par les **disparités** $\hat{d}_{ii'}$, définies par la relation f telle que $f(\delta_{ii'}) = \hat{d}_{ii'}$ appelée **fonction de représentation**, constitue une des spécifications essentielles du modèle de positionnement multidimensionnel.

Si cette relation f, est une fonction linéaire, soit :

- $f(\delta_{ii'}) = b\delta_{ii'}$ le modèle sans constante proposé initialement par Richardson en 1938 ; ce modèle est adapté aux échelles de mesure de type ratio ;
- $f(\delta_{ii'}) = a + b\delta_{ii'}$ le modèle avec constante proposé par Torgerson en 1952 ; ce modèle est adapté aux échelles de mesure de type intervalle ;

alors le modèle de positionnement multidimensionnel est qualifié de **métrique**.

Sinon, on peut simplement imposer que la relation f soit une transformation monotone (croissante ou décroissante) afin de respecter les proximités de classement des similarités ou des dissimilarités. En outre, on peut préciser de façon plus ou moins restrictive le caractère monotone (**monotonicité**) de la relation, soit respectivement :

- $\delta_{ii'} \prec \delta_{jj'} \Rightarrow \hat{d}_{ii'} = f(\delta_{ii'}) \prec f(\delta_{jj'}) = \hat{d}_{jj'}$ (monotonicité forte) ;
- $\delta_{ii'} \prec \delta_{jj'} \Rightarrow \hat{d}_{ii'} = f(\delta_{ii'}) \leq f(\delta_{jj'}) = \hat{d}_{jj'}$ (monotonicité faible) ;

(les symboles $\prec$ et $\leq$ désignent ici une relation d'ordre stricte, respectivement large).

Dans ce cas, le positionnement multidimensionnel est qualifié de **non-métrique**. Le positionnement multidimensionnel est également non-métrique si la représentation fournie ne précise que le rang et non la position des stimuli pour chaque dimension.

Des spécifications complémentaires peuvent être introduites afin de pouvoir représenter une configuration euclidienne qui soit un compromis entre les jugements individuels exprimés par différents experts, par exemple, en termes de dissimilarités : le modèle appelé **INDSCAL** (pour *Individual Differences Scaling*) introduit des pondérations distinctes selon les individus sujets de l'expérience sur chacune des dimensions de l'espace euclidien.

En fonction du nombre de matrices de dissimilarités, du niveau de mesure des données et du type de représentation choisi, on est amené à choisir un modèle spécifique parmi un ensemble de techniques de représentation géométrique regroupées sous l'appellation générique de positionnement multidimensionnel. La terminologie utilisée permet de distinguer parmi ces techniques quatre

grandes familles de modèles selon le choix de l'espace de représentation et le nombre de matrices de dissimilarités à représenter :

- le modèle de positionnement multidimensionnel classique (**PMC**) utilisant un espace euclidien pour représenter une matrice unique de dissimilarités ;
- le modèle de positionnement multidimensionnel répété (**PMR**) proposant une représentation unique de plusieurs matrices de dissimilarités dans un espace euclidien ;
- le modèle de positionnement multidimensionnel pondéré (**PMP)**, popularisé sous le label *INDSCAL*, utilisant des poids pour représenter des matrices de dissimilarités distinctes dans un espace euclidien ;
- le modèle de positionnement multidimensionnel généralisé (**PMG**) visant également à représenter plusieurs matrices de dissimilarités dans un espace euclidien.

Certaines distinctions sont également introduites par la terminologie suivant la nature des dissimilarités à représenter : si les dissimilarités sont exprimées selon une échelle ordinale (rangs ou classements), les modèles de positionnement multidimensionnels utilisés sont qualifiés de **non-métriques** ; s'il s'agit d'une échelle de mesure (intervalles, ratios), alors les modèles de positionnement multidimensionnel sont qualifiés de **métriques**.

La procédure **PROXIMITIES** de *SPSS* permet de calculer les écarts ou les proximités entre individus sur la base d'indices de similarité ou de dissimilarité. La procédure **ALSCAL** de *SPSS* permet de construire des structures géométriques multidimensionnelles, le plus souvent bi ou tridimensionnelles, s'ajustant au mieux aux similarités ou dissimilarités observées ou calculées. L'algorithme de positionnement multidimensionnel utilisé par ALSCAL est directement issu des travaux de Forrest W. Young, professeur émérite de psychométrie à l'Université de Caroline du Nord à Chappel Hill [Young, Takane et Lewyckyj, 1978].

Portrait de Forrest W. Young
par Patricia M. Young

3. Un exemple démonstratif : la topographie des parcours routiers

L'objectif du positionnement multidimensionnel est de construire une représentation d'un **ensemble *I*** d'objets telle que les positions relatives de ces objets représentés traduisent les dissemblances ou ressemblances existant entre ces objets en termes d'éloignement ou respectivement de proximité. Ce problème de représentation est très proche de celui auquel est confronté l'arpenteur qui doit déduire une carte topographique des différents relevés de distance effectués entre des lieux distincts. Les cartes routières où figure souvent en annexe une table des distances à parcourir entre les différentes villes de la région décrite peuvent constituer un premier exemple susceptible d'éclairer cette analogie. En utilisant une carte routière de la France, on peut établir le relevé des distances kilométriques à parcourir pour joindre par la route les principales métropoles régionales (cf. tableau 1). Les villes figurent les objets et les relevés kilométriques mesurent les dissimilarités entre objets.

La distance par la route pour aller d'Angers (ligne v02) à Amiens (colonne V01) est de 399 kilomètres (contenu de la case [2 , 1]). Cette distance est la même que pour aller d'Amiens (ligne v01) à Angers (colonne V02), soit 399 km (case [1 , 2]). Cette observation résulte de la **propriété de symétrie** de l'indice de dissimilarité *d* utilisé :

$$d(i,i') = d(i',i) \qquad [1]$$

i figurant l'index des lignes et *i'* celui des colonnes. Elle se traduit par la symétrie lignes/colonnes constatée dans ce tableau : le contenu de chaque case [*i*, *i'*] est égal au contenu de la case symétrique [*i'*, *i*]. Cette propriété caractérise également les indices de similarité entre objets

$$s(i,i') = s(i',i) \qquad [1']$$

En outre, on constate que les distances routières entre villes figurant dans ce tableau sont des nombres strictement positifs à l'exception des cases[*i*, *i*] de la diagonale dont le contenu est nul (case [*2* , *2*], la distance d'Angers à Angers est égale à 0). Cela traduit la **propriété d'identifiabilité** de l'indice de dissimilarité *s*:

$$d(i,i') = 0 \quad \Leftrightarrow \quad i = i' \qquad [2]$$

Les indices de similarités partagent une propriété similaire :

$$s(i,i') = Max \quad \Leftrightarrow \quad i = i' \qquad [2']$$

la borne maximum *Max* pouvant être égale par exemple à 1 ou bien à l'infini selon la définition de l'indice retenu.

Une dernière constatation moins élémentaire peut être faite en examinant le tableau plus soigneusement : la distance kilométrique du trajet direct Angers-Paris est inférieure à n'importe laquelle des distances parcourues sur un itinéraire indirect passant par une ville-tiers. Prenons le trajet indirect Angers-Tours-Paris comme exemple : Angers-Tours – 106 km, plus Tours-Paris – 234 km, soit 340 km distance supérieure à celle parcourue lors d'un trajet direct Angers-Paris, 303 km). Ceci reste vrai quel que soit le couple de villes à relier. Cette observation empirique « la ligne droite est le plus court chemin pour aller d'un point à un autre » traduit une propriété géométrique caractéristique des espaces métriques, dénommée **l'inégalité triangulaire** :

$$d(i,i') \le d(i,i'') + d(i'',i') \qquad \forall\, i, i', i'' \in I \qquad [3]$$

Tableau 1: la France en automobile	Villes	Amiens	Angers	Biarritz	Bordeaux	Brest	Calais	Cherbourg	Clermond-Ferrand	Dijon	Genève	Grenoble	Le Havre	Lille	Lyon	Marseille	Montpellier	Nancy	Nantes	Nice	Paris	Perpignan	Reims	Rennes	Rouen	Saint-Etienne	Strasbourg	Toulouse	Tours	Vichy
Villes		V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	V29
Amiens	v01	0	399	862	679	619	159	366	524	417	614	674	180	113	587	910	880	358	489	1067	130	1043	167	411	117	660	498	827	352	485
Angers	v02	399	0	518	335	371	494	290	402	498	636	644	297	503	538	842	692	602	90	999	303	762	454	126	282	530	742	552	106	399
Biarritz	v03	862	518	0	183	817	997	808	555	815	889	820	800	975	735	697	534	1001	512	879	744	445	885	618	785	705	1141	308	510	609
Bordeaux	v04	679	335	183	0	634	814	625	369	632	713	655	617	792	549	662	499	818	329	844	561	460	702	435	602	515	958	250	327	442
Brest	v05	619	371	817	634	0	714	402	752	812	988	996	468	723	890	1194	1048	886	305	1351	581	1094	734	245	502	900	1026	884	456	761
Calais	v06	159	494	997	814	714	0	461	672	556	753	840	275	109	753	1076	1046	469	584	1233	274	1209	278	506	212	800	609	967	487	633
Cherbourg	v07	366	290	808	625	402	461	0	647	661	858	874	215	470	768	1072	998	655	310	1229	340	1068	481	204	249	860	795	858	351	639
Clermond-Ferrand	v08	524	402	555	369	752	672	647	0	280	334	286	586	613	180	463	367	472	469	620	282	464	463	507	498	145	574	389	296	59
Dijon	v09	417	498	815	632	812	556	661	280	0	197	284	542	471	197	520	490	192	588	677	323	653	278	567	454	240	309	664	392	220
Genève	v10	614	636	882	713	988	753	858	334	197	0	147	739	668	154	435	405	355	726	484	520	568	475	743	651	210	382	654	530	295
Grenoble	v11	674	644	820	655	996	840	874	286	284	147	0	776	755	106	277	286	476	713	337	557	449	562	751	688	140	505	535	538	265
Le Havre	v12	180	297	800	617	468	275	215	586	542	739	776	0	284	676	980	950	511	357	1137	211	1062	320	269	88	710	651	852	290	547
Lille	v13	113	503	975	792	723	109	470	613	471	668	755	284	0	668	991	961	384	593	1148	223	1124	193	515	221	736	524	940	465	574
Lyon	v14	587	538	735	549	890	753	768	180	197	154	106	676	668	0	323	293	389	607	480	470	456	475	645	588	57	434	542	432	159
Marseille	v15	910	842	697	662	1194	1076	1072	463	520	435	277	980	991	323	0	163	712	911	197	793	326	798	949	892	330	757	412	736	457
Montpellier	v16	880	692	534	499	1048	1046	998	367	490	405	286	950	961	293	163	0	682	743	345	763	163	768	818	862	323	727	249	647	427
Nancy	v17	358	602	1001	818	886	469	655	472	192	355	476	511	384	289	712	682	0	692	869	308	845	191	659	423	460	140	856	491	412
Nantes	v18	489	90	512	329	305	584	310	469	588	726	713	357	593	607	911	743	692	0	1068	393	789	544	106	372	607	832	579	196	466
Nice	v19	1067	999	879	844	1351	1233	1229	620	677	484	337	1137	1148	480	197	345	869	1068	0	950	508	955	1106	1049	427	842	594	893	614
Paris	v20	130	303	744	561	581	274	340	382	323	520	557	211	223	470	793	763	308	393	950	0	926	155	360	123	517	448	681	234	343
Perpignan	v21	1043	762	445	460	1094	1209	1068	464	653	568	499	1062	1124	456	326	163	845	789	508	926	0	931	895	974	481	820	210	717	590
Reims	v22	167	454	885	702	734	278	481	463	278	475	562	320	193	475	798	768	191	544	955	155	931	0	511	232	537	331	806	375	424
Rennes	v23	411	126	618	435	285	506	204	507	567	743	751	269	515	645	949	818	659	106	1106	360	895	511	0	294	712	799	685	211	516
Rouen	v24	117	282	785	602	502	212	249	498	454	651	688	88	221	588	892	862	423	372	1049	123	974	232	294	0	640	563	764	275	459
Saint-Etienne	v25	660	530	705	515	900	800	860	145	240	210	140	710	736	57	330	323	460	607	427	517	481	537	712	640	0	705	545	472	145
Strasbourg	v26	498	742	1141	958	1026	609	795	574	309	382	505	651	524	434	757	727	140	832	842	448	890	331	299	563	705	0	976	631	511
Toulouse	v27	827	552	308	250	884	967	858	389	664	654	535	852	940	542	412	249	856	579	594	681	210	806	685	764	545	976	0	507	443
Tours	v28	352	106	510	327	456	487	351	296	392	530	538	290	465	432	736	647	491	196	893	234	717	375	211	275	472	631	507	0	293
Vichy	v29	485	399	609	442	761	633	639	59	220	295	265	547	574	159	457	427	412	466	614	343	590	424	516	459	145	514	443	293	0

Muni de ces trois propriétés, l'indice de dissimilarité *d* constitue alors une instance d'une classe particulière de fonctions du produit cartésien $I \times I$ (l'ensemble des couples $(i,i') \quad i,i' \in I$), à valeurs dans l'ensemble des réels positifs $\Re^+$ que l'on appelle une **distance** (entre les **objets *i***) ou encore une **métrique** (de leurs écarts). Les ensembles d'objets *I* muni d'une métrique sont appelés **espaces métriques.**

Si les objets sont représentés par des **vecteurs** d'observations, alors ces **espaces vectoriels** peuvent être munis d'une géométrie où les notions usuelles d'angle et d'orthogonalité se transcrivent en termes de **produit scalaire**. Si les observations sont des nombres réels, on peut construire une **représentation cartésienne** des objets dont les coordonnées sont définies par rapport à un **repère orthonormé**, c'est à dire relativement à une origine ***O*** et à des axes orthogonaux dont la direction et la métrique des écarts sont définies par des **vecteurs unitaires** (i.e. de longueur ou **norme** 1).

Dans ces espaces vectoriels, les **distances euclidiennes**, sont définies comme la racine carrée de la somme des carrés des écarts, terme à terme, entre coordonnées cartésiennes. Cette définition généralise à un espace multidimensionnel (i.e. à plusieurs variables - $\Re^p$) le calcul élémentaire effectué dans le plan ($\Re^2$) de la longueur de l'hypoténuse du triangle rectangle (cf. figure 1). Ainsi, les distances euclidiennes définissent des géométries vectorielles qui sont invariantes par translation, rotation ou **déplacement** (composée d'une translation et d'une rotation) et qui coïncident avec notre perception de l'espace à 3 dimensions. Un espace muni d'une distance euclidienne est appelé un **espace euclidien**.

A titre illustratif, on peut utiliser le principe du positionnement multidimensionnel pour obtenir une **représentation euclidienne** des parcours routiers entre les principales agglomérations françaises (cf . figure 3) où les villes se positionnent en fonction de leurs éloignements respectifs.

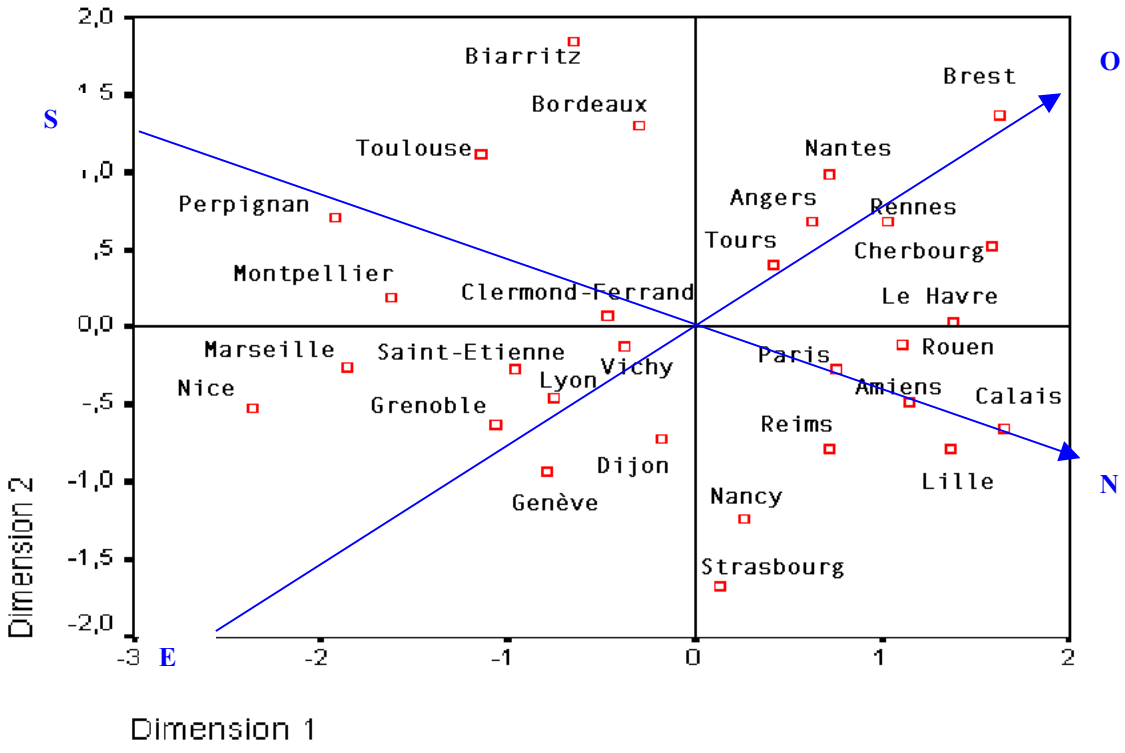


Figure 3 : topographie des parcours routiers entre les principales villes françaises.

Le premier axe (« Dimension 1 ») de la représentation est toujours celui correspondant aux écarts observés les plus importants en termes de disparités. Le second axe (« Dimension 2 ») est orienté selon une direction orthogonale au premier, correspondant aux écarts les plus importants parmi ceux qui ne relèvent pas du premier axe. Et ainsi de suite, l'ensemble des axes s'ordonne selon l'application de cette règle d'extraction.
Une rotation appropriée de ce système d'axes orthogonaux nous conduit à interpréter la première dimension comme un axe Nord-Sud et la seconde dimension comme un axe Est-Ouest. La topographie obtenue diffère sensiblement de la représentation cartographique qui nous est familière : les déformations sont principalement imputables au fait que le réseau routier dans certaines régions impose des trajets où la distance parcourue en automobile est notablement supérieure à la distance géographique (« à vol d'oiseau »).
Une telle représentation est obtenue par l'intermédiaire d'un algorithme itératif permettant de minimiser l'écart entre les distances qui se déduisent de la représentation euclidienne recherchée et les **disparités** $\hat{d}_{ii'}$, définies comme des fonctions des dissimilarités observées : $\hat{d}_{ii'} = f(\delta_{ii'})$, fonctions convenablement choisies pour respecter soit les écarts soit l'ordre entre les dissimilarités observées.
En effet, **l'analyse des proximités** mis au point par Shepard en 1962 a montré qu'il est possible d'obtenir des représentations métriques ajustées à partir de relations ordinales entre distances et proximités. Cette méthode est généralisée par Kruskal qui propose en 1964 une approche en termes d'optimisation construite autour de la définition d'un critère d'ajustement.

Définition de la fonction objectif

Afin de pouvoir conduire l'estimation selon une procédure d'optimisation, Kruskal propose de définir la **fonction objectif** qui constitue le critère d'ajustement à minimiser. Il s'agit d'une procédure des moindres carrés où l'on cherche à minimiser la somme des carrés des écarts entre les distances issues de la représentation euclidienne et les disparités, fonction des proximités observées : $\sum_{i=1}^{n}\sum_{i'=1}^{n}(d_{ii'} - f[\delta_{ii'}])^2$

Pour assurer la comparabilité entre des ensembles de stimuli distincts, ces écarts peuvent être ramenés à un facteur d'échelle commun : $\sum_{i=1}^{n}\sum_{i'=1}^{n} {d_{ii'}}^2$, ce qui conduit au ratio suivant : $\dfrac{\sum_{i=1}^{n}\sum_{i'=1}^{n}(d_{ii'} - f[\delta_{ii'}])^2}{\sum_{i=1}^{n}\sum_{i'=1}^{n} {d_{ii'}}^2}$ et en termes d'optimisation à une première définition du critère d'ajustement de la fonction de représentation *f* appelé le **f-stress** :

$$f - stress = \sqrt{\frac{\sum_{i=1}^{n}\sum_{i'=1}^{n}(d_{ii'} - f[\delta_{ii'}])^2}{\sum_{i=1}^{n}\sum_{i'=1}^{n} {d_{ii'}}^2}}$$

De par sa définition, le stress d'une fonction de représentation est positif ou nul. Plus le *f*-stress est élevé, moins la configuration spatiale *X*, définie par la matrice **X** des coordonnées des points de la solution étudiée, est adaptée à l'ensemble Δ des proximités observées.
Si le *f*-stress est nul, la configuration spatiale *X* s'adapte parfaitement à l'ensemble Δ des disparités, c'est à dire : $d_{ii'} = f(\delta_{ii'}) \quad \forall (i,i') \in I \times I$

Le **stress** d'une configuration spatiale *X*, pour un l'ensemble Δ de disparités est définie par l'optimun du critère d'ajustement pour l'ensemble des fonctions de représentation considérées : $stress\,(\Delta, X) = \min_{f} \{f - stress(\mathbf{\Delta}, \mathbf{X}, f)\}$.

La fonction objectif ainsi définie est appelée le **stress de type I**. Le stress est un indicateur normalisé variant entre 0 et 1, la valeur nulle indiquant un ajustement parfait.

Ainsi que le montre l'historique des itérations de l'algorithme du positionnement multidimensionnel pour la topographie des parcours routiers, la convergence est atteinte au bout de quatre itérations fixant la valeur finale du stress de la configuration à 0,06, valeur faible indiquant un très bon ajustement ;

```
Iteration history for the 2 dimensional solution (in squared distances)

                Young's S-stress formula 1 is used.
              Iteration     S-stress      Improvement
                  1           ,11701
                  2           ,08476         ,03225
                  3           ,08211         ,00265
                  4           ,08191         ,00020

                       Iterations stopped because
                S-stress improvement is less than   ,001000

           Stress and squared correlation (RSQ) in distances

RSQ values are the proportion of variance of the scaled data (disparities)
           in the partition (row, matrix, or entire data) which
            is accounted for by their corresponding distances.
             Stress values are Kruskal's stress formula 1.

                For  matrix
    Stress  =   ,06013      RSQ =  ,98008
_
```

Tableau 2 *: itérations de l'algorithme d'optimisation et valeurs du stress de type I pour la topographie des parcours routiers.*

La qualité de cet ajustement peut être précisée grâce à un indicateur calculé pour la configuration finale, le R^2, coefficient de corrélation au carré (« **squared correlation (RSQ)** ») entre les distances et les disparités. Ce coefficient s'interprète en termes de pourcentage de variabilité expliquée. Dans cet exemple, 98 % de la variabilité des distances issues de la configuration euclidienne est expliquée par les disparités, qui en raison du modèle choisi (modèle euclidien respectant les écarts), sont des fonctions linéaires affines des dissimilarités observées, soit : $f(\delta_{ii'}) = b + a\delta_{ii'}$.

Le diagramme de Shepard constitue pour sa part un outil graphique de vérification de la qualité de cet ajustement en permettant de visualiser la relation entre distances et disparités sous la forme d'un diagramme bi-dimensionnel :

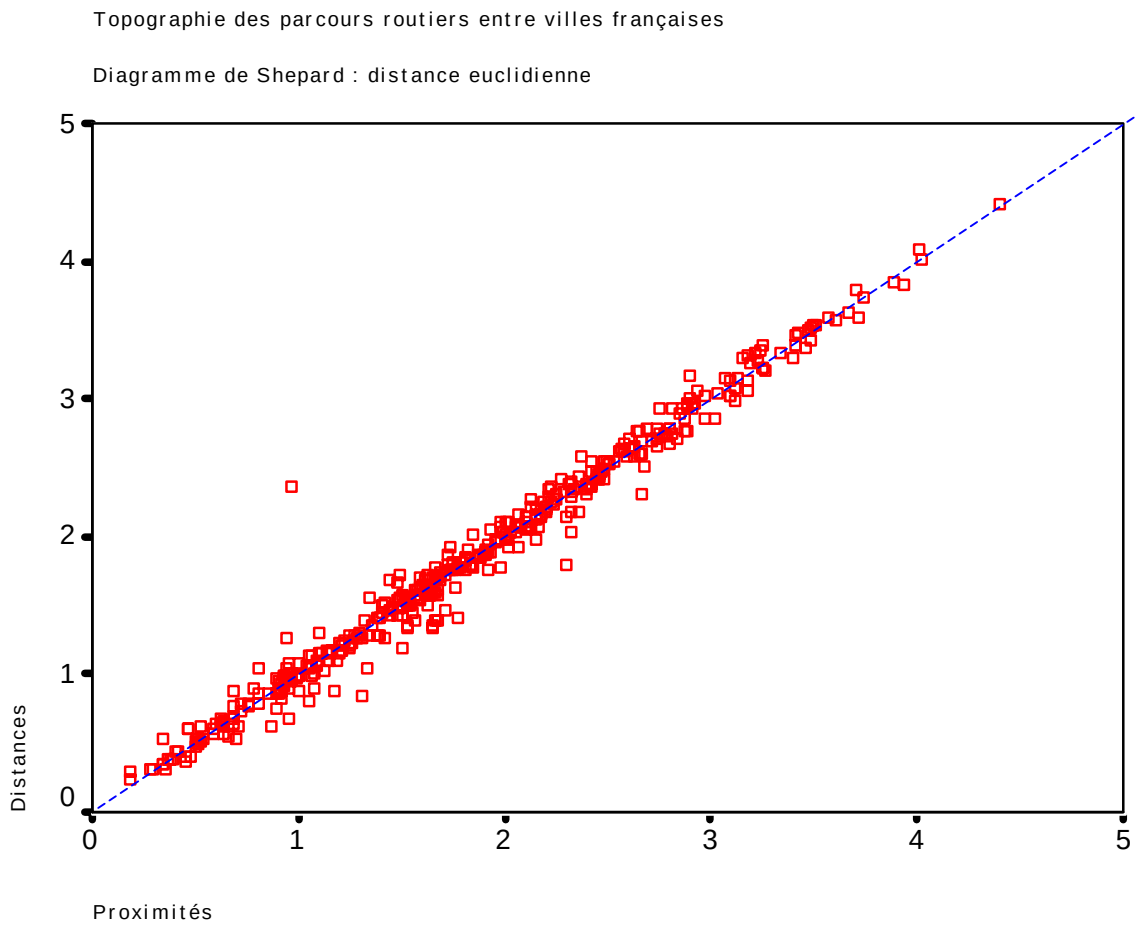


Figure 4 : le diagramme de Shepard permettant de visualiser la relation entre distances et disparités.

Le diagramme de Shepard de la figure 4 montre que la nature de la relation entre distances et disparités est linéaire et croissante.

Ile de Samos, patrie de Pythagore, « Images of Mathematicians on Postage Stamps » Jeff Miller (Webring « Mathematics on Stamps »).

4. Le modèle classique, analyse factorielle sur tableau de distances

Cette présentation théorique du modèle classique n'est pas indispensable à la compréhension globale de l'exposé en première lecture pour un lecteur principalement intéressé par les applications de la méthode. La structure et les notations de la section sont reprises de l'ouvrage [Cox et Cox , 2001]
Le modèle classique du positionnement multidimensionnel (« *classical multidimensional scaling* ») a été initialement développé par Torgerson en 1952 comme outil d'analyse des résultats d'expérimentation à caractère psychométrique (cf. par exemple, l'étude en 1957 de Rothkopf sur la confusion des signaux Morse), appliquant la méthode de Young et Householder (1938) fondée sur la décomposition spectrale d'une matrice symétrique à termes réels pour trouver une représentation euclidienne d'un ensemble de points à partir de leurs distances respectives. Le modèle classique du positionnement multidimensionnel est présenté par [Cailliez et Pages 1976] sous le terme francophone d'analyse factorielle sur tableau de distances (AFTD)

Décomposition spectrale d'une matrice symétrique.
Soit $\mathbf{A}$, une matrice carrée $(n \times n)$, symétrique à termes réels ($a_{ii'} \in \Re$).
Cette matrice est diagonalisable.
Désignons l'ensemble des valeurs propres par $\{\lambda_i ; i = 1, \cdots, n\}$ dont les vecteurs propres associées $\{\mathbf{v}_i ; i = 1, \cdots, n\}$ sont orthonormés ($\mathbf{v}_i{}^t \times \mathbf{v}_{i'} = 0 \quad \forall\, i \neq i'$ et $\mathbf{v}_i{}^t \times \mathbf{v}_i = 1 \quad \forall\, i$).
Alors, la matrice $\mathbf{A}$ peut s'écrire sous la forme : $\mathbf{A} = \mathbf{V}\ \mathbf{\Lambda}\ \mathbf{V}^t = \sum_{i=1}^{n} \lambda_i \bullet \left(\mathbf{v}_i \otimes \mathbf{v}_i{}^t\right)$

avec $\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & 0 & \vdots \\ \vdots & \ddots & \lambda_i & \ddots & \vdots \\ \vdots & 0 & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \lambda_n \end{bmatrix} = \mathbf{diag}\left(\lambda_1, \cdots, \lambda_i, \cdots, \lambda_n\right)$;

$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_i & \cdots & \mathbf{v}_n \end{bmatrix}$ tel que $\mathbf{V}\ \mathbf{V}^t = \mathbf{V}^t\ \mathbf{V} = \mathbf{I}_{n \times n}$ (matrice orthonormale) ;
et $\otimes$ symbolisant le produit de Kronecker de deux matrices.

4.1. Représentation euclidienne d'un tableau de distances

Soit un ensemble de *n* objets à représenter dans un espace euclidien de dimension *p*. Chaque objet *i* est représenté par un point construit à partir du vecteur de coordonnées

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{ip} \end{bmatrix} = \begin{bmatrix} x_{i1}, \cdots, x_{ij}, \cdots, x_{ip} \end{bmatrix}^t \text{ avec } \left\langle \mathbf{x_i}, \mathbf{x}_{i'} \right\rangle = \begin{bmatrix} x_{i1}, \cdots, x_{ij} \cdots, x_{ip} \end{bmatrix} \times \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{ip} \end{bmatrix} = \sum_{j=1}^{p} x_{ij} x_{i'j} = \mathbf{x_{i'}}{}^t \times \mathbf{x_i}$$

définissant le produit scalaire entre les deux vecteurs $\mathbf{x}_i$ et $\mathbf{x}_{i'}$.

La distance euclidienne entre deux points *i* et *i'* est donnée par la norme de la somme des carrés des écarts :

$$d^2(i,i') = \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2 = \left(\mathbf{x_i} - \mathbf{x_{i'}}\right)^t \times \left(\mathbf{x_i} - \mathbf{x_{i'}}\right) = \left\|\mathbf{x_i} - \mathbf{x_{i'}}\right\|^2 \qquad [4]$$

Soit la matrice **B** des produits scalaires définie par $b_{ii'} = \mathbf{x}_i^{\,t} \times \mathbf{x}_{i'}$, la procédure de positionnement multidimensionnel simple consiste à calculer **B** à partir du carré des distances observées puis à extraire de la matrice **B** les coordonnées $\mathbf{x}_i$ du point *i*, solutions du problème posé.

4.2. Calcul de la matrice des produits scalaires

La solution étant déterminée à une translation près, on choisit de situer l'origine du repère euclidien au barycentre du nuage des points. Dans ce repère affine, les coordonnées des points sont centrées : $\sum_{i=1}^{n} x_{ij} = 0$.

Le produit scalaire étant une forme bilinéaire symétrique, en développant l'équation [4], on trouve : $d^2(i,i') = \mathbf{x}_i^{\,t} \times \mathbf{x}_i + \mathbf{x}_{i'}^{\,t} \times \mathbf{x}_{i'} - 2\mathbf{x}_i^{\,t} \times \mathbf{x}_{i'}$ [5]

(en définissant le produit scalaire dans le plan $\Re^2$ avec le cosinus de l'angle entre deux vecteurs, soit : $u \times v = \|u\| \, \|v\| \cos\theta$, on retrouve le calcul de la longueur d'un côté du triangle à partir de celles des deux autres côtés et du cosinus de l'angle opposé).

En sommant respectivement sur chacun des indices *i* et *i'* et en utilisant le fait que les coordonnées sont centrées ($\sum_{i=1}^{n} \mathbf{x}_i^{\,t} \times \mathbf{x}_{i'} = \sum_{i'=1}^{n} \mathbf{x}_i^{\,t} \times \mathbf{x}_{i'} = 0$), on obtient :

- $\sum_{i=1}^{n} d^2(i,i') = \sum_{i=1}^{n} \mathbf{x}_i^{\,t} \times \mathbf{x}_i + n\left(\mathbf{x}_{i'}^{\,t} \times \mathbf{x}_{i'}\right)$ d'où $\mathbf{x}_{i''}^{\,t} \times \mathbf{x}_{i''} = \frac{1}{n}\sum_{i=1}^{n} d^2(i,i') - \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^{\,t} \times \mathbf{x}_i$ [6]
- $\sum_{i'=1}^{n} d^2(i,i') = \sum_{i'=1}^{n} \mathbf{x}_{i'}^{\,t} \times \mathbf{x}_{i'} + n\left(\mathbf{x}_i^{\,t} \times \mathbf{x}_i\right)$ d'où $\mathbf{x}_i^{\,t} \times \mathbf{x}_i = \frac{1}{n}\sum_{i'=1}^{n} d^2(i,i') - \frac{1}{n}\sum_{i'=1}^{n} \mathbf{x}_{i'}^{\,t} \times \mathbf{x}_{i'}$ [6']

En sommant [6'] sur l'indice *i*, on obtient :

- $\sum_{i=1}^{n}\sum_{i'=1}^{n} d^2(i,i') = n\sum_{i'=1}^{n} \mathbf{x}_{i'}^{\,t} \times \mathbf{x}_{i'} + n\sum_{i=1}^{n} \mathbf{x}_i^{\,t} \times \mathbf{x}_i = 2n\sum_{i=1}^{n} \mathbf{x}_i^{\,t} \times \mathbf{x}_i$ [7]

Pour chaque élément $b_{ii'} = \mathbf{x}_i^{\,t} \times \mathbf{x}_{i'}$ de la matrice des produits scalaires, on obtient en substituant d'après [6] et [6'] :

$$b_{ii'} = -\frac{1}{2}\left[d^2(i,i') - \mathbf{x}_i^{\,t} \times \mathbf{x}_i - \mathbf{x}_{i'}^{\,t} \times \mathbf{x}_{i'}\right] = -\frac{1}{2}\left[d^2(i,i') - \frac{1}{n}\sum_{i'=1}^{n} d^2(i,i') - \frac{1}{n}\sum_{i=1}^{n} d^2(i,i') + \frac{2}{n}\sum_{i=1}^{n} \mathbf{x}_i^{\,t} \times \mathbf{x}_i\right]$$

et d'après [7] :

$$b_{ii'} = -\frac{1}{2}\left[d^2(i,i') - \frac{1}{n}\sum_{i'=1}^{n} d^2(i,i') - \frac{1}{n}\sum_{i=1}^{n} d^2(i,i') + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n} d^2(i,i')\right]$$

En posant :

$$a_{ii'} = -\frac{1}{2}d^2(i,i') \qquad a_{i..} = \frac{1}{n}\sum_{i'=1}^{n} a_{i,i'} \qquad a_{.i'} = \frac{1}{n}\sum_{i=1}^{n} a_{i,i'} \quad a_{..} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n} a_{i,i'}$$

il est possible d'exprimer la matrice $\mathbf{B}$ des produits scalaires en fonction de la matrice $\mathbf{A}$, définie par $[\mathbf{A}]_{i,i'} = a_{i,i'} = -\frac{1}{2}d^2(i,i')$, sous la forme du produit matriciel suivant :

$$\boxed{\mathbf{B} = \mathbf{H}\,\mathbf{A}\,\mathbf{H}} \qquad [8]$$

où $\mathbf{H}$ est l'opérateur de centrage défini par $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\otimes\mathbf{1}^t$,

$\mathbf{1}$ étant le vecteur constant $\mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{bmatrix}$ de dimension *n*,

et $\mathbf{1}\otimes\mathbf{1}^t$, la matrice carrée de dimension (*n*x*n*) $\quad \mathbf{1}\otimes\mathbf{1}^t = \begin{bmatrix} 1 & \cdots & 1 & \cdots & 1 \\ \vdots & \ddots & & & \vdots \\ 1 & & 1 & & 1 \\ \vdots & & & \ddots & \vdots \\ 1 & \cdots & 1 & \cdots & 1 \end{bmatrix}$.

La matrice $\mathbf{B}$ des produits scalaires s'obtient donc, à un facteur $-\frac{1}{2}$ près, à partir de la matrice des distances au moyen d'une double opération de centrage, effectuée d'une part sur les lignes et d'autre part sur les colonnes de cette matrice, qui consiste à enlever la moyenne des lignes et la moyenne des colonnes, puis à rajouter la moyenne générale du tableau.

4.3. Extraction des coordonnées

La matrice $\mathbf{B}$ des produits scalaires, s'exprime dans le repère cartésien barycentrique comme le produit matriciel du tableau (*n*x*p*) des coordonnées centrées $\mathbf{X}$ par son transposé $\mathbf{X}^t$:

$$\mathbf{B} = \mathbf{X}\mathbf{X}^t$$

La matrice $\mathbf{B}$ est symétrique, semi-définie positive et de rang *p* :

$$p = r(\mathbf{B}) = r(\mathbf{X}\mathbf{X}^t) = r(\mathbf{X})$$

Elle possède donc *p* valeurs propres λ_j non nulles et $n - p$ valeurs propres nulles et peut être exprimée dans les termes de sa décomposition spectrale :

$$\mathbf{B} = \mathbf{V}\,\mathbf{\Lambda}\,\mathbf{V}^t$$

où $\mathbf{\Lambda}$ est la forme diagonale de la matrice dans la base des vecteurs propres, avec :

$\mathbf{\Lambda} = \mathbf{diag}(\lambda_1,\cdots,\lambda_j,\cdots,\lambda_p)$ et $\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_j & \cdots & \mathbf{v}_p \end{bmatrix}$, la matrice de passage dont les colonnes sont les vecteurs propres associés.

Par identification, on en déduit la matrice $\mathbf{X}$ des coordonnées centrées :

$$\mathbf{X} = \mathbf{V}\,\mathbf{\Lambda}^{1/2}$$

avec $\mathbf{\Lambda}^{1/2} = \mathbf{diag}\left(\sqrt{\lambda_1},\cdots,\sqrt{\lambda_j},\cdots,\sqrt{\lambda_p}\right)$

4.4. Représentation euclidienne d'un tableau de dissimilarités

En pratique, l'utilisation du positionnement multidimensionnel s'effectue plus souvent à partir d'un tableau de dissimilarités $[\delta_{ii'}]_{I\times I}$ qu'à partir d'un tableau de distances $[d_{ii'}]_{I\times I}$.

Si la matrice **B** des produits scalaires, formée à partir de la matrice **A**, définie par $[\mathbf{A}]_{i,i'} = a_{i,i'} = -\frac{1}{2}\delta^2(i,i')$, est définie positive de rang *p*, alors on peut trouver une représentation euclidienne correspondante dans un espace de dimension *p*, et la distance $d(i,i')$, entre les points *i* et *i'* dans cet espace euclidien, est égale à la dissimilarité $\delta(i,i')$.

La réciproque est également vraie : une matrice **B** de produits scalaires formée à partir de distances euclidiennes dans un espace de dimension *p* est semi-définie positive.

Etant donnée une matrice **B** semi-définie positive, formée à partir d'un tableau de dissimilarités entre les *n* éléments d'un ensemble d'objets, il existe une représentation euclidienne de dimension $n-1$ telle que les distances entre les points soient égales aux dissimilarités entre objets

Si la matrice **B** n'est pas semi-définie positive, la plus simple des solutions envisageables est de se limiter au valeurs propres positives. Un autre type de solution, utilisé dans l'implantation de l'algorithme *ALSCAL* réalisée par *SPSS*, est d'ajouter une constante à l'ensemble des éléments du tableau des dissimilarités $\delta(i,i')$, à l'exception des termes diagonaux, pour former un nouveau système de dissimilarités $\delta^*(i,i')$:

$$\delta^*(i,i') = \delta(i,i') + c\left(1-\kappa^{ii'}\right)$$

avec $\kappa^{ii'} = \begin{cases} 1 \; si \; i = i' \\ 0 \; si \; i \neq i' \end{cases}$, le symbole de Kronecker

telle que la matrice **B** correspondante soit semi-définie positive.

Le problème de la détermination de la valeur minimale de cette constante est traité par [Cailliez, 1983] qui recommande plutôt la transformation suivante du tableau des dissimilarités : $\delta^*(i,i') = \sqrt{\delta^2(i,i') + c\left(1-\kappa^{ii'}\right)}$

4.5. Choix du nombre d'axes pour la représentation euclidienne

Si la matrice **B** est semi-définie positive, alors la dimension *p* de la représentation euclidienne est égale au nombre de valeurs propres non nulles. Si **B** n'est pas semi-définie positive, alors la dimension *p* de l'espace euclidien est égale au nombre de valeurs propres positives. Il s'agit de la dimension maximale de la solution euclidienne.

En pratique, le nombre *q* d'axes retenus pour la représentation euclidienne doit être plutôt faible. On utilise généralement les axes correspondant aux deux ou trois premières valeurs propres.

Le choix du nombre *q* optimal d'axes (le nombre de « dimensions ») à retenir pour la représentation euclidienne peut s'effectuer en tenant compte du pourcentage de variance expliquée. La somme des distances au carré entre les points est fonction de la trace de la matrice **B** (somme des valeurs propres).

Si **B** est semi-définie positive, la trace de **B** est égale à : $tr(\mathbf{B}) = \sum_{i=1}^{n-1}\lambda_i = \frac{1}{2n}\sum_{i=1}^{n}\sum_{i'=1}^{n} d_{ii'}^2$

La proportion de variance expliquée par un sous-espace de dimension *q* est alors égale à :

$$\tau(q) = \frac{\sum_{j=1}^{q} \lambda_j}{\sum_{i=1}^{n-1} \lambda_i}$$

Si **B** n'est pas semi-définie positive, soit on ne prend en compte dans le calcul que les valeurs propres positives (solution simplifiée), soit on choisit comme indicateur (addition d'une constante) :

$$\tau(q) = \frac{\sum_{j=1}^{q} \lambda^*_j}{\sum_{i=1}^{n-1} \left|\lambda^*_i\right|}$$

Dans l'exemple de la topographie des parcours routiers entre villes françaises, le pourcentage de variance expliquée des proximités par les distances issues de la représentation euclidienne est de 98 % (`"RSQ=0,98008"`, cf. tableau 2)

4.6. Algorithme du positionnement multidimensionnel classique

A partir de ce qui précède, l'algorithme du positionnement multidimensionnel classique peut se résumer ainsi :

i) lire la matrice des dissimilarités $\mathbf{\Delta} = \left[\delta_{ii'}\right]_{I \times I}$;

ii) calculer la matrice $\mathbf{A} = \left[-\frac{1}{2}\delta_{ii'}{}^2\right]_{I \times I}$;

iii) calculer la matrice $\mathbf{B} = \left[a_{ii'} - a_{i.} - a_{.i'} + a_{..}\right]_{I \times I}$;

iv) extraire les valeurs propres non nulles $\{\lambda_1 \ \cdots \ \lambda_i \ \cdots \ \lambda_{n-1}\}$ et les vecteurs propres associés $[\mathbf{v}_1 \ \cdots \ \mathbf{v}_i \ \cdots \ \mathbf{v}_{n-1}]$ sous la contrainte de normalisation $\mathbf{v}_i{}^t \times \mathbf{v}_i = \lambda_i$

s'il y a des valeurs propres négatives, soit :

a) les annuler (version simplifiée) et aller à l'étape *v)*

b) transformer les dissimilarités en ajoutant une constante de valeur minimale (e.g. fonction $\delta^*(i,i') = \sqrt{\delta^2(i,i') + c\left(1 - \kappa^{ii'}\right)}$) et retourner à l'étape *ii) ;*

v) choisir un nombre *q* adéquat de dimensions en utilisant le pourcentage de variance expliqué $\tau(q)$ comme critère ;

vi) extraire les coordonnées euclidiennes $x_{ij} = \sqrt{\lambda_j} v_{ij}$ des points *i* pour chaque dimension *j*.

La détermination de la valeur de la constante dans la procédure *SPSS* n'est pas explicitée mais, d'après la documentation disponible, l'ajout de la constante est effectué selon la fonction $\delta'(i,i') = \delta'(i,i') + c\left(1 - \kappa^{ii'}\right)$.

4.7. Lien avec l'analyse en composantes principales

En partant du tableau $\mathbf{X}$ des données centrées, on peut calculer la matrice de variance-covariance empirique : $\mathbf{S} = \frac{1}{(n-1)}\mathbf{X}^t\mathbf{X}$.

L'analyse en composantes principales (**ACP**) s'effectue en diagonalisant l'opérateur d'inertie $\mathbf{S} = \mathbf{\Psi}\ \mathbf{M}\ \mathbf{\Psi}^t$, avec $\mathbf{M} = \mathbf{diag}\left(\mu_1, \cdots, \mu_j, \cdots, \mu_p\right)$ ce qui permet d'extraire les valeurs propres $\left\{\mu_j,\ j = 1, \cdots, p\right\}$ et les vecteurs propres $\left\{\boldsymbol{\psi}_j,\ j = 1, \cdots, p\right\}$.

Les composantes principales sont alors obtenues par projection : $c_j = \boldsymbol{\psi}_j^t\mathbf{x}$.

Soit $\mathbf{B} = \mathbf{X}\mathbf{X}^t$, la matrice des produits scalaires. Les vecteurs propres de $\mathbf{B}$ sont définis par :

$$\mathbf{B}\mathbf{v}_j = \mathbf{X}\mathbf{X}^t\mathbf{v}_j = \lambda_j\mathbf{v}_j$$

En prémultipliant par $\mathbf{X}^t$, on obtient : $\mathbf{X}^t\mathbf{X}\mathbf{X}^t\mathbf{v}_j = \lambda_j\mathbf{X}^t\mathbf{v}_j$ or $\mathbf{X}^t\mathbf{X}\boldsymbol{\psi}_j = \mu_j\boldsymbol{\psi}_j$.

Par identification, on en déduit $\lambda_j = \mu_j$ et $\mathbf{v}_j = \boldsymbol{\psi}_j$.

Il y a donc équivalence entre le positionnement multidimensionnel classique (analyse des proximités avec une métrique euclidienne) et l'ACP. Les p coordonnées principales des stimuli obtenues en positionnement multidimensionnel classique sont simplement les p composantes principales des individus dans une ACP.

Portrait d'Euclide, « Images of Mathematicians on Postage Stamps » Jeff Miller (Webring « Mathematics on Stamps »).

5. Un exemple illustratif : les céréales du petit déjeuner

Cet exemple est extrait d'une comparaison des céréales du petit déjeuner présentée en 1993 à une réunion sur les graphiques en statistique, organisée par l'*ASA* (*American Statistical Association*). Cette expérimentation portait initialement sur 77 produits jugés selon 11 critères. L'extrait suivant concerne uniquement les céréales produites par un des fabricants, soit 23 produits distincts. Les critères étudiés sont le nombre de calories (<calories>), les teneurs en protéines (<protein>), en graisses (<fat>), en sodium (<sodium>), en fibres (<fiber>), en complexes carbohydratés (<carbo>), en sucres (<sugars>), le numéro d'étagère (<shelf>) compté à partir du sol, les teneurs en potassium (<potass>) et en vitamines (<vitamins>).

cereales.sav - Editeur de données SPSS

Fichier Edition Affichage Données Transformer Analyse Graphes Outils Fenêtre Aide

1 : name AllB

	name	calories	protein	fat	sodium	fiber	carbo	sugars	shelf	potass	vitamins
1	AllB	70,00	4,00	1,00	260,00	9,00	7,00	5,00	3,00	320,00	25,00
2	AllF	50,00	4,00	,00	140,00	14,00	8,00	,00	3,00	330,00	25,00
3	AppJ	110,00	2,00	,00	125,00	1,00	11,00	14,00	2,00	30,00	25,00
4	CorF	100,00	2,00	,00	290,00	1,00	21,00	2,00	1,00	35,00	25,00
5	CorP	110,00	1,00	,00	90,00	1,00	13,00	12,00	2,00	20,00	25,00
6	Crac	110,00	3,00	3,00	140,00	4,00	10,00	7,00	3,00	160,00	25,00
7	Cris	110,00	2,00	,00	220,00	1,00	21,00	3,00	3,00	30,00	25,00
8	FrM	100,00	3,00	,00	,00	3,00	14,00	7,00	2,00	100,00	25,00
9	FroF	110,00	1,00	,00	200,00	1,00	14,00	11,00	1,00	25,00	25,00
10	Froo	110,00	2,00	1,00	125,00	1,00	11,00	13,00	2,00	30,00	25,00
11	FruB	120,00	3,00	,00	240,00	5,00	14,00	12,00	3,00	190,00	25,00
12	JRCN	110,00	2,00	1,00	170,00	1,00	17,00	6,00	3,00	60,00	100,00
13	JRFN	140,00	3,00	1,00	170,00	2,00	20,00	9,00	3,00	95,00	100,00
14	MuC	160,00	3,00	2,00	150,00	3,00	17,00	13,00	3,00	160,00	25,00
15	NGA	140,00	3,00	2,00	220,00	3,00	21,00	7,00	3,00	130,00	25,00
16	Nut&	120,00	2,00	1,00	190,00	,00	15,00	9,00	2,00	40,00	25,00
17	NutW	90,00	3,00	,00	170,00	3,00	18,00	2,00	3,00	90,00	25,00
18	Prod	100,00	3,00	,00	320,00	1,00	20,00	3,00	3,00	45,00	100,00
19	RaBr	120,00	3,00	1,00	210,00	5,00	14,00	12,00	2,00	240,00	25,00
20	Rais	90,00	2,00	,00	,00	2,00	15,00	6,00	3,00	110,00	25,00
21	RiKr	110,00	2,00	,00	290,00	,00	22,00	3,00	1,00	35,00	25,00
22	Sma	110,00	2,00	1,00	70,00	1,00	9,00	15,00	2,00	40,00	25,00
23	Spec	110,00	6,00	,00	230,00	1,00	16,00	3,00	1,00	55,00	25,00

Figure 5 : le tableau des données d'enquête.

En utilisant le modèle de positionnement multidimensionnel classique, on constate sur cet exemple, en situation réelle d'application de la méthode, que l'ajustement obtenu est beaucoup moins bon que sur les données topograhiques. En effet, le pourcentage de variabilité expliquée n'est plus que de 69 % (cf. tableau 3, `"RSQ= ,69342"`). L'indicateur de la qualité de l'ajustement que constitue le stress a augmenté de manière notable : 0,27 (`"Stress= ,26845"`) contre 0,06 obtenu avec l'exemple précédent (en fait, une illustration ad hoc de la méthode).

De même si l'on consulte le diagramme de Shepard associé à ce résultat (cf. figure 8), on constate qu'un certain nombre de disparités s'éloignent très significativement d'un modèle de représentation linéaire affine, voire s'affranchissent du modèle d'une relation monotone entre distances et disparités. De l'examen du diagramme de Shepard, on conclut que le modèle métrique n'est pas adapté à cet exemple et qu'il conviendrait de poursuivre l'analyse de cet exemple avec un modèle non métrique.

```
Iteration     S-stress      Improvement
                    1           ,37423
                    2           ,31488         ,05935
                    3           ,30540         ,00948
                    4           ,30413         ,00127
                    5           ,30393         ,00020
                         Iterations stopped because
                   S-stress improvement is less than   ,001000

             Stress and squared correlation (RSQ) in distances

RSQ values are the proportion of variance of the scaled data
(disparities)
            in the partition (row, matrix, or entire data) which
             is accounted for by their corresponding distances.
              Stress values are Kruskal's stress formula 1.
                 For  matrix
    Stress  =   ,26845      RSQ =  ,69342
```

Tableau 3 : itérations de l'algorithme d'optimisation et valeurs du stress de type I les céréales du petit déjeuner.

Cependant, si l'on compare avec une méthode factorielle, l'analyse en composantes principales (ACP) du tableau **X** des données, on constate que les projections sur les deux premiers axes factoriels (53% de variance expliquée) donne une configuration (cf. figure 7) très proche de la représentation euclidienne fournie par l'algorithme ALSCAL pour les options correspondant au positionnement multidimensionnel classique : les projections sur l'axe F2 de l'ACP sont de signe contraire à celles sur la dimension D2 du positionnement multidimensionnel. On vérifie donc ici l'équivalence entre le modèle de positionnement multidimensionnel classique et l'analyse en composantes principales.

En fait, le résultat fourni par la procédure ALSCAL correspond à un positionnement multidimensionnel réalisé sans transformation monotone donnant une approximation de la solution du positionnement classique que constitue l'analyse factorielle du tableau de distances ; les différences entre les deux solutions tiennent à la nature de l'algorithme utilisé puisqu'une analyse factorielle réalise une projection (opération contractant les distances) tandis que le positionnement multidimensionnel sans transformation monotone réalise une approximation où les disparités pourront être soit supérieures soit inférieures aux distances à reconstituer ; dans le cas d'une véritable AFTD, cela se traduit par un diagramme de Shepard où tous les points sont situés sous la diagonale tandis que dans notre exemple les points se situent de part et d'autre de la diagonale.

Autre distinction entre les deux méthodes : dans l'AFDT, la solution à *n-1* dimensions est incluse dans la solution à *n* dimensions en raison du processus de recherche des directions propres, ce qui n'est pas forcément vrai en positionnement multidimensionnel où la recherche d'une solution à trois dimensions se fait indépendamment de la recherche d'une solution à deux dimensions.

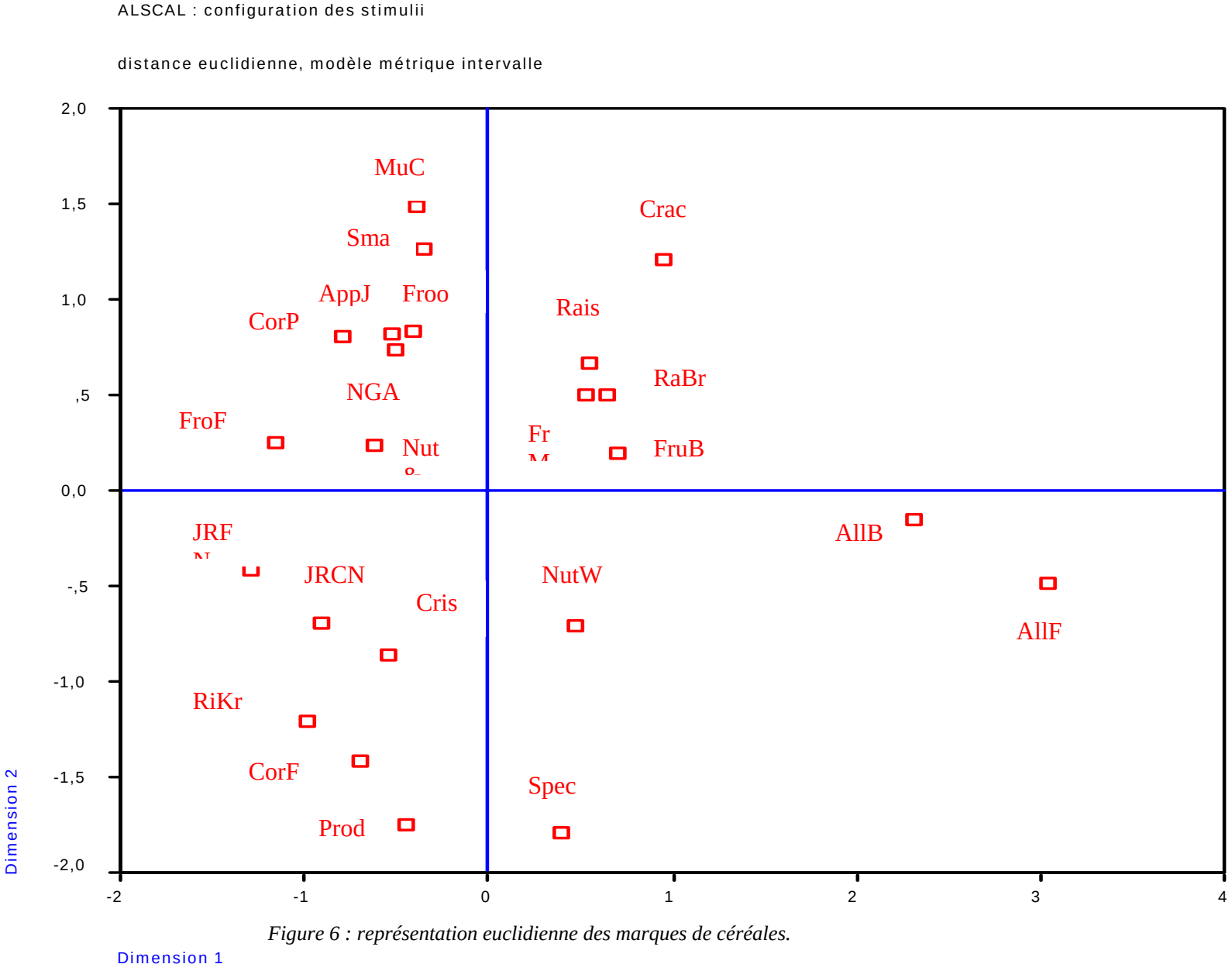


Figure 6 : représentation euclidienne des marques de céréales.

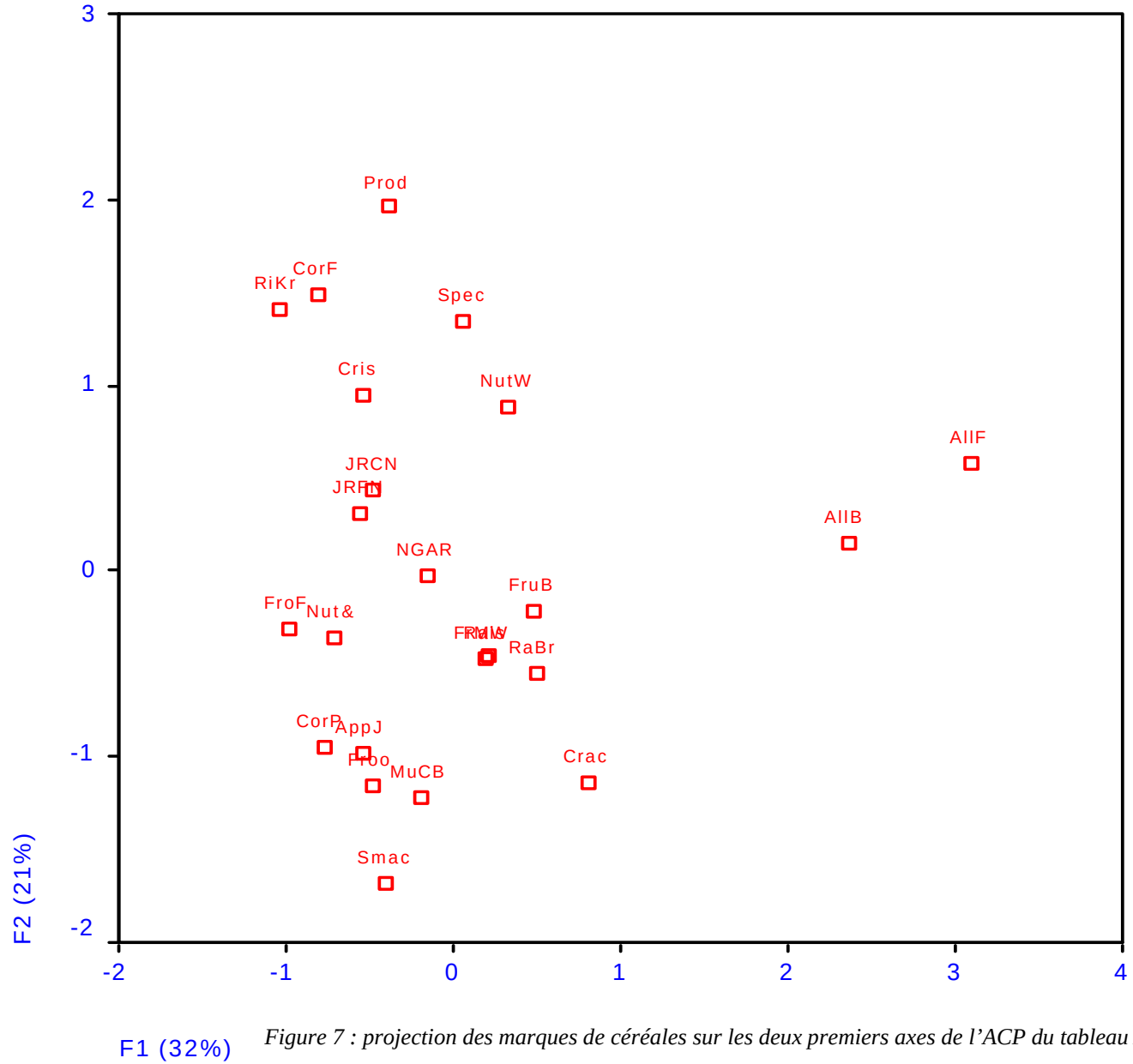


Figure 7 : projection des marques de céréales sur les deux premiers axes de l'ACP du tableau **X** .

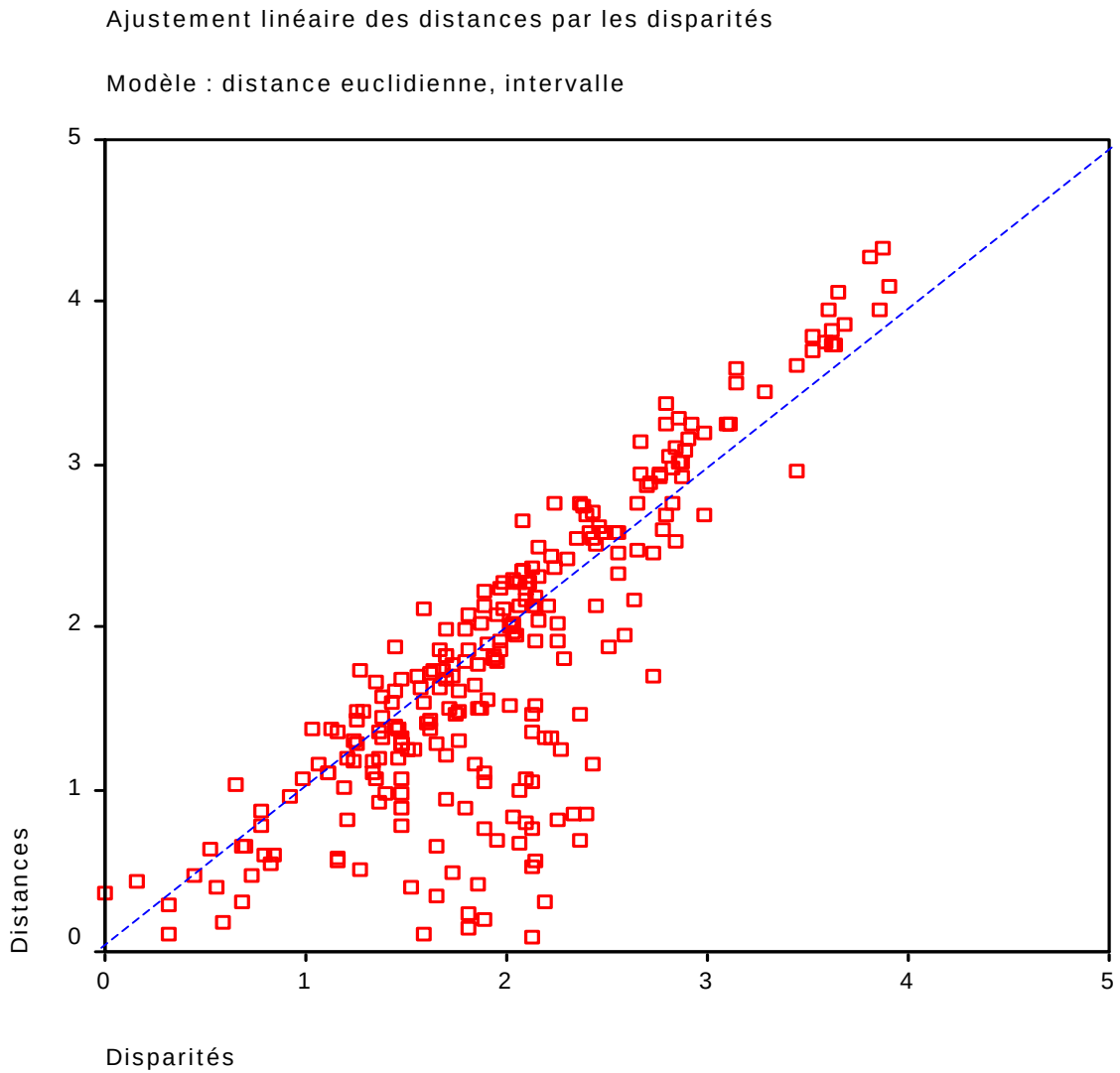
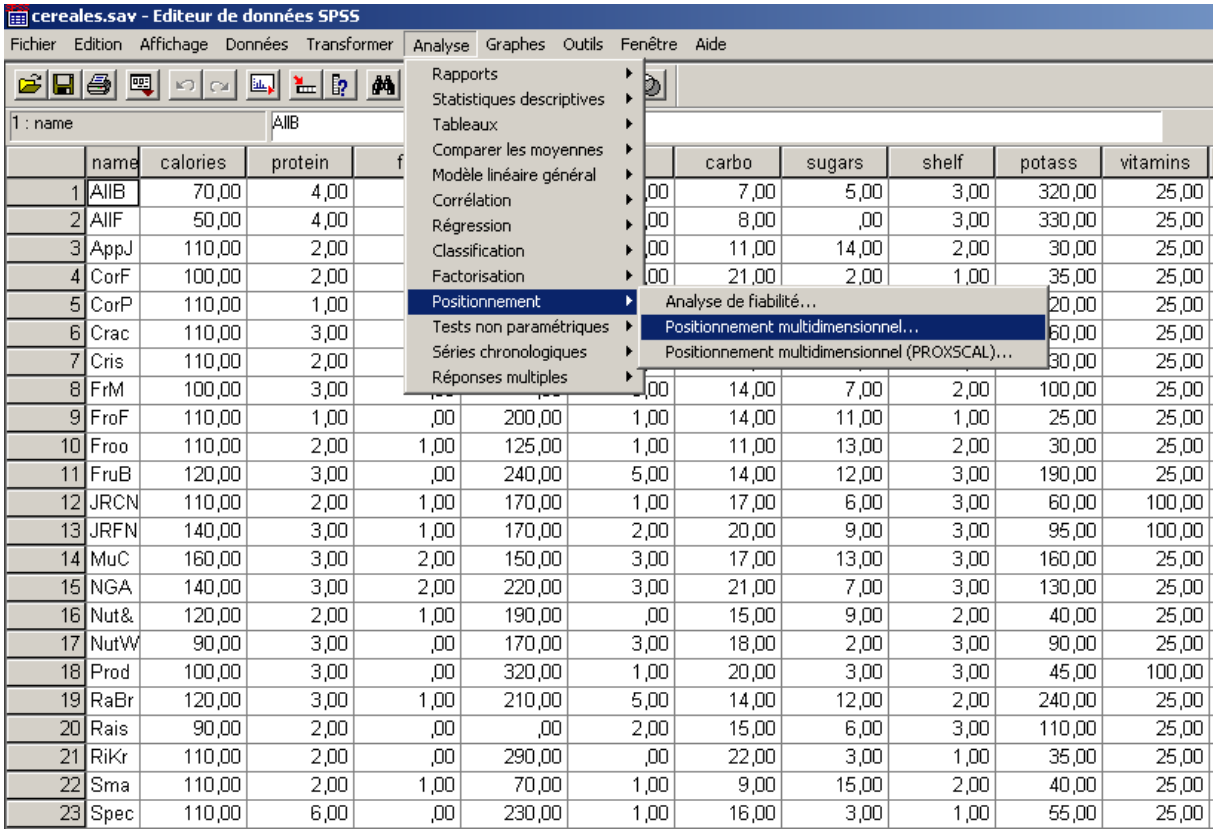


Figure 8 : diagramme de Shepard pour la représentation euclidienne des marques de céréales.

6. Spécification des paramètres du positionnement multidimensionnel

Pour effectuer un positionnement multidimensionnel, il convient de sélectionner la procédure Positionnement multidimensionnel de l'option Positionnement du menu Analyse afin d'obtenir la boîte de dialogue permettant de spécifier les principaux paramètres du positionnement multidimensionnel.



	name	calories	protein				carbo	sugars	shelf	potass	vitamins
1	AllB	70,00	4,00			,00	7,00	5,00	3,00	320,00	25,00
2	AllF	50,00	4,00			,00	8,00	,00	3,00	330,00	25,00
3	AppJ	110,00	2,00			,00	11,00	14,00	2,00	30,00	25,00
4	CorF	100,00	2,00			,00	21,00	2,00	1,00	35,00	25,00
5	CorP	110,00	1,00							20,00	25,00
6	Crac	110,00	3,00							60,00	25,00
7	Cris	110,00	2,00							30,00	25,00
8	FrM	100,00	3,00			,00	14,00	7,00	2,00	100,00	25,00
9	FroF	110,00	1,00	,00	200,00	1,00	14,00	11,00	1,00	25,00	25,00
10	Froo	110,00	2,00	1,00	125,00	1,00	11,00	13,00	2,00	30,00	25,00
11	FruB	120,00	3,00	,00	240,00	5,00	14,00	12,00	3,00	190,00	25,00
12	JRCN	110,00	2,00	1,00	170,00	1,00	17,00	6,00	3,00	60,00	100,00
13	JRFN	140,00	3,00	1,00	170,00	2,00	20,00	9,00	3,00	95,00	100,00
14	MuC	160,00	3,00	2,00	150,00	3,00	17,00	13,00	3,00	160,00	25,00
15	NGA	140,00	3,00	2,00	220,00	3,00	21,00	7,00	3,00	130,00	25,00
16	Nut&	120,00	2,00	1,00	190,00	,00	15,00	9,00	2,00	40,00	25,00
17	NutW	90,00	3,00	,00	170,00	3,00	18,00	2,00	3,00	90,00	25,00
18	Prod	100,00	3,00	,00	320,00	1,00	20,00	3,00	3,00	45,00	100,00
19	RaBr	120,00	3,00	1,00	210,00	5,00	14,00	12,00	2,00	240,00	25,00
20	Rais	90,00	2,00	,00	,00	2,00	15,00	6,00	3,00	110,00	25,00
21	RiKr	110,00	2,00	,00	290,00	,00	22,00	3,00	1,00	35,00	25,00
22	Sma	110,00	2,00	1,00	70,00	1,00	9,00	15,00	2,00	40,00	25,00
23	Spec	110,00	6,00	,00	230,00	1,00	16,00	3,00	1,00	55,00	25,00

Figure 9 : appel de la procédure SPSS de positionnement multidimensionnel simple.

Les spécifications requises concernent en premier lieu les dissimilarités (Bloc de spécification <Distances>) qui peuvent être lues soit directement sous forme de forme de matrice carrée (<Données en matrice(s)>) soit calculées à partir de tableaux de données rectangulaires croisant observations en lignes et variables en colonnes (<Calculées à partir des données>).
Ici (cf. figure 10), le choix a été de calculer ces dissimilarités entre les lignes du tableau de données (option <Entre observations>du choix de <Calcul des indices> à partir des valeurs centrées réduites (choix <Centrer-réduire> de la liste <Standardiser> du bloc <Transfomer les valeurs>). La liste des variables soumises à ce calcul est définie par leur sélection dans la liste spécifique <Variables : :>.
La seconde étape (cf. figure 11) consiste à spécifier le modèle d'ajustement des dissimilarités aux distances afin de pouvoir en déduire une représentation euclidienne. Le choix effectué est celui correspondant à un ajustement linéaire de type affine (choix <Intervalle> du bloc <Niveau de mesure> pour un modèle de positionnement multidimensionnel classique (choix <Distance euclidienne> du bloc <Modèle de positionnement>). Dans le présent contexte d'un modèle de dimensionnement simple, c'est l'option par défaut (<Matrice>) de conditionnement (bloc <Conditionnement>) qui est retenue puisque l'ensemble des dissimilarités entre les marques de céréales est obtenu à partir d'une même échelle de mesure.

Enfin, on demande (bloc <Dimension>, option par défaut <Minimum=2>, <Maximum=2>) que la représentation euclidienne s'effectue dans un espace de dimension 2.

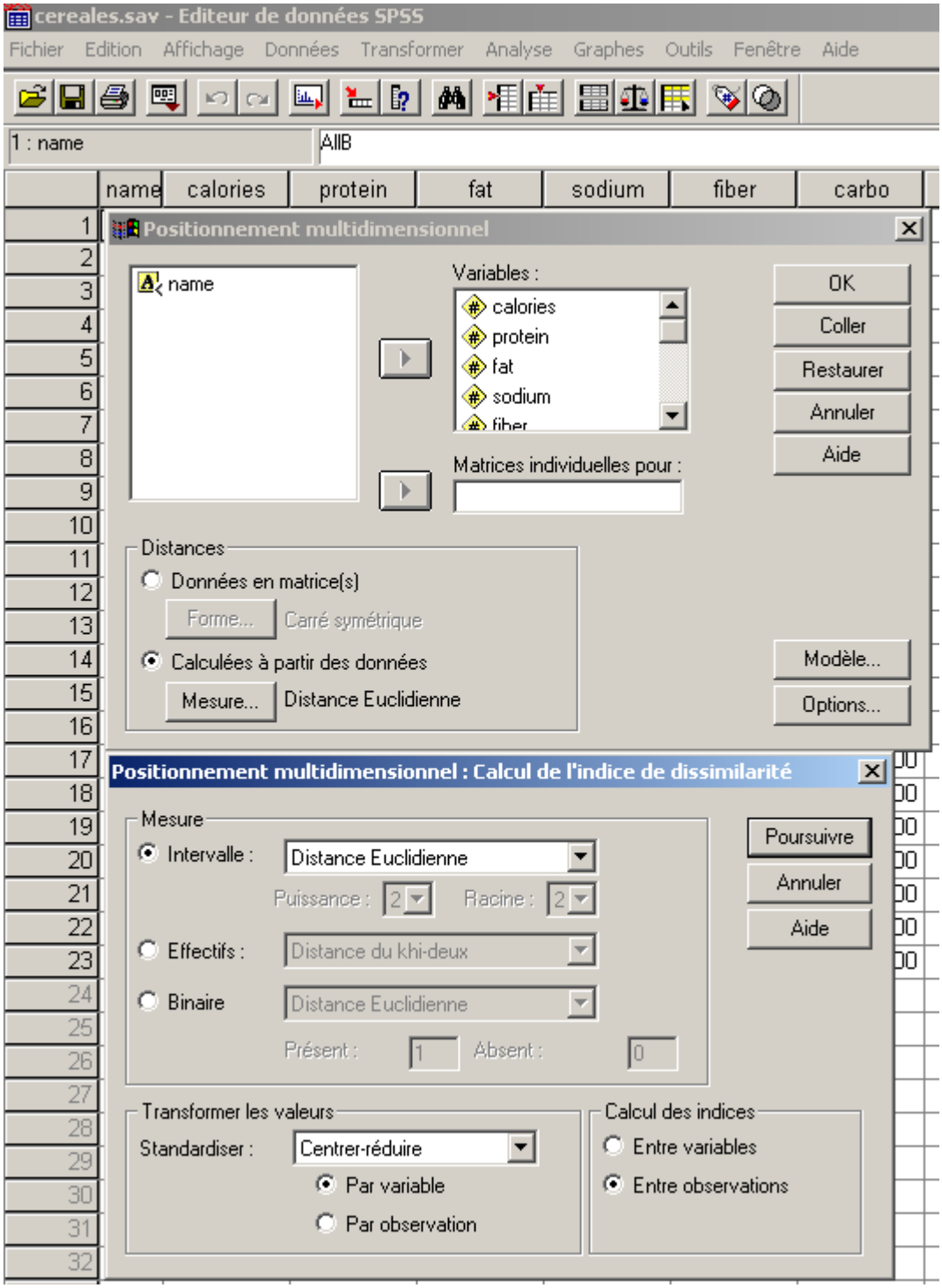


Figure 10 : calcul de l'indice de dissimilarité entre objets.

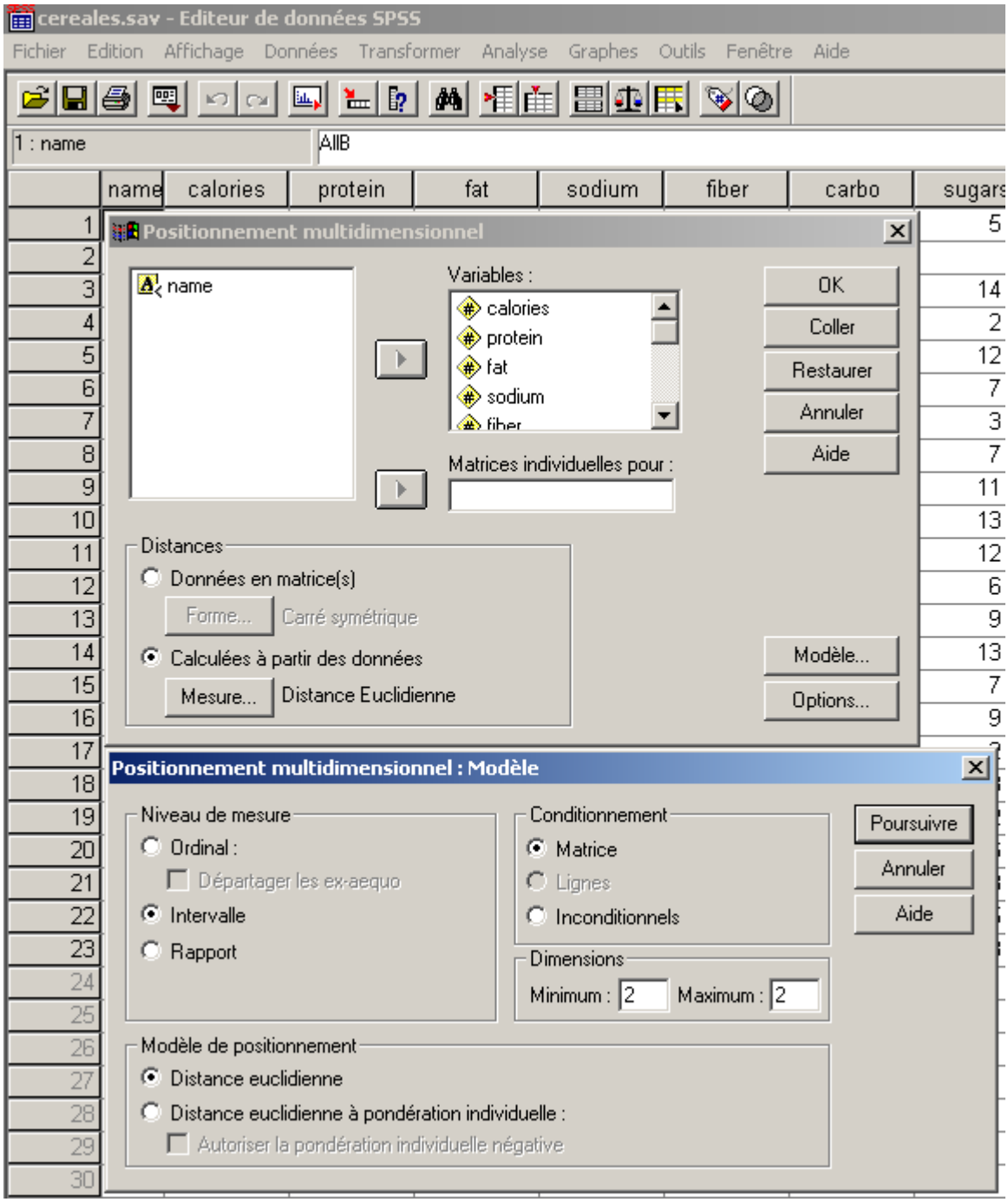


Figure 11 : choix du modèle d'ajustement des dissimilarités aux distances.

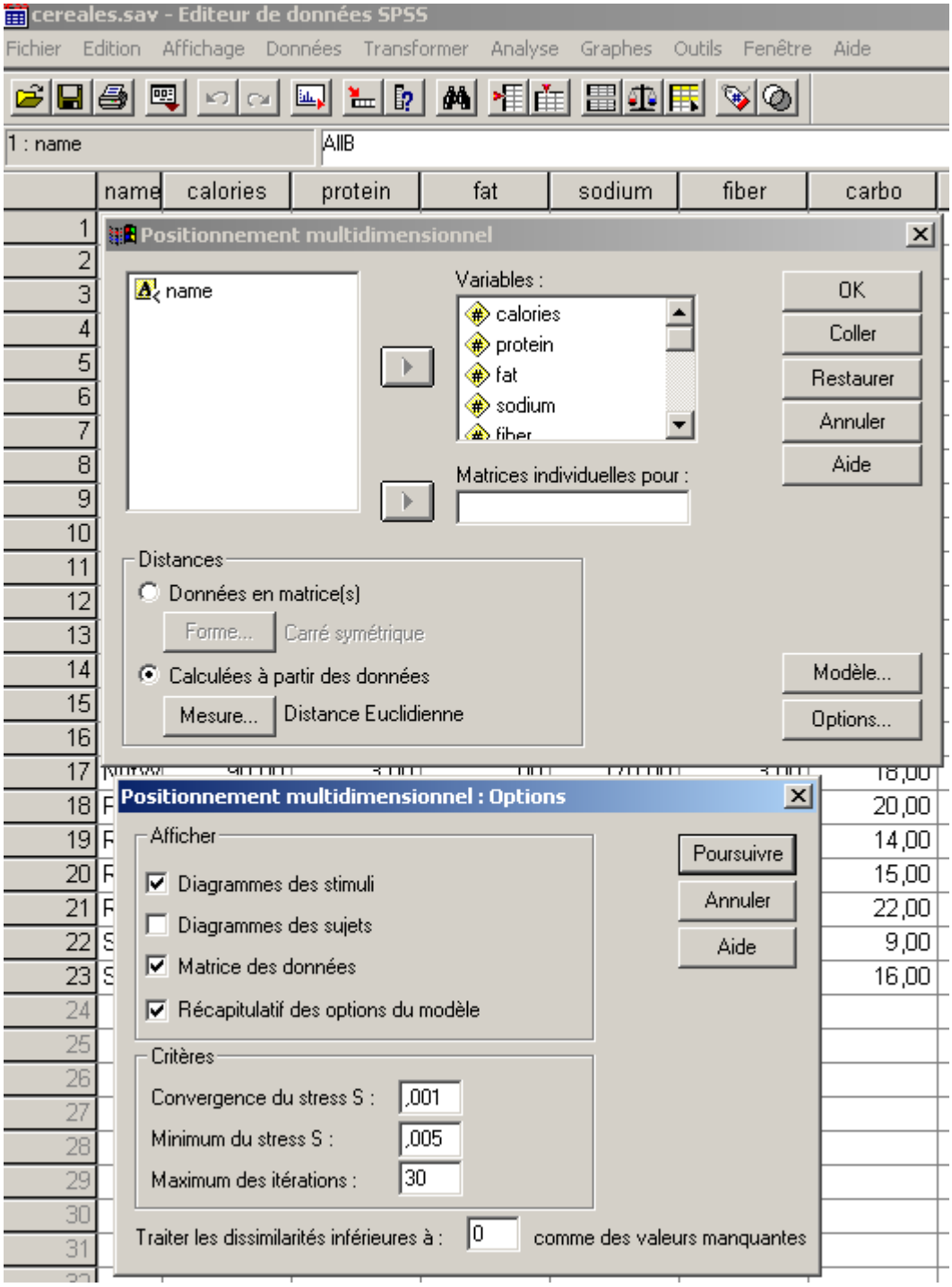


Figure 12 : options du traitement, critères de convergence et résultats.

Concernant les diverses options proposées de traitement (cf. figure 12), on demande d'afficher (bloc <Afficher>) la représentation euclidienne des objets (choix <Diagramme des stimuli>), la matrice symétrique des proximités calculées (choix <Matrice des données>) qui constitue les « données » à partir duquel s'est effectué l'ajustement linéaire affine et la sélection des paramètres de la procédure (choix <Récapitulatif des options du modèle> qui permettra de se repérer ultérieurement parmi les listings des différents essais effectués.

L'ajustement des disparités aux distances s'effectuant à partir d'un algorithme numérique de recherche du minimum de la fonction objectif (S-STRESS), un certain nombre de paramètres permettent de contrôler la convergence de cet algorithme itératif : figurent dans le bloc <Critères>, les valeurs par défaut d'arrêt de l'algorithme (<Convergence du stress S=0,001>), de borne minimale de la fonction objectif (<Minimum du stress S=0,005>), du maximum d'itérations autorisées (<Maximum des itérations=30>. De même, on définit une borne minimale autorisée pour les dissimilarités, option par défaut (<=0>), en assimilant les valeurs inférieures à des valeurs manquantes.
Toutes ces valeurs par défaut peuvent bien sûr être modifiées dans les essais que l'on est amené à effectuer pour rechercher la convergence de l'algorithme itératif.

Références bibliographiques

Les références ci-dessous correspondent soit à des travaux cités dans le texte de cette note, soit à des articles princeps, soit à des ouvrages de synthèse. Consultée lors de la révision de cette note, la synthèse récente « Positionnement multidimensionnel et quantification vectorielle » de Gérard d'Aubigny (2003) permettra au lecteur intéressé d'approfondir ce sujet.

Cox T.F., Cox M.A.A. (2001) *Multidimensional Scaling (2nd Ed.),* Chapman & Hall/CRC, Londres.

Cailliez F. (1983) « The Analytical Solution of the Additive Constant Problem », ***Psycho****metrika*, **48**, 305-308.

Cailliez F., Pages J.P. (1976) *Introduction à l'analyse des données*, SMASH, Paris.

Aubigny (d') G. (2003) « Positionnement multidimensionnel et quantification vectorielle», in *Traitement du signal et de l'image. Analyse des Données* Govaert G. (dir.) Hermès, 105-150.

Richardson M.W. (1938) « Multidimensional Psychophysics », *Psychological Bulletin* **35**, 659-660.

Rothkopf E.W. (1957) « A Measure of Stimulus Similarity and Errors in some Paired-Associate Leaning Tasks », *J. of Experimental Psychology* **53**, 94-101.

Shepard R.N. (1962) « The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function », *Psychometrika* **27**, 125-140, 219-246.

Kruskal J.B. (1964a) « Multidimensional Scaling by Optimizing Goodness of Fit to a Non Metric Hypothesis », *Psychometrika*, **29**, 1-27.

Kruskal J.B. (1964b) « Non Metric Multidimensional Scaling: a Numerical Method », *Psychometrika*, **29**, 115-129.

Kruskal J.B., Wish M. (1984) « Multidimensional Scaling », *Quantitative Applications in the Social Sciences*, **11**, Sage University Paper.

Pagès J. (2003) « Recueil direct de distances sensorielles: application à l'évaluation de dix vins blancs du Val de Loire ». *Sciences des aliments*, **23**, 679-688.

Torgerson W.S. (1952) « Multidimensional Scaling: I Theory and Method », *Psychometrika*, **17**, pp. 401-419.

Tournois J., Dycke P. (1993) *Pratique de l'échelonnement multidimensionnel*, De Boeck-Wesmael, Bruxelles.

Young F.W., Harris D.F. (1994. « Chapter 7: Multidimensional Scaling », in *SPSS Professional Statistics 6.1*, Marijja J. Norusis (dir.), Chicago, IL, USA, 155-222.

Young F.W., Takane Y., Lewykyj R. (1978) « ALSCAL: A Nonmetric Multidimensional Scaling Program With Several Difference Options », *Behavioral Research Methods and Instrumentation*, **10**, pp. 451-453.
Young F.W., Lewykyj R. (1979) *ALSCAL-4 User's Guide*, Data Analysis and Theory Associates, Carborro, NC, USA.
Young G., Householder A.S. (1938) « Discussion of a Set of Points in Terms of their Mutual Distances », *Psychometrika*, **3**, pp. 19-22.

Portrait d'Alston Scott Householder (1904-1993),
Extrait de "The History of Mathematics Archive"
John O'Connoret Edmund Robertson, School of Mathematics and Statistics, University of St Andrews

Remerciements

L'auteur remercie Francis Caillez, Jérôme Pagès et Gilbert Saporta pour leurs remarques critiques et les conseils prodigués à la lecture de la version initiale de cette note, cependant les éventuelles erreurs et omissions qui pourraient demeurer relèvent de sa seule responsabilité.