Cah.Techn. I.N.R.A., 2004 53, 35-58

UNE INTRODUCTION A L'ANALYSE D'HOMOGENEITE AVEC SPSS POUR WINDOWS

Dominique Desbois [1]

RESUME :
L'analyse d'homogénéité est une méthode exploratoire multidimensionnelle qui fournit une représentation synthétique des catégories issues d'une batterie de critères qualitatifs, référentiel d'un protocole d'expérimentation ou d'enquête. Cette note a pour but d'aider les utilisateurs de *SPSS pour Windows* dans la mise en oeuvre de l'analyse d'homogénéité au moyen de la procédure HOMALS du logiciel *SPSS pour Windows*. Cette mise en oeuvre concerne l'analyse de tableaux de données construits à partir de variables nominales. L'équivalence entre l'analyse d'homogénéité et les principales méthodes factorielles, en particulier l'analyse des correspondances multiples, est illustrée à partir d'exemples complétant l'exposé théorique consacré aux méthodes.

MOTS CLEFS :
Analyse d'homogénéité, analyse des correspondances multiples, logiciel statistique, mise en oeuvre.

HOMALS[2] [Gifi, 1990] est une procédure itérative basée sur la technique des moindres carrés alternées permettant de réaliser une analyse d'homogénéité. L'une des options particulières de cette procédure fournit les facteurs d'une analyse des correspondances multiples. L'objectif de cette note est donc de présenter l'analyse d'homogénéité pour les utilisateurs francophones de SPSS afin qu'ils puissent utiliser plus aisément cette procédure pour dépouiller leurs données d'enquête de façon pertinente, en réalisant des analyses de correspondances multiples.

1. L'ANALYSE D'HOMOGENEITE, POUR UNE REPRESENTATION OPTIMALE DES CATEGORIES.

Soit un ensemble d'observations décrivant des **objets** au moyen de **catégories** issues d'une batterie de critères qualitatifs (**variables catégorielles**). L'analyse d'homogénéité est une technique exploratoire d'analyse des données permettant de décrire les relations existant entre deux ou plusieurs de ces variables catégorielles en fournissant une représentation graphique de leurs catégories, sous la forme d'un nuage de points (**points-catégories**) projetés dans un sous-espace de faible dimension.
Cette représentation graphique, effectuée dans un système d'axes orthonormés appelés « **dimensions** » est optimale au sens où elle maximise l'écart entre les positions des différentes catégories. Dans ce sous-espace particulier, on peut également représenter les objets soumis à l'observation (**points-objets**) en liant leur représentation à celle des catégories de référence de l'étude. Pour chaque variable, les catégories d'une même variable scindent le nuage des points représentant les objets en sous-nuages de points qui rassemblent les objets partageant la même catégorie. Les points représentant les catégories sont situés au centre du sous-nuage des points représentant les objets qui appartiennent à la même catégorie. Les proximités entres objets reflètent les similarités ou les dissimilarités entre leurs configurations respectives de réponse à la batterie de critères qualitatifs. Ainsi, les objets partageant un

[1] *INRA-ESR Nancy et SCEES - 251, rue de Vaugirard, 75732 Paris Cedex 15.*
Courriel :Dominique.Desbois@agriculture.gouv.fr Fax : +33 1 49 55 85 00
[2] *Homogeneity Analysis by Alternating Least Squares*– Analyse d'homogénéité par les moindres carrés alternés

même profil de réponse sont projetés en un même point. Cependant, la réciproque n'est pas forcément vérifiée : deux objets dont les scores (valeurs de la projection selon les dimensions) sont proches ne sont pas nécessairement similaires. Si une variable possède un bon **pouvoir discriminant**, les objets se situeront à proximité des catégories auxquelles ils appartiennent. Idéalement, les objets classés dans la même catégorie doivent se situer à proximité les uns des autres, leurs scores étant similaires. Les catégories appartenant à des variables différentes sont situées à proximité les unes des autres si elles caractérisent les mêmes sous-ensembles d'objets. Ainsi, deux objets ayant des scores similaires pour un critère particulier doivent posséder des scores similaires pour les variables qui lui sont **homogènes**.

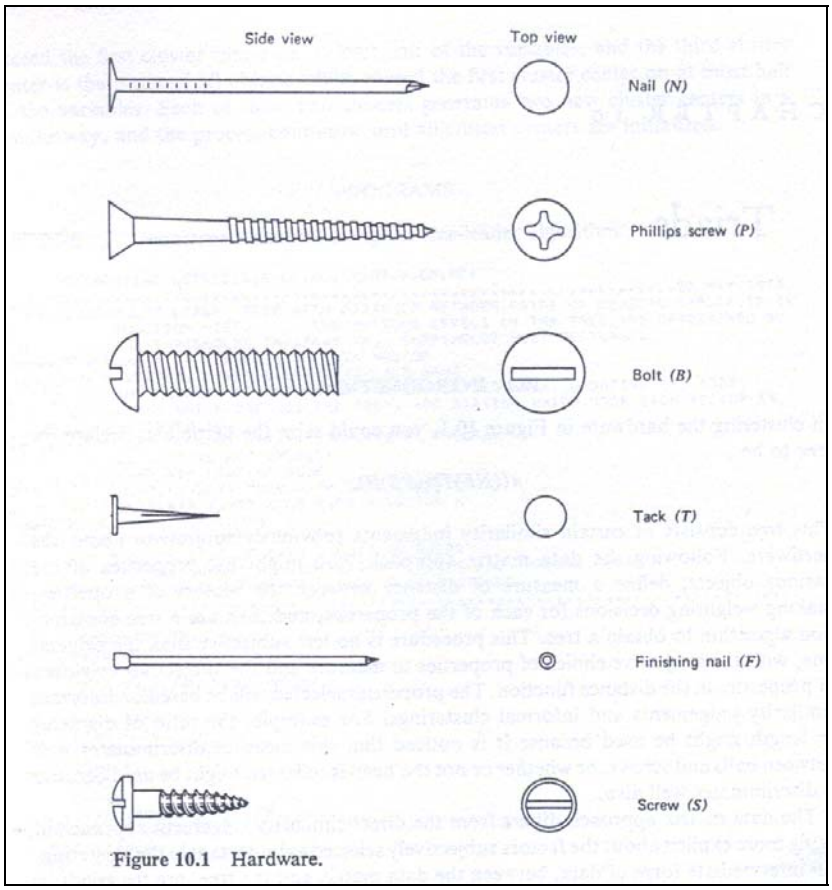
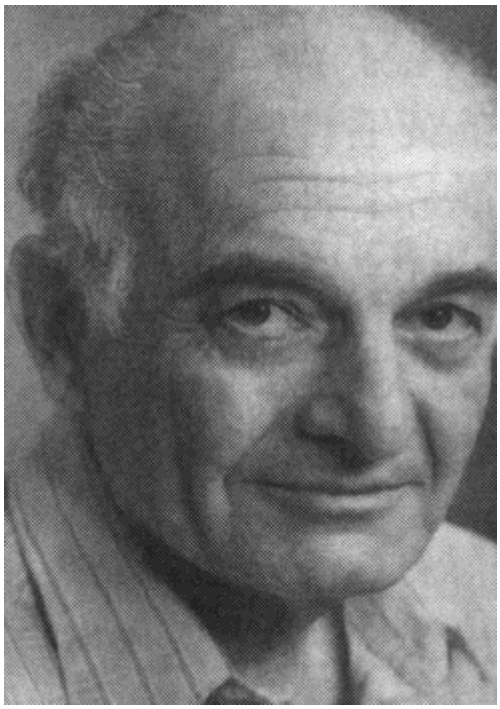


Figure 1 : *visualisation des objets, face et profil du petit matériel de quincaillerie (extrait de l'ouvrage [Hartigan 1975])***.**

Le terme d'**homogénéité** se réfère donc à une situation où les variables fournissent une partition de l'ensemble des objets selon les mêmes catégories ou des catégories similaires. Historiquement, le concept d'homogénéité est associé à un paradigme selon lequel des variables distinctes peuvent mesurer le même phénomène. Par exemple, pour les psychométriciens, les performances intellectuelles sont approchées à travers une batterie de tests qualifiés d'homogènes, au sens ou la somme des scores obtenus à un sens car elle fournit une mesure de ces performances.

De façon plus formelle, on peut définir l'analyse d'homogénéité, *stricto sensu*, comme un programme de minimisation d'une fonction-objectif, la **perte d'homogénéité** (cf. infra § 3 pour une définition), permettant d'obtenir une représentation graphique des catégories qui corresponde à la solution optimale présentée antérieurement. La généralisation de cette définition fournit un cadre méthodologique où le terme d'analyse d'homogénéité se réfère à une famille de techniques d'analyse multivariée partageant, selon différentes formes de codage des données et sous des formulations diverses du critère d'optimalité, un paradigme commun d'optimisation de l'homogénéité des variables.

L'analyse d'homogénéité peut être également présentée comme la solution d'un problème de décomposition en valeurs propres et en valeur propres singulières, et peut de ce fait être rattachée aux méthodes factorielles : ainsi, pour deux critères qualitatifs, l'analyse d'homogénéité est équivalente à l'analyse des correspondances ; pour plusieurs critères, elle est équivalente à l'analyse des correspondances multiples. A ce titre, elle peut également être présentée comme une méthode de positionnement multidimensionnel travaillant à partir d'un tableau de « dissimilarités » constitué par les distances du Khi-Deux entre profils-lignes issus d'un tableau disjonctif complet codant, pour la **population *I*** des objets, les caractéristiques observées selon **l'ensemble *J*** des **modalités** ou catégories d'observation. L'analyse d'homogénéité peut également être considérée comme une analyse en composantes principales sur données nominales (modèle de Guttman). Lorsqu'il n'y a pas de relations linéaires entre variables ou lorsque les variables sont nominales, l'analyse d'homogénéité est préférable à une analyse en composantes principales normée (i.e. effectuées sur variables centrées et réduites).



Portrait de Louis GUTTMAN, 1916-1987
(Materials for the History of Statistics, The University of York)

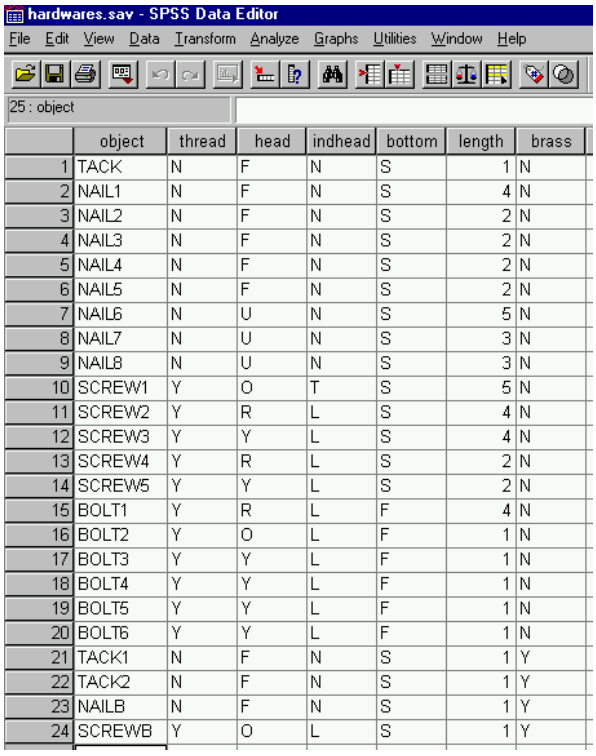
2. UN EXEMPLE D'ANALYSE D'HOMOGENEITE : les petits articles de quincaillerie.

Ce premier exemple illustratif de l'analyse d'homogénéité est basé sur des données décrivant de petits articles de quincailleries (clous, vis, boulons, etc.) à l'aide de variables catégorielles [Hartigan, 1975] décrivant leur forme et leur dimension. Il y a ***n=24*** objets ou observations et ***p=6*** variables descriptives catégorielles, la variable *OBJECT* identifiant les 24 observations.

Nom	Valeur	*Etiquette*	Position
OBJECT		*Objet*	1
THREAD		*Pointe*	2
	N	*non*	
	Y	*oui*	
HEAD		*Forme de la tête*	3
	F	*plate*	
	O	*conique*	
	R	*ronde*	
	U	*coupe*	
	Y	*cylindre*	
INDHEAD		*Indentation de la tête*	4
	L	*fente*	
	N	*aucune*	
	T	*étoile*	
BOTTOM		*Forme de la base*	5
	F	*plate*	
	S	*tranchante*	
LENGTH		*Longueur en demi-pouces*	6
	1	*0,5"*	
	2	*1"*	
	3	*1,5"*	
	4	*2"*	
	5	*2,5"*	
BRASS		*Cuivré*	7
	N	*non*	
	Y	*oui*	

Tableau 1 : *descriptif des données et détail des catégories*

Ci-dessous figure, dans l'éditeur de données *SPSS*, le tableau de ces données descriptives sous forme alphanumérique :



	object	thread	head	indhead	bottom	length	brass
1	TACK	N	F	N	S	1	N
2	NAIL1	N	F	N	S	4	N
3	NAIL2	N	F	N	S	2	N
4	NAIL3	N	F	N	S	2	N
5	NAIL4	N	F	N	S	2	N
6	NAIL5	N	F	N	S	2	N
7	NAIL6	N	U	N	S	5	N
8	NAIL7	N	U	N	S	3	N
9	NAIL8	N	U	N	S	3	N
10	SCREW1	Y	O	T	S	5	N
11	SCREW2	Y	R	L	S	4	N
12	SCREW3	Y	Y	L	S	4	N
13	SCREW4	Y	R	L	S	2	N
14	SCREW5	Y	Y	L	S	2	N
15	BOLT1	Y	R	L	F	4	N
16	BOLT2	Y	O	L	F	1	N
17	BOLT3	Y	Y	L	F	1	N
18	BOLT4	Y	Y	L	F	1	N
19	BOLT5	Y	Y	L	F	1	N
20	BOLT6	Y	Y	L	F	1	N
21	TACK1	N	F	N	S	1	Y
22	TACK2	N	F	N	S	1	Y
23	NAILB	N	F	N	S	1	Y
24	SCREWB	Y	O	L	S	1	Y

Figure 2 : *le tableau des données alphanumériques*

2.1. Pouvoir explicatif des dimensions de la solution

La représentation graphique que l'on souhaite obtenir de ces données en termes de catégories et d'objets, s'effectue dans un repère orthonormé dont on doit préciser le nombre d'axes ***a*** **,** appelé la **dimension de la solution**. La dimension maximum du sous-espace de représentation est égale soit au **nombre de catégories** (***m=19***) moins le nombre de variables sans valeurs manquantes (*p=6*), soit au nombre d'observations (*n=24*) moins un si celui-ci est inférieur, soit ***a****=min{13,23}****=13***. En pratique, le nombre d'axes utilisé pour la représentation est généralement très inférieur à ce maximum car souvent une solution comportant deux ou trois dimensions suffit pour synthétiser les traits essentiels de l'information contenue dans le tableau des données, l'information additionnelle apportée par des dimensions supplémentaires se révélant marginale.

Les **valeurs propres** permettent de rendre compte de l'importance relative de chaque dimension dans la part d'information statistique pris en compte par la solution. Ces valeurs propres prennent des valeurs dans l'intervalle $[0;1]$. La valeur 1 est atteinte par la valeur propre triviale qui correspond au vecteur propre reliant le centre de gravité du nuages des profils catégoriels et l'origine du repère. Les valeurs propres nulles correspondent à des directions indéterminées de la solution[3].

Eigenvalues

Dimension	Eigenvalue
1	,621
2	,368

Tableau 2 : *les deux premières valeurs propres.*

Leur rapport avec la somme totale des valeurs propres, appelé le **taux d'inertie** en analyse des correspondances, constitue une mesure pessimiste de la part de variabilité globale prise en compte. La procédure *HOMALS* de *SPSS* étant limitée à 10 dimensions, le calcul est effectué dans ce sous-espace. Néanmoins, les valeurs propres d'ordre supérieur ayant une valeur résiduelle, cette approximation ne change pas fondamentalement l'estimation des taux d'inertie.

Dimension	Valeur propre	Taux d'inertie	Inertie cumulée
1	0,621	0,287	0,287
2	0,368	0,170	0,457
3	0,328	0,151	0,608
4	0,279	0,129	0,737
5	0,197	0,091	0,828
6	0,128	0,059	0,887
7	0,086	0,040	0,927
8	0,084	0,039	0,966
9	0,056	0,026	0,991
10	0,019	0,009	1,000

Tableau 3 : *taux d'inertie associés au valeurs propres.*

[3] tout vecteur est solution de l'équation aux valeurs propres, donc vecteur propre.

Ainsi, les deux dimensions retenues permettent de prendre en compte 46% de l'inertie totale à travers une représentation graphique plane interprétable en termes de distances entre observations.

2.2. Représentation graphique des objets à partir des scores

Les ***scores*** (coordonnées des objets selon les premières dimensions de la solution) permettent de repérer les valeurs extrêmes (« *outlier* ») : l'objet projeté à l'extrémité négative de la dimension 2 (**D2<0**) peut être considéré comme une valeur atypique ou aberrante et, de ce fait, éventuellement exclu lors d'une analyse ultérieure (cf. infra).

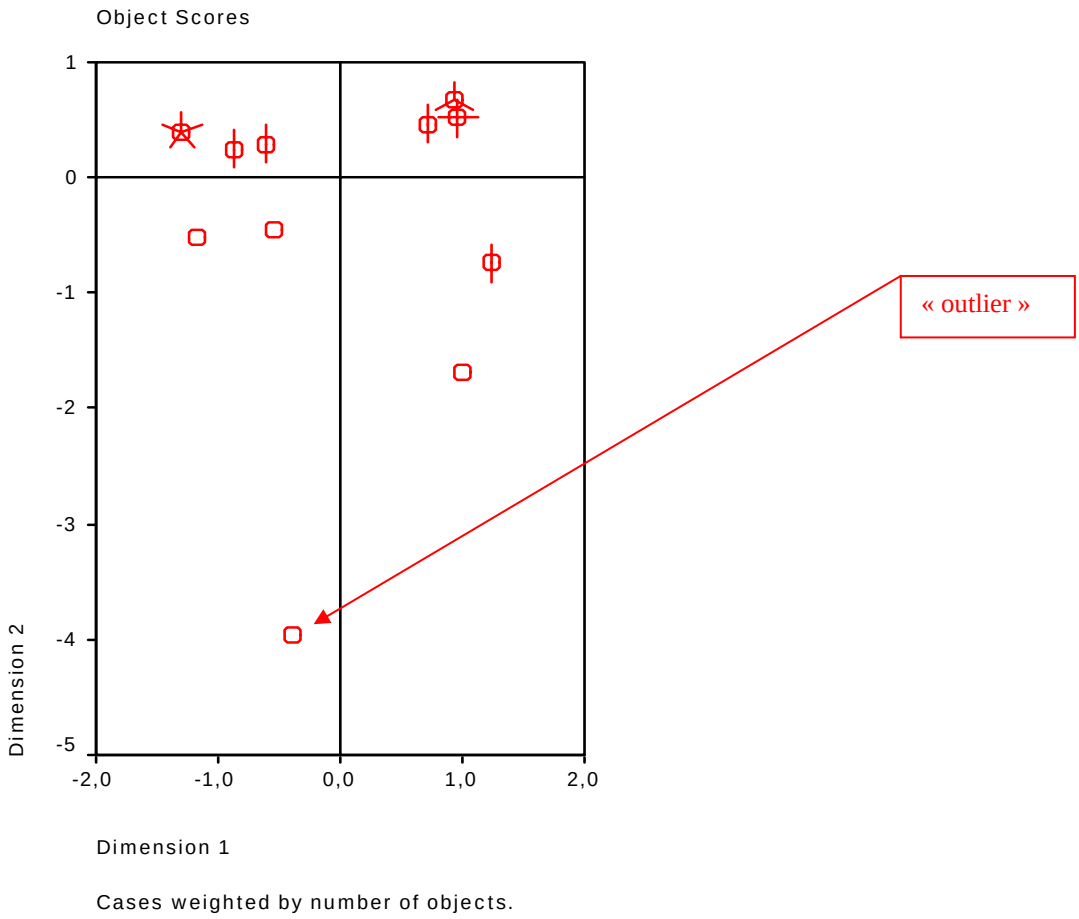


Figure 3 : *projection des objets dans le plan des deux premières dimensions.*

Cette représentation des objets sous forme de **tournesol** (le nombre de pétales du tournesol est proportionnel au nombre d'objets) est bien adaptée aux ensembles d'objets dont la cardinalité *n* est importante car elle permet de rendre compte des différences de densité au sein du nuage des points-objets.

Si le nombre d'observations est suffisamment faible, il est alors possible de projeter chacune des observations avec leur identifiant. Cela permet de vérifier la configuration de réponses fournies par des sous-ensembles particuliers d'objets. Ce graphique permet de constater que la **première dimension** (axe horizontal **D1**) sépare les vis (*screw*) et les boulons (*bolt*), qui ont un filetage (*thread*), des clous (*nail*) et des punaises (*tack*) qui n'en ont pas. De façon moins prononcée, cette première dimension instaure une séparation entre les boulons (*bolt*) qui ont

une base plate et tous les autres objets (qui ont une base pointue). La **seconde dimension** (axe vertical **D2**) sépare les objets *screw1* et *nail6* de l'ensemble des autres objets : ces deux objets sont les plus longs (cf. figure 2). Notons également que *screw1* apparaît comme l'objet le plus éloigné de l'origine : la configuration des caractéristiques de cet objet apparaît comme très spécifique puisqu'elle n'est partagée par aucun autre objet.

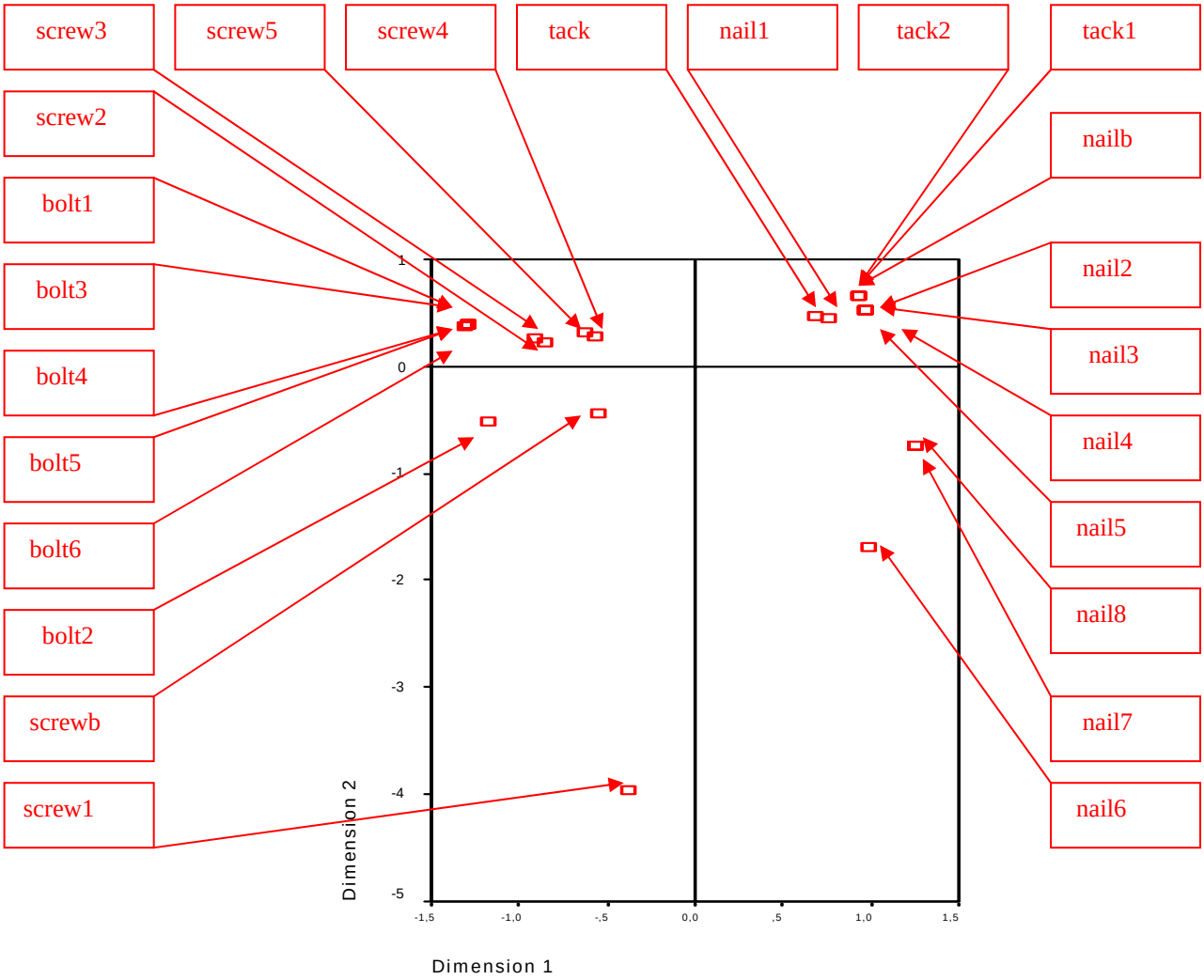


Figure 4 : *étiquetage des objets dans le plan des deux premières dimensions.*

Cependant, la pratique des variables illustratives (cf. infra § 2.5) dans l'établissement des graphiques facilite la synthèse de ces informations : pour chacun de ces graphiques illustratifs, les objets sont étiquetés à partir de la palette de valeurs catégorielles issue de la variable illustrative sélectionnée.
La procédure *HOMALS* permet de spécifier les variables illustratives utilisées pour produire une représentation graphique de la densité des différentes modalités de réponse.

2.3. Mesures du pouvoir discriminant

La mesure du pouvoir discriminant d'une variable relativement à une dimension peut se définir comme le pourcentage de variance de la dimension expliqué par cette variable, c'est à dire dans le langage de l'analyse des correspondances, la contribution de cette variable à l'inertie d'un axe factoriel. La valeur maximum de cet indicateur est égale à 1 si tous les objets se répartissent sur l'ensemble de ces catégories (caractère complet de la nomenclature des catégories) et si les objets appartenant à la même catégorie se révèlent identiques en termes de configuration descriptive relativement aux autres critères. S'il y a des données manquantes dans le tableau analysé, l'indice du pouvoir discriminant de la variable peut être supérieur à 1.

Cette mesure du pouvoir discriminant étant calculée comme la somme des carrés des coordonnées des catégories (*scores*), elle est d'autant plus élevée que les catégories de la variable considérée présentent une dispersion importante de leurs coordonnées selon la dimension examinée. La moyenne des indices de discrimination sur l'ensemble des variables est égale pour chaque dimension à la valeur propre correspondante, exprimant ainsi la variance de cette dimension.

Les dimensions sont ordonnées dans l'ordre décroissant de leur variance , les valeurs propres étant extraites par ordre d'importance décroissant : la direction de la première dimension correspond au vecteur propre associé à la première valeur propre (la plus élevée) ; la direction de la seconde dimension correspond au second vecteur propre associé à la seconde valeur propre en importance ; etc.

Le diagramme des mesures du pouvoir discriminant indique que la première dimension est constituée par une synthèse des variables *thread* (présence d'une pointe) et *bottom* (forme de la base) : les deux variables présentent des niveaux d'indice de discrimination importants pour la 1ère dimension et faibles pour la 2nde dimension. Ainsi, les catégories de ces variables sont bien dispersées selon l'axe *D1* et peu dispersées selon l'axe *D2*.

Inversement, la variable *length* présente une valeur élevée de l'indice de discrimination selon l'axe *D2* et une valeur faible pour l'axe D1. En conséquence, l'angle entre le vecteur correspondant à cette variable et la 2nde dimension est faible, la valeur de l'indice selon l'axe *D2* correspondant au carré du cosinus de l'angle. Cet indice, assimilable au carré d'un coefficient de corrélation (R^2), exprime la similarité entre les deux directions, et reflète la ségrégation observée selon la 2nde dimension sur le diagramme des objets entre les objets les plus longs (situés dans le demi-plan *D2<0*) et l'ensemble des autres objets (situés dans le demi-plan *D2>0*).

Remarquons également que les variables concernant la forme et l'indentation de la tête présentent des valeurs importantes de leurs indices de discrimination selon les deux dimensions.

Par contre la variable *brass* située près de l'origine du graphique n'apparaît pas comme discriminante dans ce plan des deux premières dimensions, l'ensemble des objets pouvant posséder ou non le caractère cuivré. Pour la même raison, la variable *length* ne peut être liée à la 1ère dimension puisqu'elle ne discrimine les objets que dans la 2nde dimension.

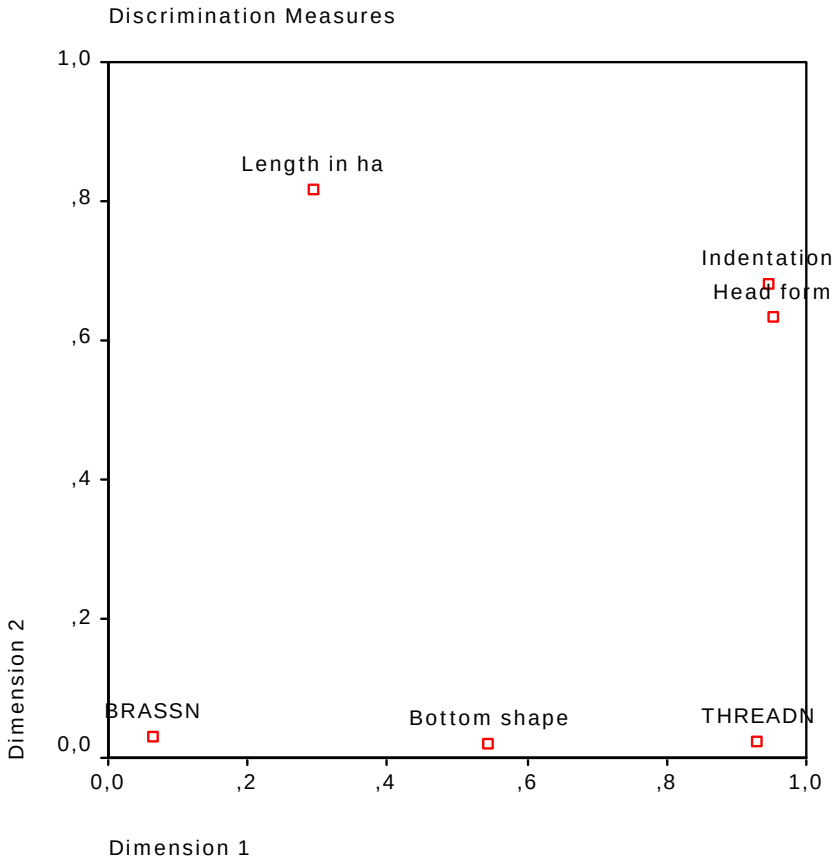


Figure 5 : *mesure du pouvoir discriminant selon les deux premières dimensions.*

Si l'indice de discrimination indique quelle est la part de variance expliquée par une variable pour chaque dimension, il ne permet pas de distinguer entre les variables dont les catégories présentent une dispersion moyenne selon une dimension et celles dont la plupart des catégories ont des coordonnées similaires à l'exception de certaines d'entre elles très différentes.

2.4. Quantifications des catégories

En revanche, les projections graphiques des catégories permettent de caractériser précisément les relations entre catégories d'une même variable mais aussi entre catégories de variables distinctes, en situant chaque catégorie sur un même graphique au moyen de leurs *quantifications* selon chaque dimension (équivalent des coordonnées factorielles des profils catégoriels dans l'analyse des correspondances multiples).

Ainsi, la variable *length* possède cinq catégories dont trois sont localisées dans la partie supérieure du graphique (demi-plan *D2>0)* et les deux autres (soit 1,5'' et 2,5'') se situent dans la partie inférieure du graphique (demi-plan *D2<0*).

En outre, la catégorie étiquetée *2_1/2_in* (soit 2,5'') située à l'extrémité négative de la 2nde dimension, se singularise très nettement par rapport à l'ensemble des autres catégories, rejoignant en cela la catégorie *STAR* (tête en étoile ou cruciforme) de la variable *Indentation of head* (indentation de la tête). En fait, la catégorie *2_1/2_in* est située au point moyen

(barycentre) des localisations des deux objets qui partagent cette spécificité, soit *screw1* et *nail6*.
La catégorie *STAR* se situe exactement au lieu géométrique de projection de l'objet *screw1* qui est le seul à présenter cette indentation cruciforme de la tête. Cette catégorie *STAR* se différencie des deux autres catégories (*SLIT* – fente et *NONE* – sans indentation) selon la 2nde dimension.

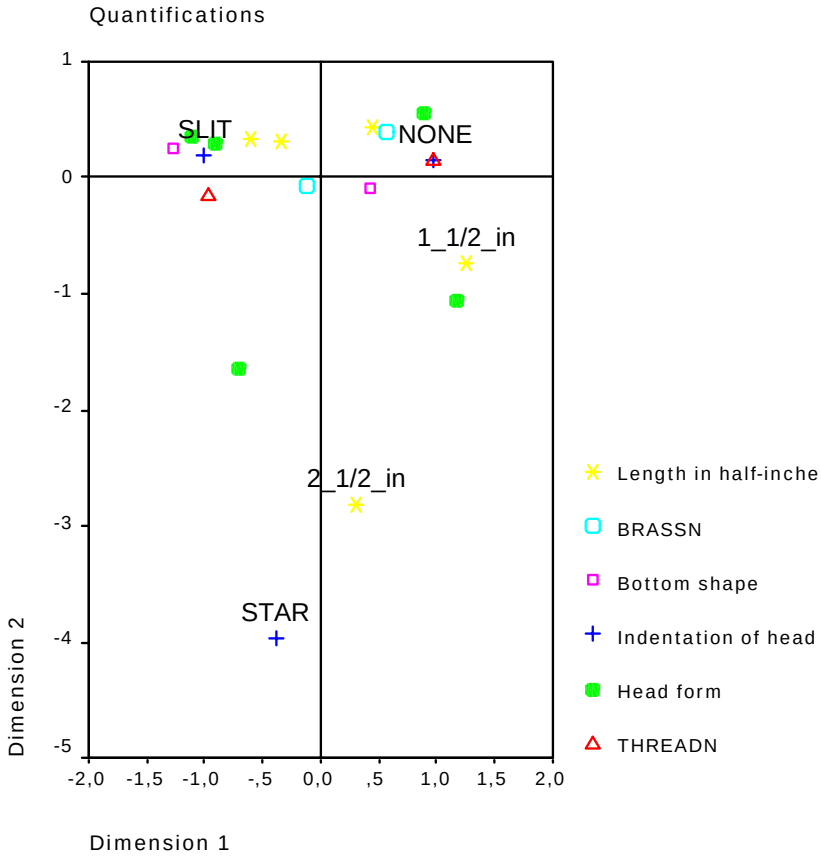


Figure 6 : *quantification des catégories.*

La dispersion des catégories d'une variable selon une dimension particulière reflète la variabilité de la configuration des réponses et constitue un indicateur de son pouvoir discriminant relatif à cette dimension.
Ainsi, selon l'axe horizontal *D1*, les catégories de la variable *THREADN* (codage numérique de la variable *thread*) sont très dispersées alors qu'elles ne le sont pas selon l'axe vertical *D2*. Il s'en suit que la variable *thread* discrimine mieux les objets selon la 1ère dimension que selon la 2nde dimension.
En revanche, les catégories de la forme de la tête (*Head form*) sont autant dispersées selon l'axe *D1* que selon l'axe *D2*. On en conclut que le pouvoir discriminant de cette variable est équivalent selon les deux dimensions.
Une variable dont les catégories sont plus dispersées selon une dimension possède un pouvoir discriminant plus important selon cette dimension qu'une autre variable dont les catégories sont projetées de façon moins dispersées. Par exemple, selon la 1ère dimension, les deux catégories de la variable *BRASSN* (codage numérique de la variable *brass* - caractère cuivré)

sont beaucoup moins dispersées que les deux catégories de la variable *THREADN*, indiquant que la variable *thread* possède un pouvoir discriminant plus important que celui de *brass* selon cette dimension (vérifiable en figure 5, d'après les niveaux relatifs de la mesure de discrimination des deux variables considérées).

2.5. Graphiques illustratifs

On peut éventuellement pousser plus loin l'analyse en consultant les différents graphiques illustratifs projetant individuellement, pour chaque variable, les objets étiquetés par le codage des catégories.

L'utilisation de ces variables illustratives montre que la 1ère dimension sépare parfaitement le groupe des articles possédant une pointe, étiquetés *Yes_Thread* et situés dans le demi-plan [*D1<0*], du groupe de ceux qui n'ont pas de pointe, étiquetés *No_Thread* et situés dans le demi-plan [*D1>0*].

Cette différenciation parfaite en fait un indicateur bien corrélé à la 1ère dimension.

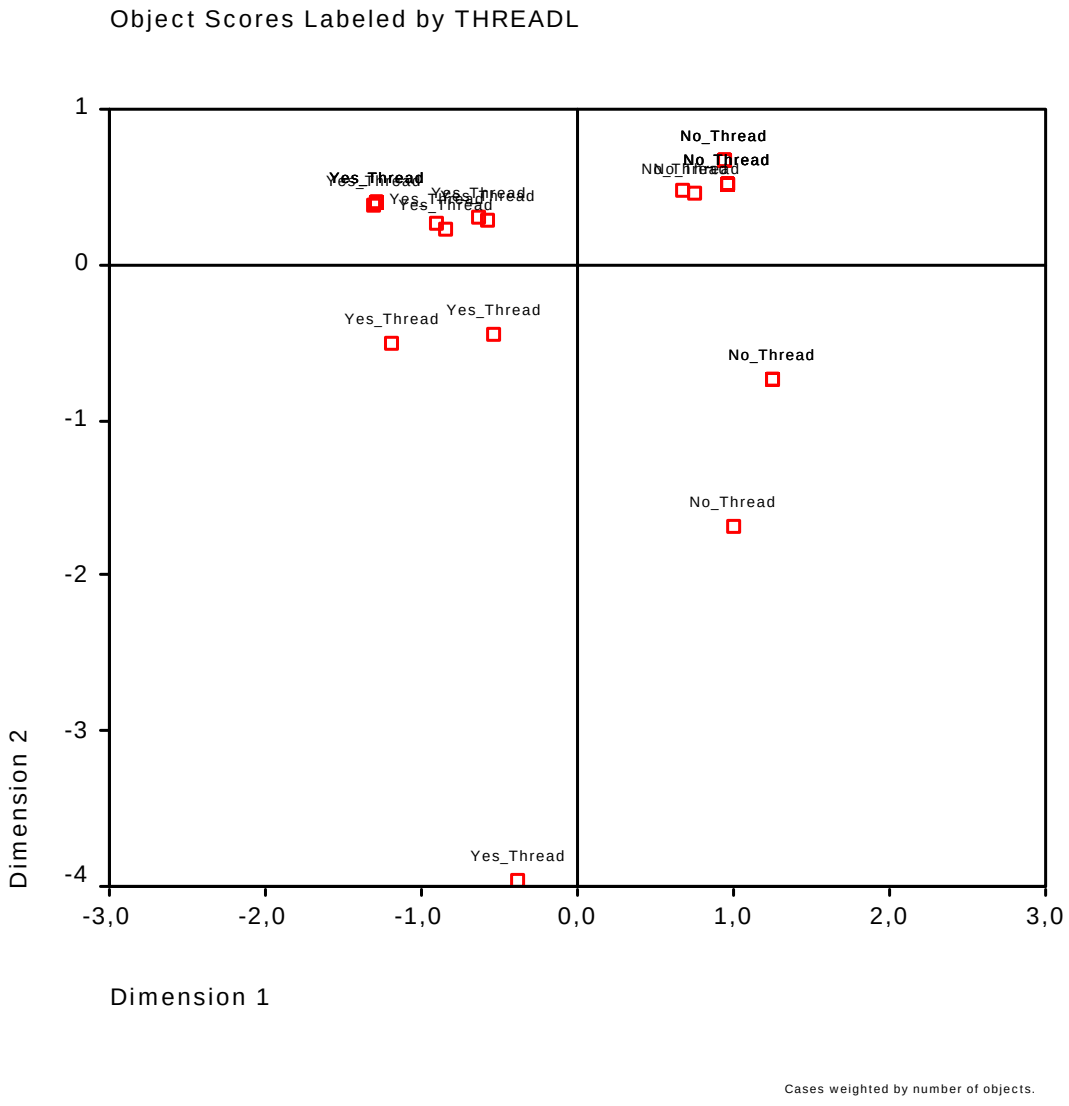


Figure 7 : *projection des objets, variable illustrative THREADL (« présence d'une pointe »).*

La projection des objets étiquetés par la forme de la tête (*Head form*) montre que celle-ci discrimine bien les articles dans les deux dimensions.
Les objets à tête plate (*FLAT*) sont situés dans le quadrant supérieur droit [*D2>0 & D1>0*] tandis que les articles dont la tête est en coupe (*CUP*) sont situés dans le quadrant inférieur droit [*D2<0 & D1>0*].
Les objets à tête conique (*CONE*) sont situés dans le quadrant inférieur gauche [*D2<0* & *D1<0*] mais on observe que ces objets sont beaucoup plus dispersés que dans les autres catégories.
Dans le quadrant supérieur gauche [D2>0 & D1<0], les objets à tête cylindrique (*CYLINDER*) ne peuvent être distingués des objets à tête ronde (*ROUND*).

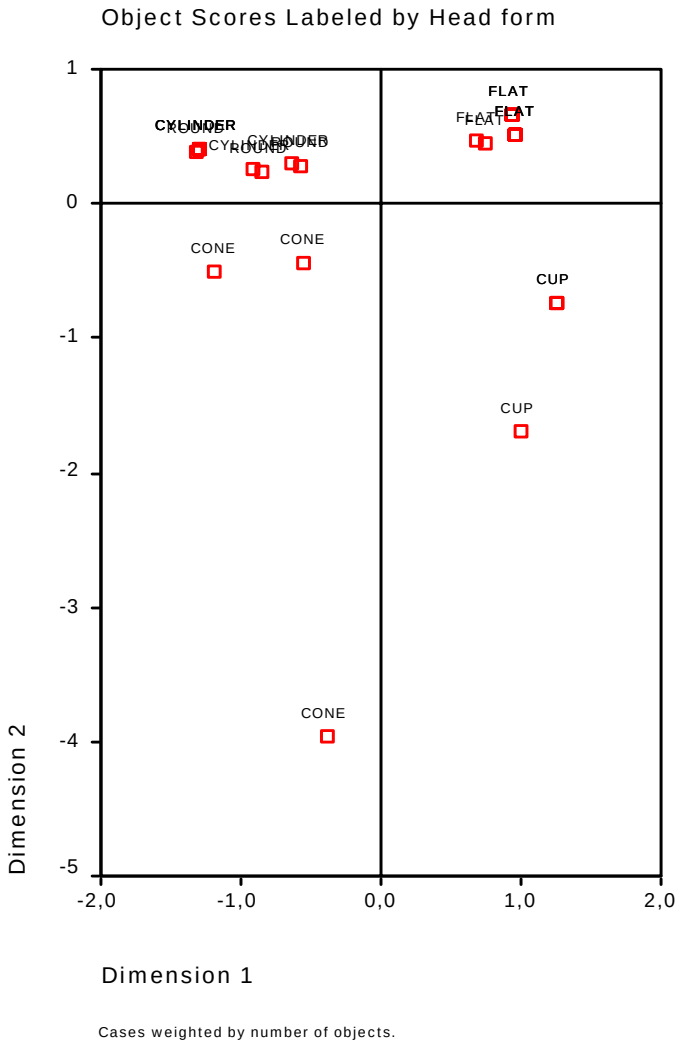


Figure 8 : *projection des objets, variable illustrative HEADL (« forme de la tête »).*

Le graphique selon les catégories de longueur montre que ces catégories se distinguent non pas selon l'axe horizontal du graphique mais plutôt selon l'axe vertical. Ce constat confirme l'analyse selon laquelle les catégories de la variable *length* ne discriminent pas les objets selon la 1ère dimension mais seulement selon la 2nde, les objets les plus courts étant situés dans le demi-plan [*D2>0*]

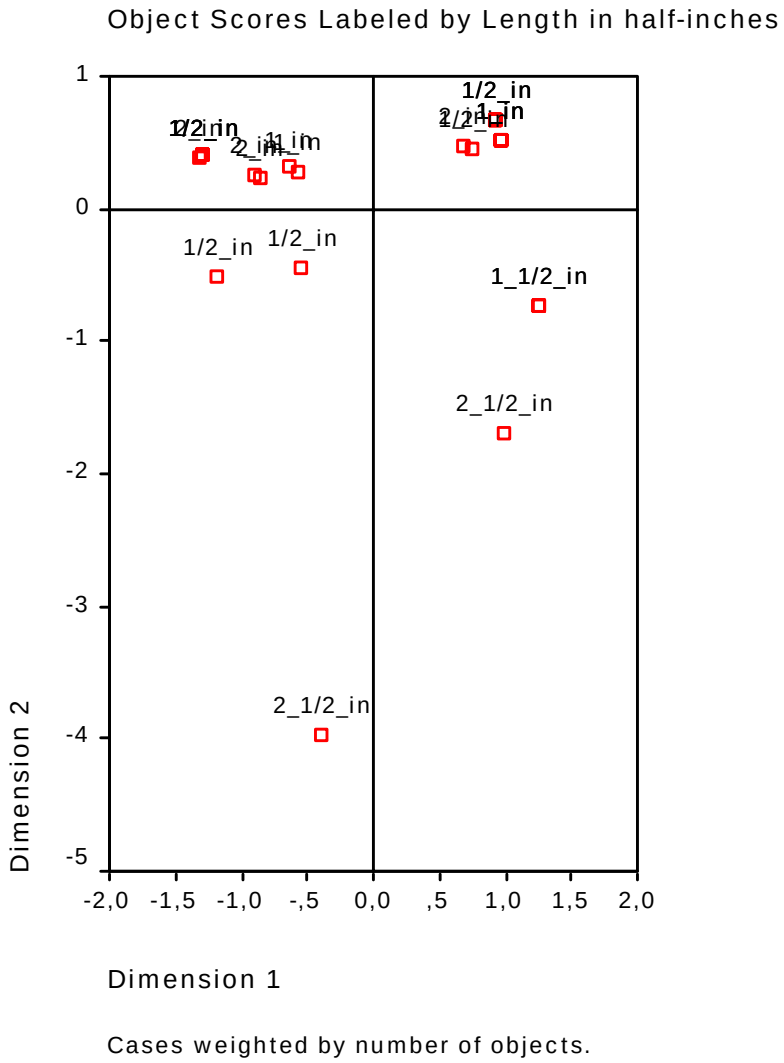


Figure 9 : *projection des objets, variable illustrative LENGHTL« longueur en pouces »*

Le graphique illustratif à partir de la variable *BRASS* (caractère cuivré ou non de l'objet) ne permet pas de mettre en évidence une différenciation nette des objets selon l'une ou l'autre des deux premières dimensions.

2.6. Filtrage des observations atypiques

Une fois identifiées les observations atypiques comportant trop de caractéristiques qui leur sont propres, on peut les exclure de l'analyse par filtrage, permettant ainsi de se focaliser sur les phénomènes dont l'occurrence n'est pas marginale. Si l'on réitère l'analyse d'homogénéité après un traitement excluant cette observation jugée atypique, on constate un léger changement au niveau des valeurs propres qui ne modifie pas de manière radicale l'ordre de grandeur de leur taux d'inertie. Pour autant, on ne doit pas conclure sans examen préalable à la quasi-équivalence des deux analyses

Le graphique des mesures de discrimination indique désormais que l'indentation de la tête (« *head indentation* ») ne discrimine plus les objets selon la 2nde dimension mais seulement selon la 1ère dimension, tandis que le caractère discriminant de la variable *brass* (cuivré ou non) se manifeste désormais selon la 2nde dimension. Les indices de discrimination des autres variables demeurent inchangés dans ces deux premières dimensions.

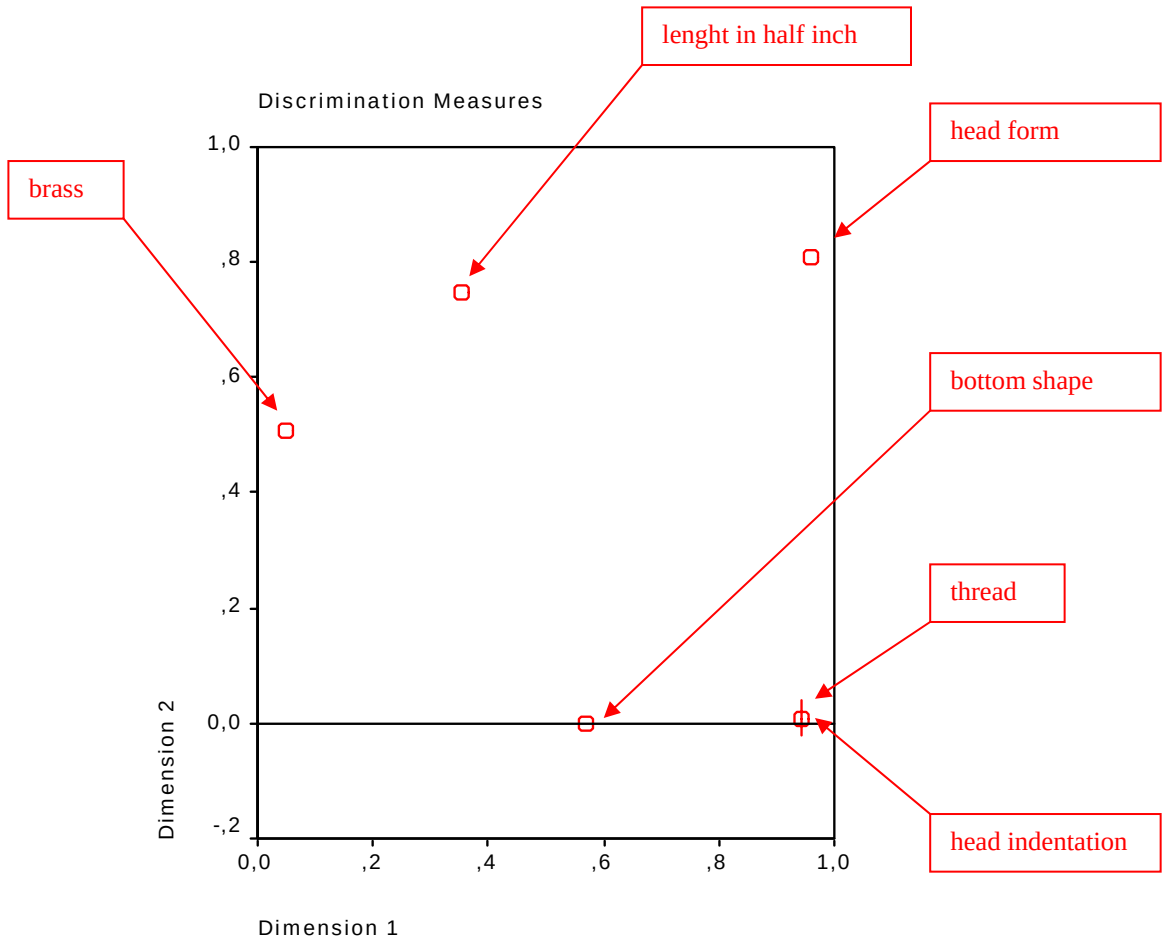


Figure 10 : *mesures de discrimination, après filtrage de l'objet atypique.*

Le graphique des objets étiquetés par la variable *brass* montre que les objets cuivrés (« *YES_Br* ») sont désormais projetés à l'extrémité négative de la 2nde dimension (zone [*-2<D2<-1*]) alors que les objets non cuivrés (« *Not_Br* ») sont projetés dans le demi-plan [*D2>-1*], confirmant ainsi le pouvoir discriminant de la variable *brass* selon la 2nde dimension.

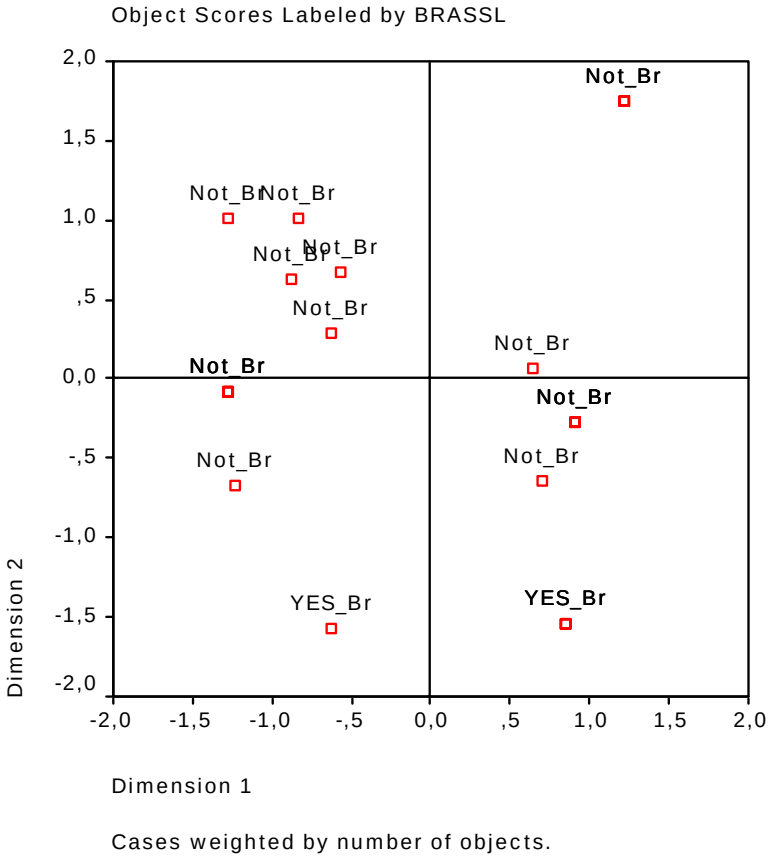
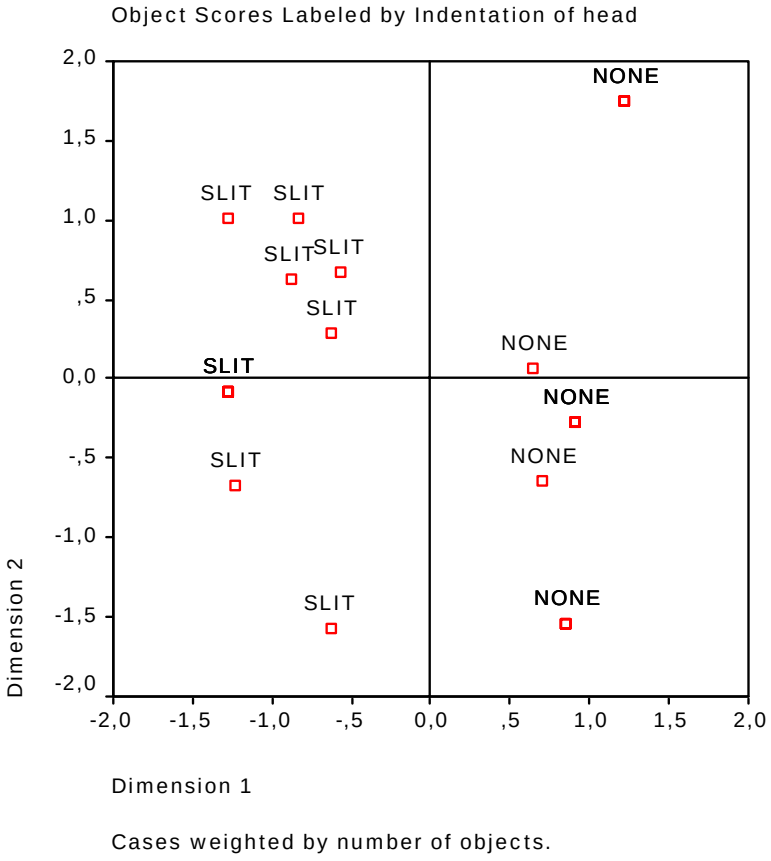


Figure 11 : *projection des objets étiquetés par BRASSL, après filtrage de l'objet atypique*

La projection illustrative des objets étiquetés par les catégories relatives à l'indentation de la tête (« *Indentation of head* ») montre que la première dimension permet de discriminer parfaitement les objets non indentés (« *NONE* ») des objets indentés (« *SLIT* »), comme dans l'analyse précédente.
Cependant, la 2nde dimension ne discrimine plus les catégories d'indentation, à l'inverse de l'analyse précédente.



.

Figure 12 : *projection des objets étiquetés par indentation de la tête (« INDHEADL»), après filtrage de l'objet atypique*

3. L'ANALYSE D'HOMOGENEITE, POUR UNE REPRESENTATION OPTIMALE DES CATEGORIES.

3.1. Le concept d'homogénéité

Développée par le groupe Albert Gifi[4], la procédure *HOMALS* se base sur le concept d'**homogénéité**, que l'on peut définir de la manière suivante.

Soit le vecteur $\mathbf{z}_j$, $j = 1, \cdots, p$, contenant les observations faites sur les *n* individus d'une population, correspondant à la variable Z_j.

Le vecteur $\mathbf{z}_j$ est ***homogène à*** $\mathbf{x}$, vecteur unitaire (de norme 1), si et seulement si après une transformation t_j de normalisation (tel que $\|t_j(\mathbf{z}_j)\| = 1$), on a $\mathbf{x} = t_j(\mathbf{z}_j)$.

Si le vecteur $\mathbf{z}_j$ n'est pas homogène à $\mathbf{x}$, on définit la ***perte d'homogénéité*** comme suit : $\sigma^2(\mathbf{x}, t) = \frac{1}{p}\sum_{j=1}^{p} {}^t(\mathbf{x} - t_j(\mathbf{z}_j))(\mathbf{x} - t_j(\mathbf{z}_j))$.

3.2. La procédure HOMALS

Soit la matrice $\mathbf{Z}_j$ des indicatrices de codage correspondant aux indicatrices de codage d'une variable Z_j qualitative à *q* modalités. La transformation t_j du vecteur $\mathbf{z}_j$ peut être définie par $t_j(\mathbf{z}_j) = \mathbf{Z}_j\mathbf{Y}_j$ où $\mathbf{Y}_j$ est une matrice à $n \times q$ coefficients.

La procédure *HOMALS* consiste à minimiser la fonction de perte suivante :

$$\sigma^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{p}\sum_{j=1}^{p} trace\left[{}^t(\mathbf{X} - \mathbf{Z}_j\mathbf{Y}_j)(\mathbf{X} - \mathbf{Z}_j\mathbf{Y}_j)\right]$$

sous les contraintes d'orthonormalisation ${}^t\mathbf{X}\mathbf{X} = nI$ et de centrage $\mathbf{1}\mathbf{X} = 0$.

3.3. Equivalence avec l'analyse des correspondances multiples

[Gifi, 1990] montre que l'analyse d'homogénéité peut être considérée comme la résolution d'un problème de décomposition spectrale, soit en valeurs singulières, soit en valeur propres et donc fournir les facteurs d'une analyse des correspondances. [Tenenhaus et Young, 1985] établit les relations entre analyse des correspondances multiples et analyse d'homogénéité en montrant l'équivalence entre les deux méthodes, l'analyse d'homogénéité pouvant être vue comme une technique de positionnement multidimensionnel restituant une image euclidienne (à partir de graphiques-plans) des « dissimilarités » constituées par les distances du Khi-Deux entre profils-lignes.

4. EFFECTUER UNE ANALYSE D'HOMOGENEITE AVEC *SPSS*

Pour obtenir une analyse d'homogénéité sous *SPSS*, il convient de créer par recodage, à partir du tableau des données alphanumériques (cf. figure 2), un tableau numérique comportant l'ensemble des variables à analyser. Pour ce faire, il faut utiliser la procédure de recodage automatique <*Automatic Recode*> du menu de transformation <*Transform*>, créant ainsi la variable *threadn* (codage numérique) à partir de la variable *thread* (codage alphanumérique) par transformation des catégories prises dans un ordre lexicographique croissant (cf. figure 13).

[4] Albert Gifi est le nom collectif des membres du *Department of Data Theory* de l'Université de Leiden (Pays-Bas). Ce groupe, constitué autour de Jan de Leeuw a mis au point un système pour l'analyse multivariée non linéaire qui recouvre de multiples techniques factorielles allant de l'analyse en composantes principales à l'analyse canonique. Le travail de ce groupe est présenté dans l'ouvrage [Gifi, 1990]

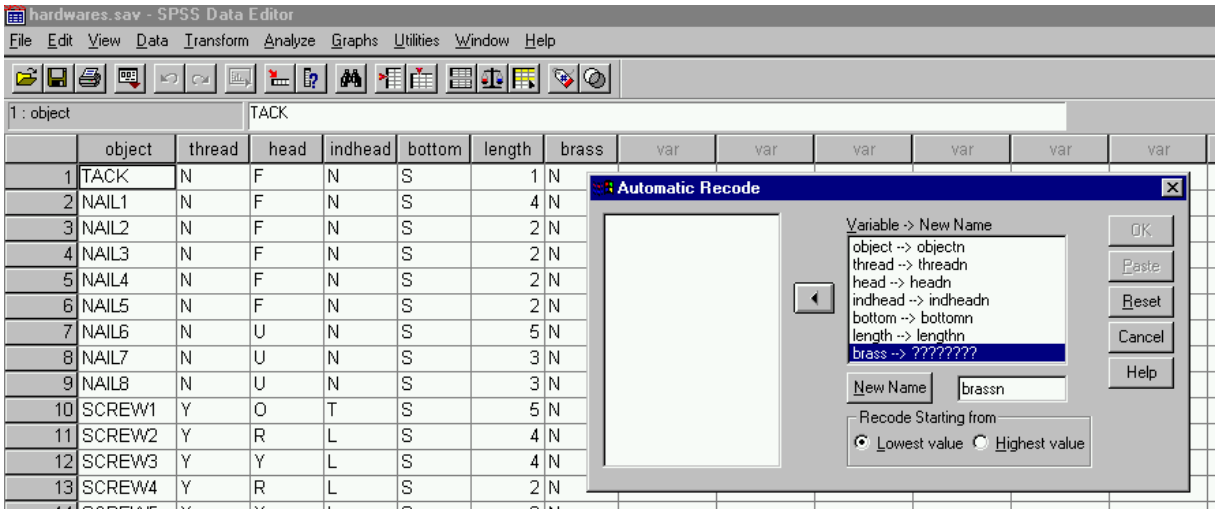


Figure 13 : *recodage des variables alphanumériques en variables numériques***.**

hardwares.sav - SPSS Data Editor

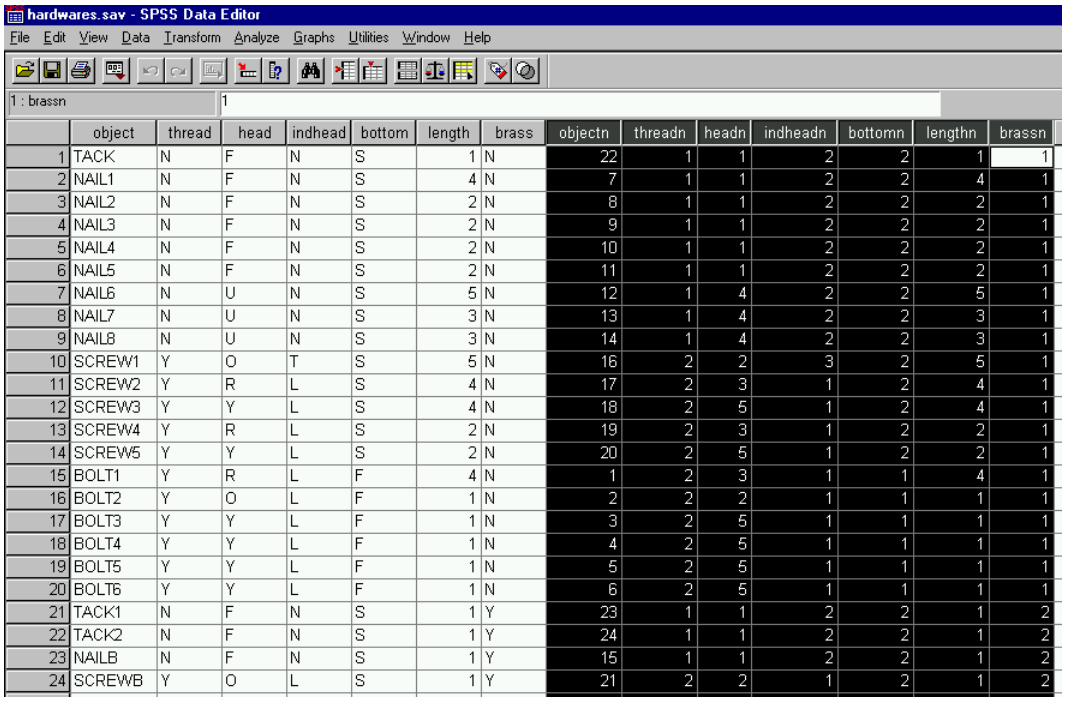
File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : objectn 22

	object	thread	head	indhead	bottom	length	brass	objectn	threadn	headn	indheadn	bottomn	lengthn	brassn
1	TACK	N	F	N	S	1	N	22	1	1	2	2	1	1
2	NAIL1	N	F	N	S	4	N	7	1	1	2	2	4	1
3	NAIL2	N	F	N	S	2	N	8	1	1	2	2	2	1
4	NAIL3	N	F	N	S	2	N	9	1	1	2	2	2	1
5	NAIL4	N	F	N	S	2	N	10	1	1	2	2	2	1
6	NAIL5	N	F	N	S	2	N	11	1	1	2	2	2	1
7	NAIL6	N	U	N	S	5	N	12	1	4	2	2	5	1
8	NAIL7	N	U	N	S	3	N	13	1	4	2	2	3	1
9	NAIL8	N	U	N	S	3	N	14	1	4	2	2	3	1
10	SCREW1	Y	O	T	S	5	N	16	2	2	3	2	5	1
11	SCREW2	Y	R	L	S	4	N	17	2	3	1	2	4	1
12	SCREW3	Y	Y	L	S	4	N	18	2	5	1	2	4	1
13	SCREW4	Y	R	L	S	2	N	19	2	3	1	2	2	1
14	SCREW5	Y	Y	L	S	2	N	20	2	5	1	2	2	1
15	BOLT1	Y	R	L	F	4	N	1	2	3	1	1	4	1
16	BOLT2	Y	O	L	F	1	N	2	2	2	1	1	1	1
17	BOLT3	Y	Y	L	F	1	N	3	2	5	1	1	1	1
18	BOLT4	Y	Y	L	F	1	N	4	2	5	1	1	1	1
19	BOLT5	Y	Y	L	F	1	N	5	2	5	1	1	1	1
20	BOLT6	Y	Y	L	F	1	N	6	2	5	1	1	1	1
21	TACK1	N	F	N	S	1	Y	23	1	1	2	2	1	2
22	TACK2	N	F	N	S	1	Y	24	1	1	2	2	1	2
23	NAILB	N	F	N	S	1	Y	15	1	1	2	2	1	2
24	SCREWB	Y	O	L	S	1	Y	21	2	2	1	2	1	2

Figure 14 : *variables numériques recodées***.**

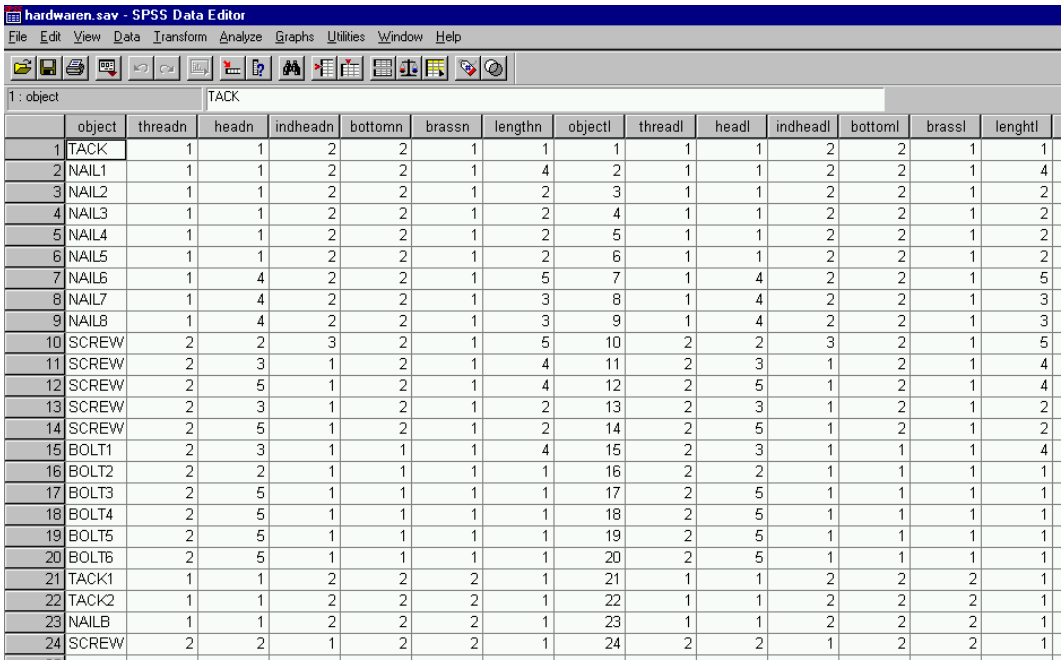
Dans une seconde étape, il faut créer par recopie autant de variables illustratives qu'il y a de critères participant à l'analyse. Pour ce faire, il suffit de sélectionner les variables recodées en cliquant avec la touche « *Control* » maintenue enfoncée (« *Ctrl+Clic* ») sur les colonnes correspondantes de l'éditeur des données (cf. figure 15).



	object	thread	head	indhead	bottom	length	brass	objectn	threadn	headn	indheadn	bottomn	lengthn	brassn
1	TACK	N	F	N	S	1	N	22	1	1	2	2	1	1
2	NAIL1	N	F	N	S	4	N	7	1	1	2	2	4	1
3	NAIL2	N	F	N	S	2	N	8	1	1	2	2	2	1
4	NAIL3	N	F	N	S	2	N	9	1	1	2	2	2	1
5	NAIL4	N	F	N	S	2	N	10	1	1	2	2	2	1
6	NAIL5	N	F	N	S	2	N	11	1	1	2	2	2	1
7	NAIL6	N	U	N	S	5	N	12	1	4	2	2	5	1
8	NAIL7	N	U	N	S	3	N	13	1	4	2	2	3	1
9	NAIL8	N	U	N	S	3	N	14	1	4	2	2	3	1
10	SCREW1	Y	O	T	S	5	N	16	2	2	3	2	5	1
11	SCREW2	Y	R	L	S	4	N	17	2	3	1	2	4	1
12	SCREW3	Y	Y	L	S	4	N	18	2	5	1	2	4	1
13	SCREW4	Y	R	L	S	2	N	19	2	3	1	2	2	1
14	SCREW5	Y	Y	L	S	2	N	20	2	5	1	2	2	1
15	BOLT1	Y	R	L	F	4	N	1	2	3	1	1	4	1
16	BOLT2	Y	O	L	F	1	N	2	2	2	1	1	1	1
17	BOLT3	Y	Y	L	F	1	N	3	2	5	1	1	1	1
18	BOLT4	Y	Y	L	F	1	N	4	2	5	1	1	1	1
19	BOLT5	Y	Y	L	F	1	N	5	2	5	1	1	1	1
20	BOLT6	Y	Y	L	F	1	N	6	2	5	1	1	1	1
21	TACK1	N	F	N	S	1	Y	23	1	1	2	2	1	2
22	TACK2	N	F	N	S	1	Y	24	1	1	2	2	1	2
23	NAILB	N	F	N	S	1	Y	15	1	1	2	2	1	2
24	SCREWB	Y	O	L	S	1	Y	21	2	2	1	2	1	2

Figure 15 : *sélection multiple par Ctrl+Clic des variables numériques recodées***.**

Ensuite, il faut sélectionner à partir du menu <*Edit*>, la commande <*Copy*> (avec le clavier, faire un <*Ctrl+C*>), pour pouvoir coller (menu <*Edit*>, commande <*Paste*>, ou équivalent-clavier faire un <*Ctrl+V*>), après avoir effectué une sélection multiple de cinq colonnes vides :



	object	threadn	headn	indheadn	bottomn	brassn	lengthn	objectl	threadl	headl	indheadl	bottoml	brassl	lenghtl
1	TACK	1	1	2	2	1	1	1	1	1	2	2	1	1
2	NAIL1	1	1	2	2	1	4	2	1	1	2	2	1	4
3	NAIL2	1	1	2	2	1	2	3	1	1	2	2	1	2
4	NAIL3	1	1	2	2	1	2	4	1	1	2	2	1	2
5	NAIL4	1	1	2	2	1	2	5	1	1	2	2	1	2
6	NAIL5	1	1	2	2	1	2	6	1	1	2	2	1	2
7	NAIL6	1	4	2	2	1	5	7	1	4	2	2	1	5
8	NAIL7	1	4	2	2	1	3	8	1	4	2	2	1	3
9	NAIL8	1	4	2	2	1	3	9	1	4	2	2	1	3
10	SCREW	2	2	3	2	1	5	10	2	2	3	2	1	5
11	SCREW	2	3	1	2	1	4	11	2	3	1	2	1	4
12	SCREW	2	5	1	2	1	4	12	2	5	1	2	1	4
13	SCREW	2	3	1	2	1	2	13	2	3	1	2	1	2
14	SCREW	2	5	1	2	1	2	14	2	5	1	2	1	2
15	BOLT1	2	3	1	1	1	4	15	2	3	1	1	1	4
16	BOLT2	2	2	1	1	1	1	16	2	2	1	1	1	1
17	BOLT3	2	5	1	1	1	1	17	2	5	1	1	1	1
18	BOLT4	2	5	1	1	1	1	18	2	5	1	1	1	1
19	BOLT5	2	5	1	1	1	1	19	2	5	1	1	1	1
20	BOLT6	2	5	1	1	1	1	20	2	5	1	1	1	1
21	TACK1	1	1	2	2	2	1	21	1	1	2	2	2	1
22	TACK2	1	1	2	2	2	1	22	1	1	2	2	2	1
23	NAILB	1	1	2	2	2	1	23	1	1	2	2	2	1
24	SCREW	2	2	1	2	2	1	24	2	2	1	2	2	1

Figure 16 : *fichier des variables numériques, actives et illustratives***.**

Pour obtenir une analyse d'homogénéité, il faut sélectionner à partir du menu <*Analyse*>, la procédure <*Optimal Scaling*> du menu <*Data Reduction*>, en choisissant les options correspondantes (options par défaut de la procédure, soit un seul ensemble de variables avec toutes les variables considérées comme nominales) :

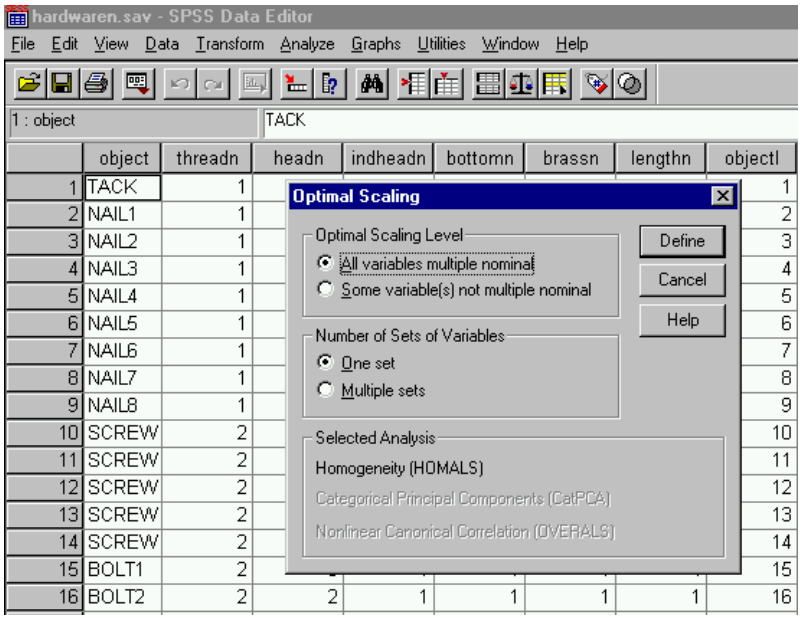


Figure 17 : *options correspondant à l'analyse d'homogénéité*

La première étape de la spécification de la procédure consiste à sélectionner les variables actives de l'analyse *(threadn, headn, indheadn, bottomn, brassn, lenghtn*) en définissant pour chacune d'entre-elles le nombre de modalités :

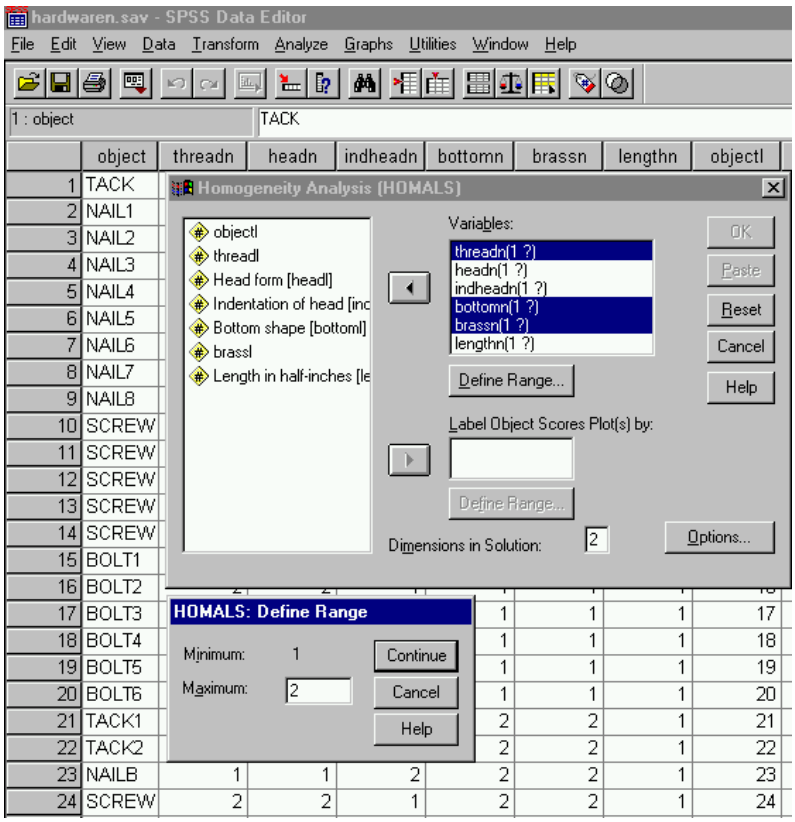


Figure 18 : *spécification des variables actives***.**

Dans la seconde étape, on spécifie les variables illustratives de l'analyse *(objectl, threadl, headl, brassl, lenghtl*) en définissant également pour chacune d'entre-elles le nombre de modalités :

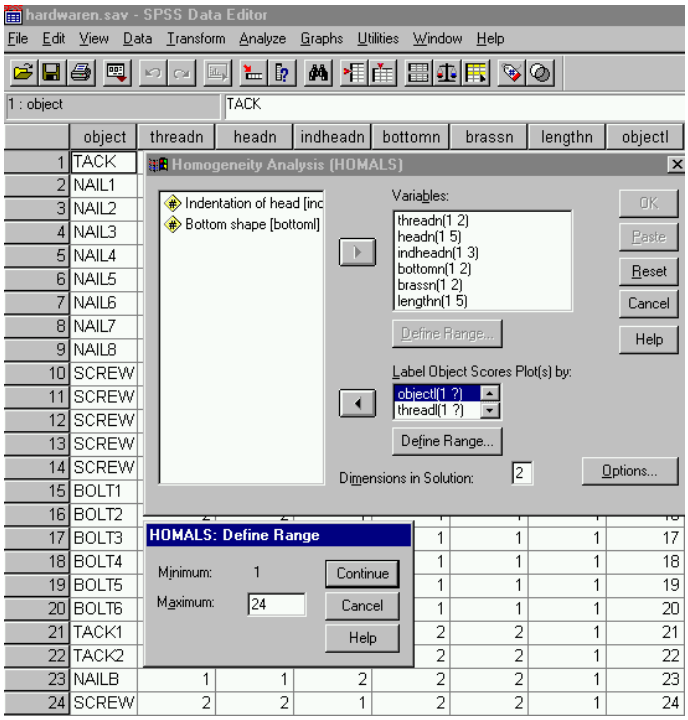


Figure 19 : *spécification des variables illustratives*

La dernière étape de cette spécification concerne le choix du nombre de dimensions (nombre d'axes factoriels) choisies pour la représentation graphique des objets, des modalités et des variables. On choisit ici une représentation graphique en deux dimensions comme solution particulière au problème d'optimisation sous contraintes que pose l'analyse formulée en terme d'homogénéité (cf. §3).
Les différentes options de traitement peuvent être choisies en utilisant le bouton <*Options…*>. Ces options portent sur les résultats (*Display*), les graphiques (*Plot*), la sauvegarde des coordonnées factorielles des objets (<*Save object scores*>) et les critères de contrôle de l'algorithme (*Criteria*).

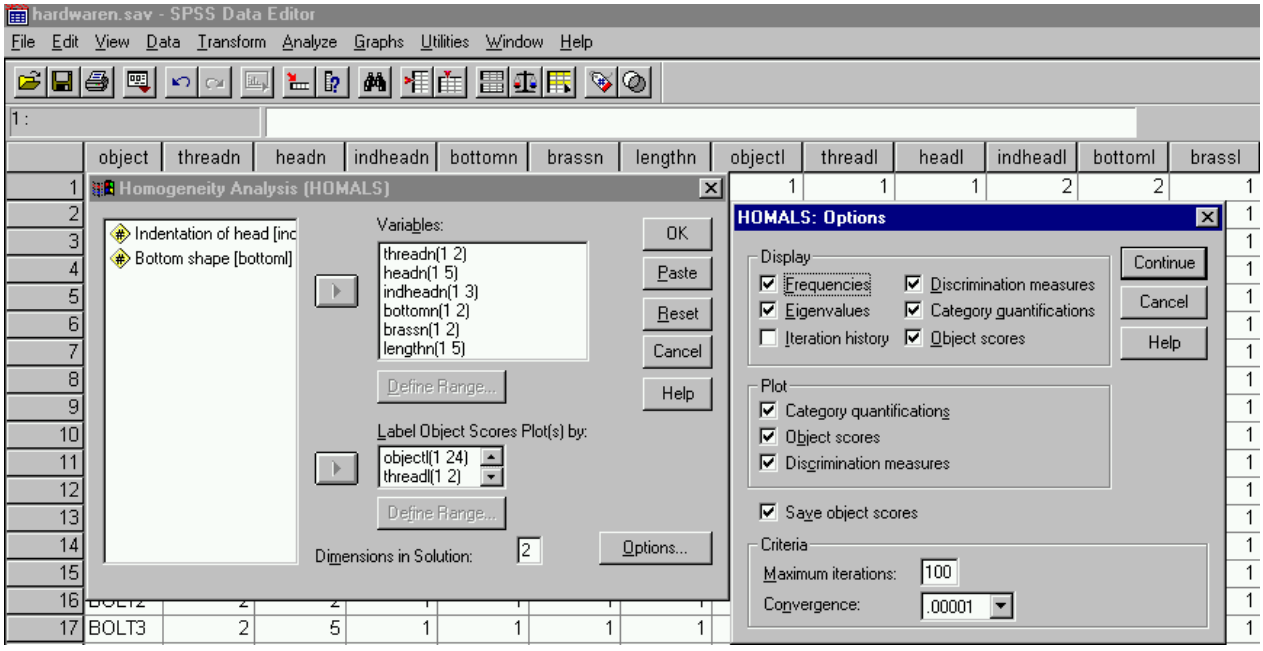


Figure 20 : *choix des options***.**

Les résultats demandés (cf. section *Display* de la figure 20) sont les distributions marginales obtenues par comptage (*Frequencies*), les valeurs propres (*Eigenvalues*), le pouvoir discriminant des variables actives (*Discrimination measures*), les coordonnées factorielles des modalités pour chaque variable (*Category quantifications*), les coordonnées factorielles des objets (*Object scores*).

Les graphiques demandés (cf. section *Plot* de la figure 20) sont le graphique factoriel des modalités de variables actives (*Category quantifications*), celui des objets (Object scores) et le diagramme du pouvoir discriminant des variables selon chacune des dimensions (*Discrimination measures*). A ces graphiques s'ajoutent autant de graphiques de densité des objets étiquetés par les modalités qu'il y a de variables illustratives.

La sauvegarde des coordonnées factorielles demandée (*Save object scores*) s'effectue dans le fichier d'origine, mais peut être ultérieurement sauvegardé dans un fichier spécifique, comme suit, pour de nouvelles analyses (classification sur axes factoriels) :

hardware_2f.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Ut

17 :

	object	hom1_1	hom2_1
1	TACK	,75	,46
2	NAIL1	,68	,47
3	NAIL2	,96	,52
4	NAIL3	,96	,52
5	NAIL4	,96	,52
6	NAIL5	,96	,52
7	NAIL6	1,00	-1,69
8	NAIL7	1,25	-,74
9	NAIL8	1,25	-,74
10	SCREW1	-,38	-3,96
11	SCREW2	-,85	,23
12	SCREW3	-,91	,26
13	SCREW4	-,57	,28
14	SCREW5	-,63	,31
15	BOLT1	-1,31	,38
16	BOLT2	-1,18	-,51
17	BOLT3	-1,30	,40
18	BOLT4	-1,30	,40
19	BOLT5	-1,30	,40
20	BOLT6	-1,30	,40
21	TACK1	,93	,67
22	TACK2	,93	,67
23	NAILB	,93	,67
24	SCREWB	-,54	-,45

Figure 21 : *sauvegarde des coordonnées factorielle des objets dans un fichier spécifique***.**

Les macro-instructions du programme *SPSS* correspondant aux options précédemment définies peuvent être sauvegardées dans un fichier de syntaxe en utilisant le bouton <*Paste*> de la boîte de dialogue :

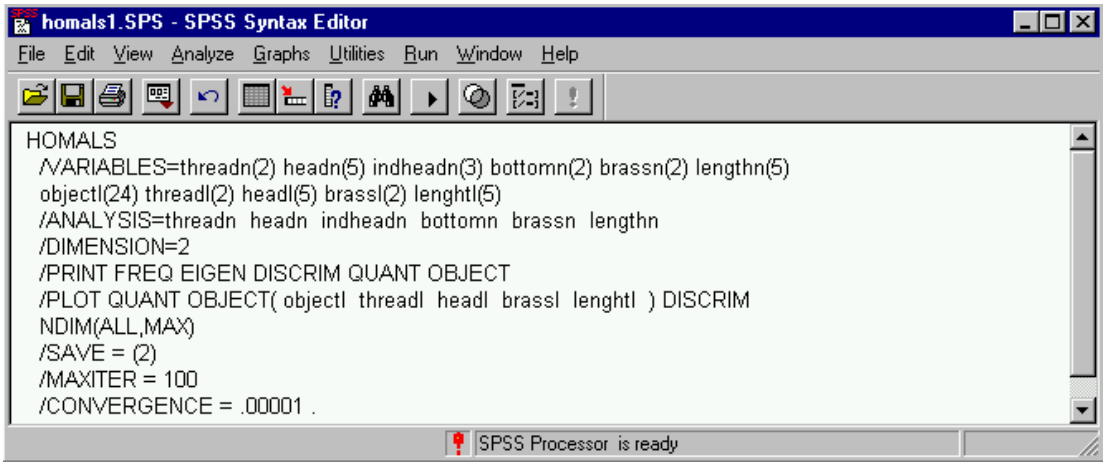


Figure 22 : *sauvegarde des macro-instructions dans un fichier programme (extension « .SPS »).*

Le seuil de convergence (*Convergence=.00001*) et le nombre maximum d'itérations (*Maximum interations=100*) permettent de contrôler l'algorithme itératif des moindres carrés alternés de la procédure *HOMALS* dans la recherche d'une solution.

Iteration History

Iteration	Fit	Difference from the Previous Iteration
1	,132757	,132757
2	,849876	,717119
3	,943649	,093773
4	,966800	,023151
5	,976822	,010022
6	,982110	,005288
7	,985104	,002993
8	,986838	,001735
9	,987851	,001013
10	,988444	,000593
11	,988793	,000349
12	,988999	,000206
13	,989122	,000123
14	,989196	,000074
15	,989241	,000045
16	,989269	,000028
17	,989287	,000018
18	,989298	,000012
19[a]	,989306	,000008

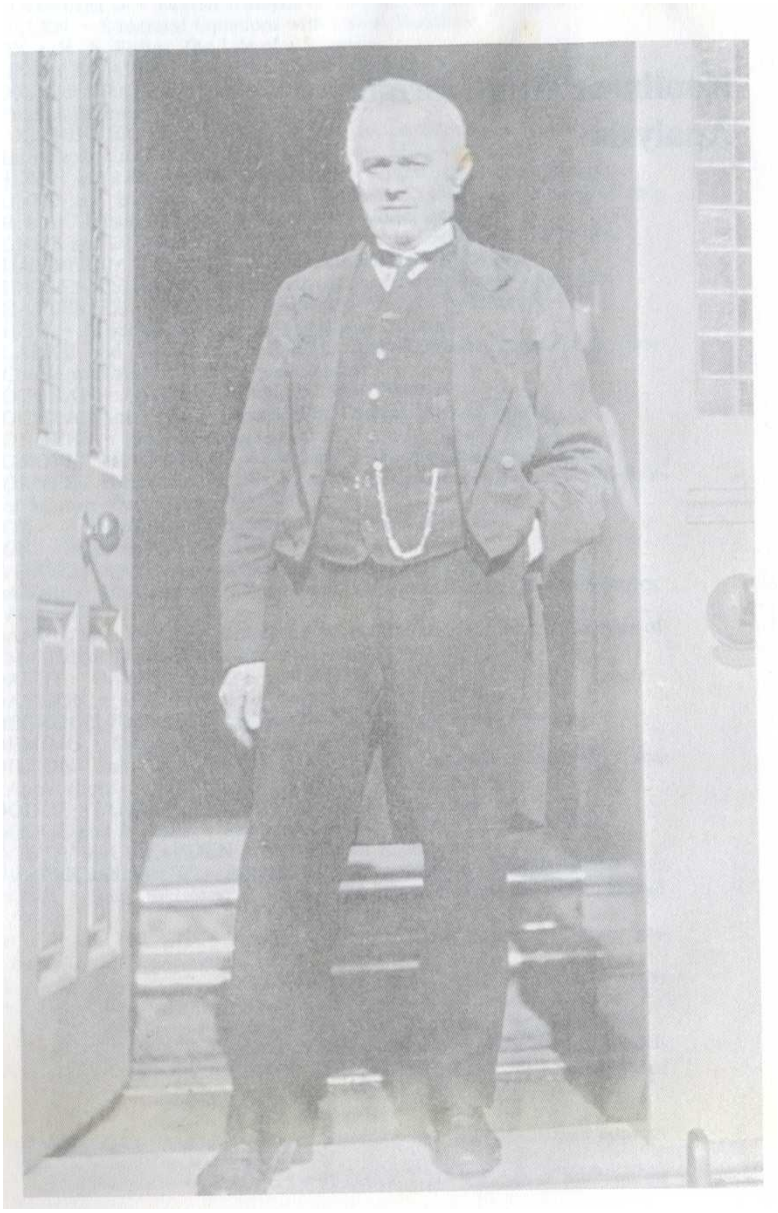
a. The iteration process stopped because the convergence test value was reached.

Tableau 4 : *historique des itérations*

Dans cet exemple, l'algorithme s'arrête à l'itération n° *19* car l'amélioration de l'indice d'ajustement (*Fit*) est devenue inférieure à la valeur du seuil de convergence.

5. RÉFÉRENCES

Gifi A. (1990) *Nonlinear Multivariate Analysis*, Wiley, 579 p.
Hartigan J.A. (1975) *Clustering Algorithms*. Wiley, New York, 351 p.
Meulman J.H., Heiser J.H. (2001*) SPSS Categories 11.0,* SPSS Inc., Chicago, 330 p.
Tenenhaus M., Young F.W. (1985) « An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis, and other methods for quantifying categorical multivariate data », *Psychometrika*, **50**, pp. 91-119.
SPSS (1994) *SPSS 6.1 Categories*, SPSS Inc., Chicago, 209 p.



Portrait d'Albert Gifi ([Gifi, 1990])