

UNIWIN VERSION 9.7.0

ANALYSE EN COMPOSANTES PRINCIPALES SYMBOLIQUE (METHODE DES CENTRES)

Révision : 02/09/2023

Définition.....	1
Entrée des données	2
Données manquantes	4
Exemple 1 : Fichier HUILES (données intervalles)	4
L'option Rapports	7
L'option Graphiques	8
Une rapide interprétation des résultats.....	11
Exemple 2 : Fichier HUILES (données individus x variables)	12
Exemple 3 : Fichier BATS (données intervalles)	13
Les variables internes créées par la procédure	16
Références	17

Définition

Les méthodes classiques d'analyse factorielle ne sont applicables qu'à des objets caractérisés par des variables monovaluées (la valeur prise par une variable pour un objet est une valeur unique).

L'ACPS est une extension de l'ACP à des objets caractérisés par des variables multivaluées décrivant de la variation ou de l'imprécision (la valeur prise par une variable pour un objet est un intervalle de valeurs).

Prenons trois exemples :

En botanique si les objets à étudier sont des plantes, la taille de la tige d'une plante est une valeur unique. Par contre, si les objets auxquels on s'intéresse sont des espèces de plantes (concepts), la taille de la tige d'une espèce définit un intervalle de valeurs. Cet intervalle représente le domaine de variation de la taille de la tige sur tous les spécimens appartenant à l'espèce en question.

En météorologie, les températures quotidiennes enregistrées en valeurs minimales et maximales offrent une vision plus réaliste des variations des conditions météorologiques par rapport aux valeurs moyennes simples.

En finance, les prix de transaction minimum et maximum, relevés quotidiennement pour un ensemble d'actions, représentent une information plus pertinente pour les experts afin d'évaluer la tendance et la volatilité des actions dans la même journée.

La procédure ACPS proposée (méthode des centres) accepte deux structures de données en entrée :

1. Un tableau de données individus x variables qui sera transformé en un tableau de données (concepts) contenant les intervalles des objets symboliques par utilisation de variables qualitatives.
2. Un tableau de données (concepts) contenant les intervalles des objets symboliques.

Les données sont automatiquement centrées et réduites.

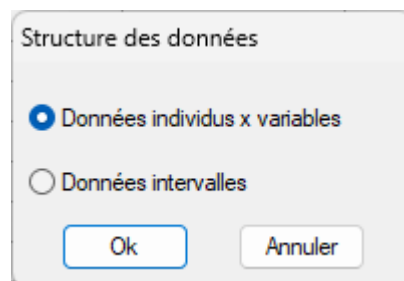
Le rapport affiche les composantes principales (approches classique et symbolique), les coordonnées des variables (approches classique et symbolique), les cosinus carrés et les contributions des concepts et des variables ainsi que les distances carrées des concepts à l'origine et les contributions des concepts à l'inertie totale.

Les graphiques proposés sont : diagramme des inerties (approche classique), cercle des corrélations des variables (approche classique), plan factoriel des variables (approche symbolique), plan factoriel des concepts (approches classique et symbolique).

Entrée des données

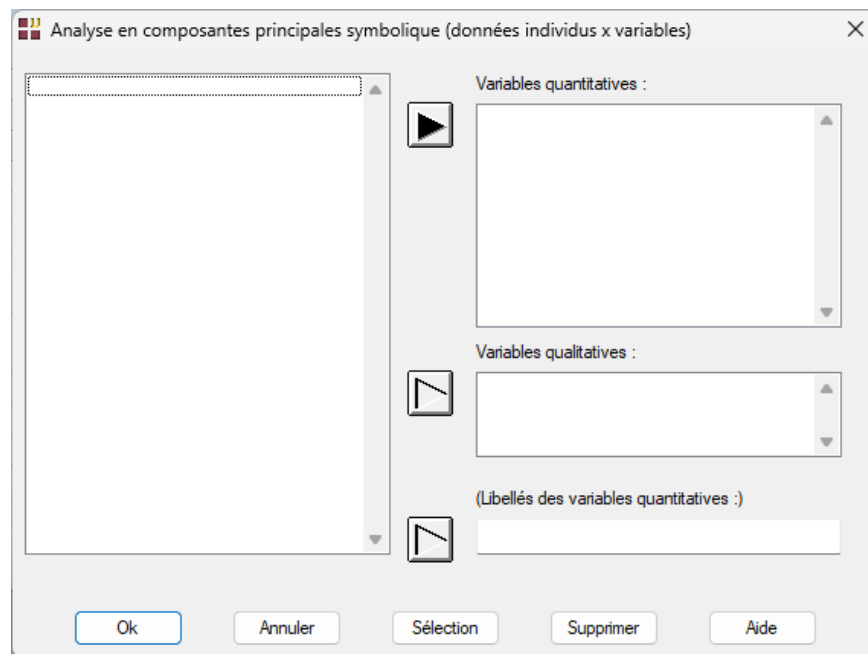
Cliquons sur l'icône ACPS dans le ruban Décrire.

La première boîte de dialogue affichée permet de préciser la structure des données à analyser :



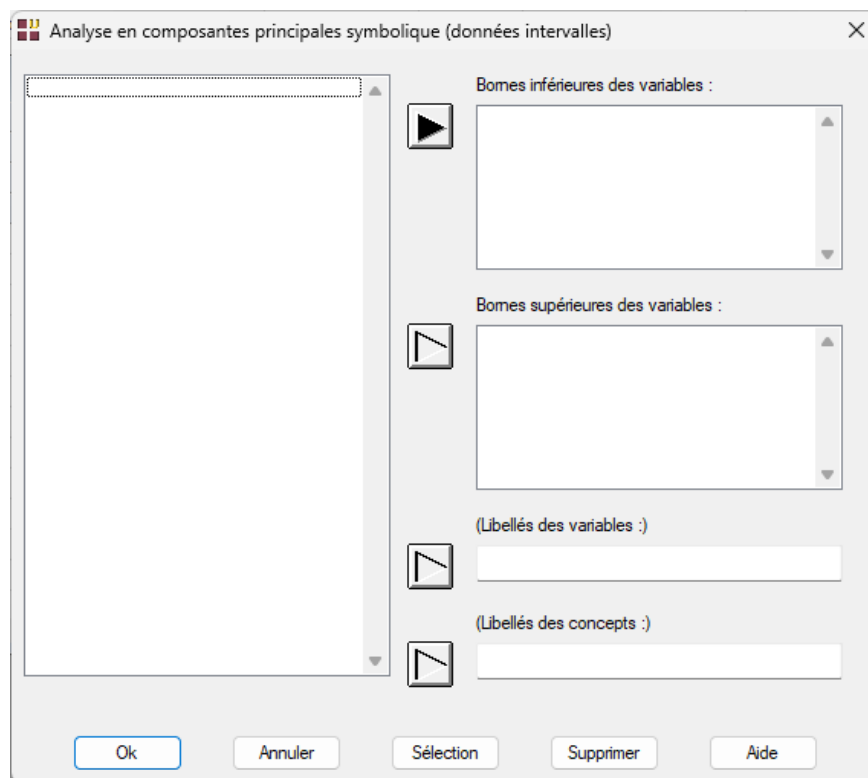
La seconde boîte de dialogue qui s'affiche ensuite dépend de la structure des données.

Données individus x variables



- Variables quantitatives : les variables contenant les données des individus.
- Variables qualitatives : les variables agrégeant les données des individus pour former les concepts (données intervalles).
- Libellés des variables quantitatives : les libellés des variables quantitatives.

Données intervalles



- Bornes inférieures des intervalles : les bornes inférieures des intervalles des variables.
- Bornes supérieures des intervalles : les bornes supérieures des intervalles des variables (dans le même ordre que pour les bornes inférieures).
- Libellés des variables : les libellés des variables.
- Libellés des concepts : les libellés des concepts.

Données manquantes

Les données manquantes ne sont pas autorisées dans cette procédure.

Exemple 1 : Fichier HUILES (données intervalles)

Nous utiliserons le fichier HUILES (Ichino) pour illustrer ce premier exemple.

Ce fichier contient des informations collectées sous forme d'intervalles concernant huit huiles (les concepts à étudier) : Linseed Oil (huile de lin), Perilla Oil (huile de périlla), Cottonseed Oil (huile de coton), Sesame Oil (huile de sésame), Camellia Oil (huile de camélia), Olive Oil (huile d'olive), Beef Tallow (suif), Hog Fat (saindoux).

GRAmin	FREmin	IODmin	SAPmin	GRAmax	FREmax	IODmax	SAPmax
Specific Gravity	Freezing Point	Iodine Value	Saponification Value	Specific Gravity	Freezing Point	Iodine Value	Saponification Value
Type = Numérique	Type = Numérique	Type = Numérique	Type = Numérique	Type = Numérique	Type = Numérique	Type = Numérique	Type = Numérique
Longueur = 8	Longueur = 8	Longueur = 8	Longueur = 8	Longueur = 8	Longueur = 8	Longueur = 8	Longueur = 8
0,930	-27	170	118	0,935	-18	204	196
0,930	-5	192	188	0,937	-4	208	197
0,916	-6	99	189	0,918	-1	113	198
0,920	-6	104	187	0,926	-4	116	193
0,916	-25	80	189	0,917	-15	82	193
0,914	0	79	187	0,919	6	90	196
0,860	30	40	190	0,870	38	48	199
0,858	22	53	190	0,864	32	77	202

Quatre variables quantitatives sont utilisées :

Specific Gravity	[GRAmin ; Gramax]	Gravité spécifique
Freezing Point	[FREmin ; FREmax]	Point de congélation
Iodine Value	[IODmin ; IODmax]	Indice d'iode
Saponification	[SAPmin ; SAPmax]	Saponification

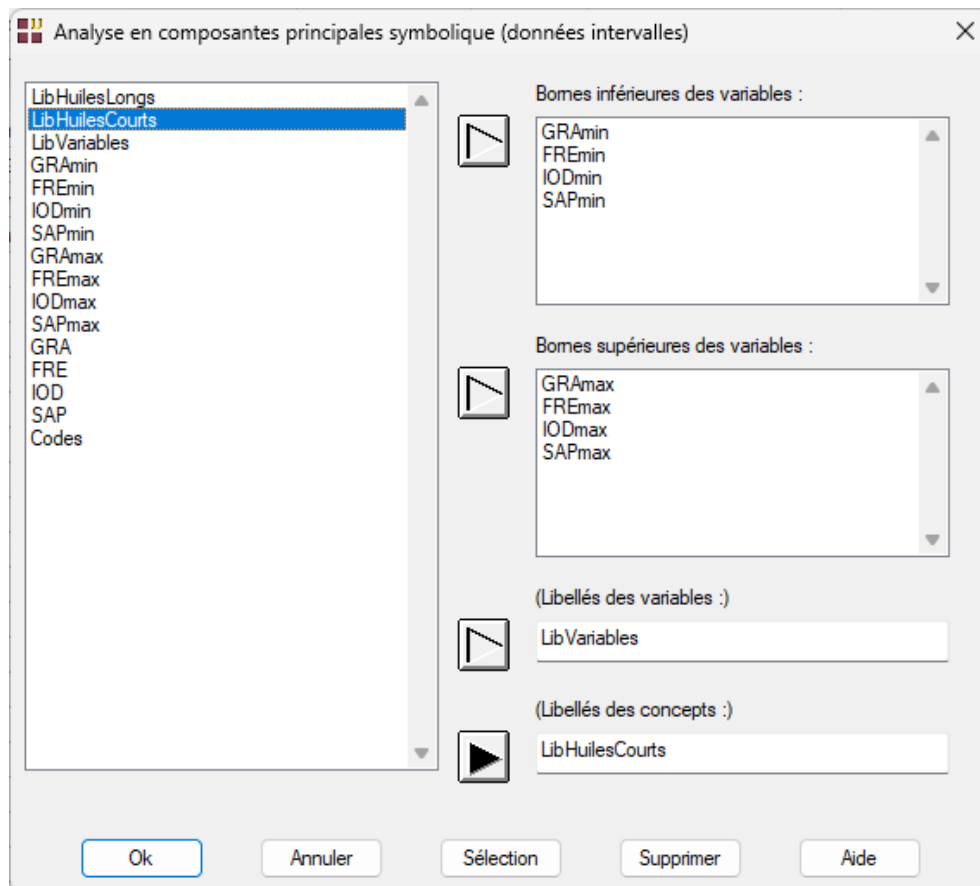
Cliquons sur l'icône ACPS dans le ruban Décrire.

Après avoir sélectionné la structure « données intervalles », la boîte de dialogue montrée ci-après s'affiche.

Sélectionnons GRamin, FREmin, IODmin et SAPmin comme variables définissant les bornes inférieures des intervalles et GRmax, FREmax, IODmax et SAPmax comme variables définissant les bornes supérieures des intervalles.

Note : il faut impérativement préciser les variables dans le même ordre pour les bornes inférieures et supérieures.

Sélectionnons LibVariables comme variable définissant les libellés des variables et LibHuilesCourts comme variable définissant les libellés des concepts.





Le bouton Sélection en pied de la boîte de dialogue permet, si cela est souhaité, de sélectionner les lignes (concepts) à étudier.

Cliquons enfin sur Ok pour exécuter le traitement de l'analyse.

Après quelques instants, la fenêtre Rapports et Graphiques s'affiche.

	1	2	3	4	5	6	7	8
1								
2	(C) UNIWIN version 9.5.1							
3								
4	DATE : 20/05/2023							
5	ORDINATEUR : LAPTOP-LEGBLO77							
6	UTILISATEUR : cchar							
7	FICHIER(S) DE DONNEES OUVERT(S) : HUILES.SGD							
8								
9	RESULTATS DE L'ANALYSE EN COMPOSANTES PRINCIPALES SYMBOLIQUE							
10								
11	Méthode mise en oeuvre : méthode des centres							
12								
13	Sélection :							
14	Aucune							
15								
16	Nombre de concepts : 8							
17								
18	Variables quantitatives :							
19	Specific Gravity							
20	Freezing Point							
21	Iodine Value							

La barre d'outils 'Rapports et Graphiques' permet par l'icône 'Données'  de rappeler la boîte de dialogue d'entrée des données.


L'icône 'Rapports'  affiche la boîte de dialogue des options pour les rapports :

Rapports

Rapport Explorateur

Rapport Général

Rapport Html

et l'icône 'Graphiques'  affiche la boîte de dialogue des options pour les graphiques.

Graphiques


Diagramme des inerties (approche classique)

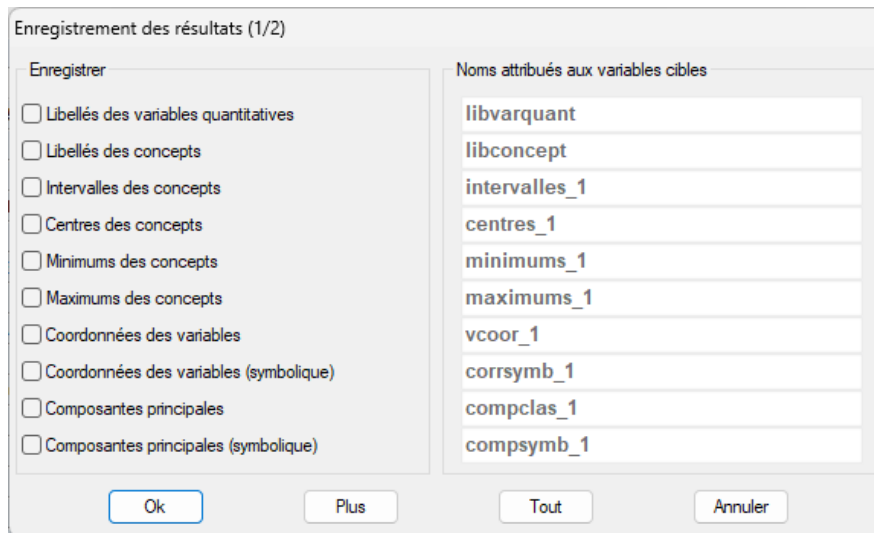
Cercle des corrélations (approche classique)

Plan factoriel des variables (approche symbolique)

Plan factoriel des concepts (approche classique)

Plan factoriel des concepts (approche symbolique)

L'icône 'Enregistrer'  permet de sélectionner les résultats de l'analyse à enregistrer dans un fichier.



Le bouton 'Plus' permet d'afficher la suite de la liste des variable et le bouton 'Tout' permet de sélectionner toutes les variables.

L'icône 'Quitter'  permet de quitter l'analyse.

L'option Rapports

Cette option permet d'obtenir le rapport à l'écran sous la forme d'un explorateur, d'un tableau ou au format HTML.

	1	2	3	4	5	6	7	8
1								
2	COMPOSANTES PRINCIPALES (SYMBOLIQUE)							
3								
4								
5		Composante 1 (min)	Composante 1 (max)	Composante 2 (min)	Composante 2 (max)	Composante 3 (min)	Composante 3 (max)	Composante 4 (min)
6	L	-4,73288	-1,27452	-1,35261	4,42813	-1,02490	1,28918	-0,9886
7	P	-1,70116	-1,05902	-1,12789	-0,34272	-1,50758	-1,04564	-0,1337
8	Co	-0,39931	0,23621	-0,96851	-0,21301	-0,17033	0,36780	-0,2459
9	S	-0,65850	-0,15369	-0,74460	-0,17862	-0,02668	0,34212	-0,3694
10	Ca	-0,61252	-0,15074	-0,88107	-0,43682	0,80670	1,20401	0,1134
11	O	-0,10045	0,59375	-0,77501	0,04294	0,01946	0,54465	-0,6448
12	B	2,22577	3,04641	0,23352	1,16192	-0,39248	0,15150	-0,5296
13	H	1,84111	2,89954	0,01950	1,13486	-0,72913	0,17131	-0,1050
14								
15								
16								
17								
18								
19								
20								
21								

Les résultats suivants sont affichés :

- Intervalles des concepts : tableau des intervalles des données utilisées
- Centres des concepts : tableau des centres des intervalles
- Tableau des inerties : tableau des valeurs propres et variances expliquées via l'ACP du nuage des centres des intervalles

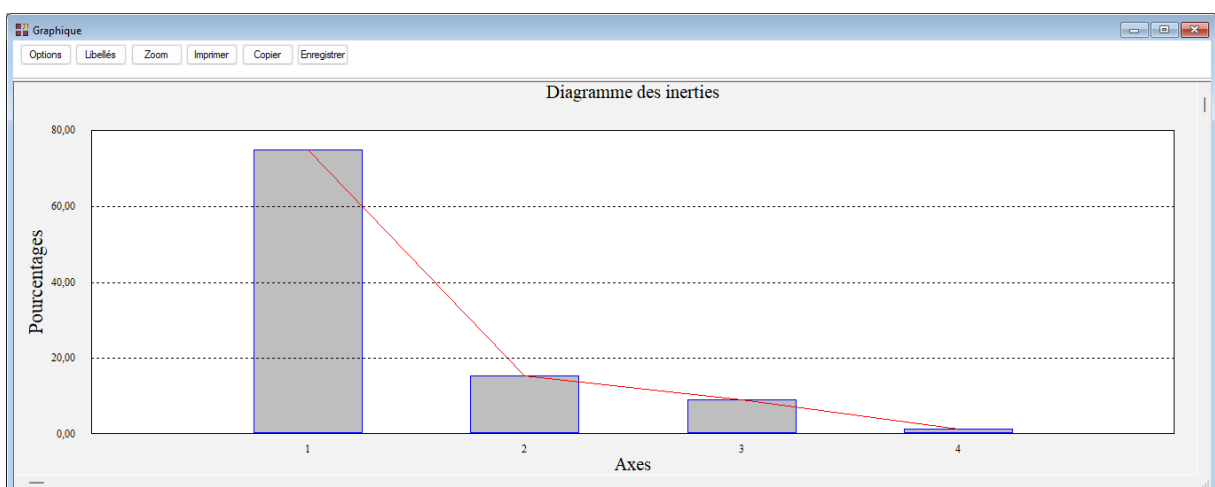
- Composantes principales (centres) : composantes principales des concepts via l'ACP du nuage des centres des intervalles
- Composantes principales (symbolique) : composantes principales des concepts via l'ACPS
- Coordonnées des variables (centres) : coordonnées des variables via l'ACP du nuage des centres des intervalles
- Coordonnées des variables (symbolique) : coordonnées des variables via l'ACPS
- Cosinus carrés des variables (centres) : cosinus carrés des variables via l'ACP du nuage des centres des intervalles
- Cosinus carrés cumulés des variables (centres) : cosinus carrés cumulés des variables via l'ACP du nuage des centres des intervalles
- Contributions des variables (centres) : contributions (%) des variables via l'ACP du nuage des centres des intervalles
- Distances carrées à l'origine (centres) : distances carrées à l'origine des concepts via l'ACP du nuage des centres des intervalles
- Cosinus carrés des concepts (centres) : cosinus carrés des concepts via l'ACP du nuage des centres des intervalles
- Cosinus carrés cumulés des concepts (centres) : cosinus carrés cumulés des concepts via l'ACP du nuage des centres des intervalles
- Contributions des concepts (centres) : contributions (%) des concepts via l'ACP du nuage des centres des intervalles
- Contributions à l'inertie totale (centres) : contributions (%) totales des concepts via l'ACP du nuage des centres des intervalles

L'option Graphiques

Cette option permet d'obtenir divers graphiques pour l'analyse ACPS.

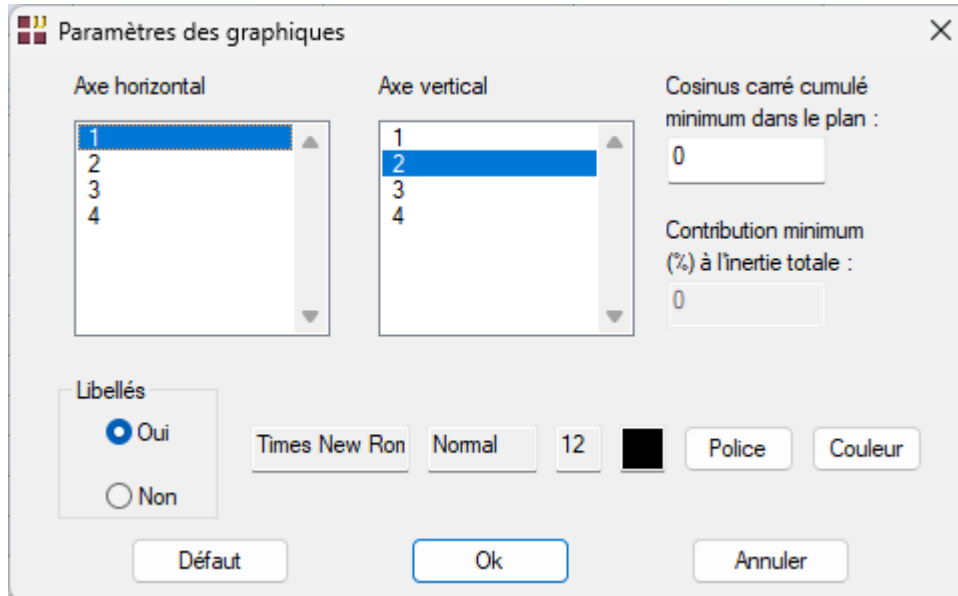
- Diagramme des inerties

Ce graphique affiche les pourcentages d'inertie pour chacun des axes factoriels.

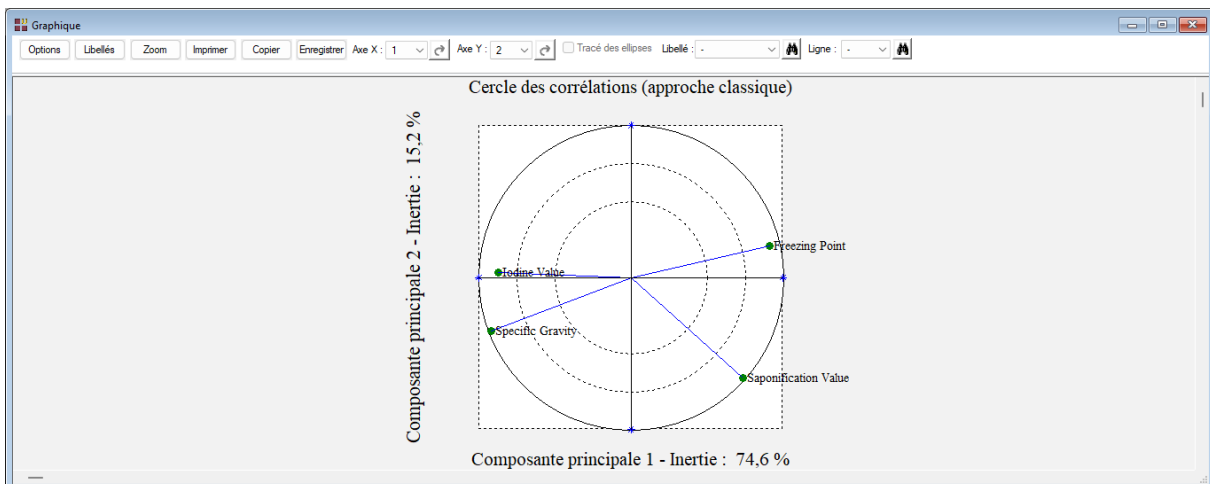


- Cercle des corrélations (approche classique)

Une boîte de dialogue s'affiche permettant de choisir le plan factoriel, de préciser si les libellés des variables sont affichés ou non et de sélectionner les variables qui seront affichées en fonction des cosinus carrés cumulés dans le plan factoriel de ces variables.



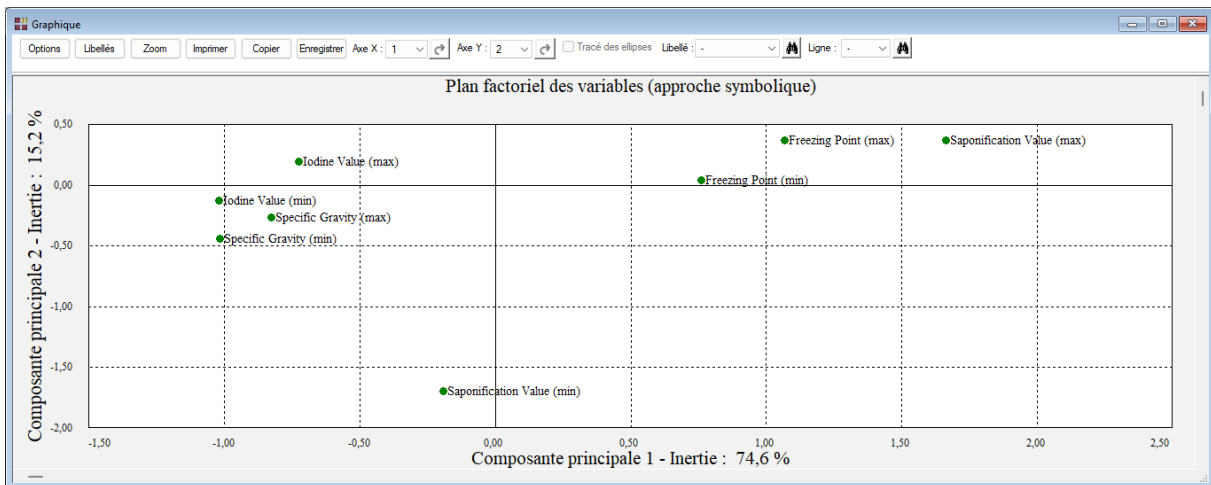
Le champ « Cosinus carré cumulé minimum dans le plan » permet de n'afficher que les variables ayant un cosinus carré cumulé dans le plan supérieur à la valeur indiquée.



Dans ce graphique, les corrélations sont calculées à partir des valeurs centrales des variables et des composantes principales, comme en ACP.

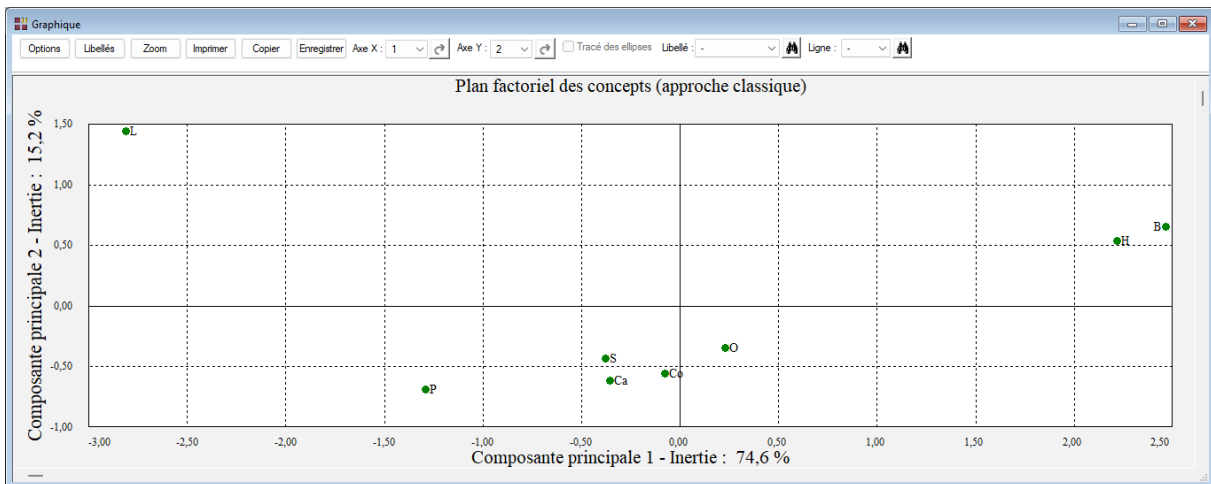
- Plan factoriel des variables (approche symbolique)

Une boîte de dialogue s'affiche permettant de choisir le plan factoriel et de préciser si les libellés des variables sont affichés ou non.



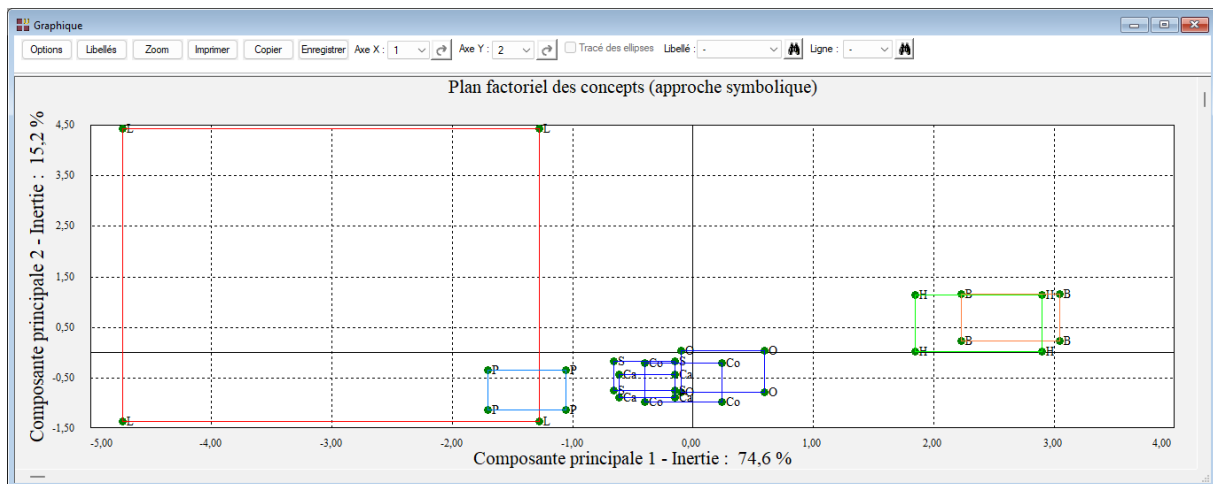
- Plan factoriel des concepts (approche classique)

Une boîte de dialogue s'affiche permettant de choisir le plan factoriel, de préciser si les libellés des concepts sont affichés ou non et de sélectionner les concepts qui seront affichées en fonction des contributions à l'inertie totale de ces concepts.



- Plan factoriel des concepts (approche symbolique)

Une boîte de dialogue s'affiche permettant de choisir le plan factoriel, de préciser si les libellés des concepts sont affichés ou non et de sélectionner les concepts qui seront affichées en fonction des contributions à l'inertie totale de ces concepts.

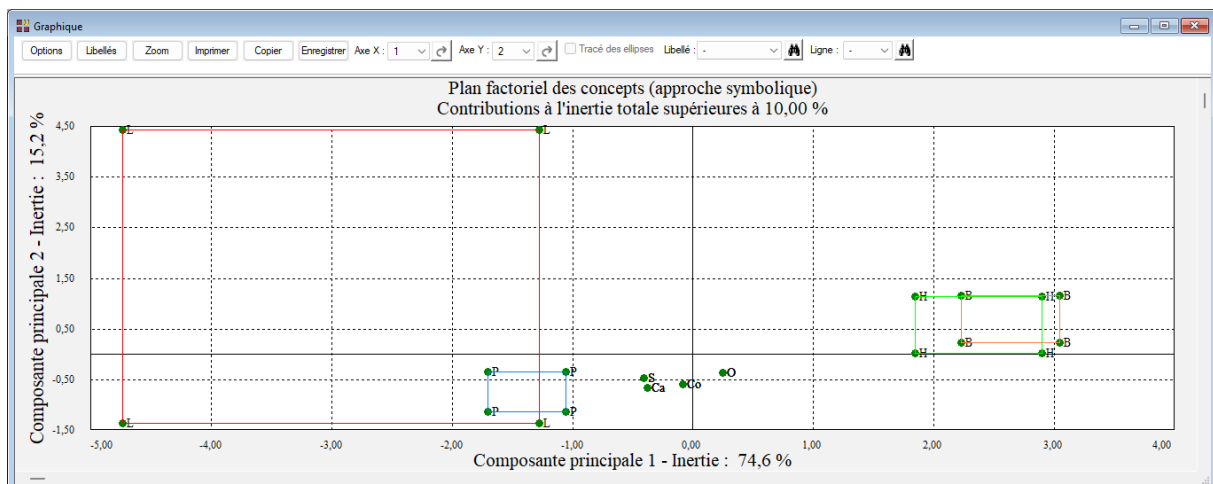


Cette représentation graphique peut rapidement devenir peu lisible, d'où l'utilité de la sélection des concepts à représenter en utilisant les contributions à l'inertie totale.

Définissons à 10 % la contribution minimum à l'inertie totale qu'un concept soit représenté par son rectangle.

Les concepts dont les contributions sont inférieures à ce seuil sont alors représentés uniquement par les centres : concepts Ca, Co, O et S.

Les rectangles sont colorés en fonction des valeurs des contributions à l'inertie totale (bleu foncé, bleu clair, vert, orange, rouge).



Une rapide interprétation des résultats

Il y a trois principaux regroupements d'huiles : {L, P}, {Ca, Co, O, S}, {B, H}.

L'analyse simultanée de plan factoriel des variables et du plan factoriel des concepts permet de fournir les variables caractérisant les regroupements d'huiles.

Ainsi, les valeurs élevées de GRA et IOD caractérisent fortement le groupe {L, P} et l'opposent au groupe {B, H} caractérisé par les valeurs élevées des variables FRE et SAP.

Si l'objectif de l'analyste est d'estimer et de connaître la tendance centrale de la dispersion des concepts, il peut quantifier chaque intervalle par son centre et utiliser les résultats de l'ACP classique appliquée aux centres des intervalles.

Si l'objectif de l'analyste consiste d'une part à étudier la dispersion globale des concepts et d'autre part à savoir comment évolue la dispersion de chaque concept quand les valeurs des variables observées varient dans leurs intervalles respectifs, il est alors nécessaire de tenir compte des valeurs de type intervalle et d'utiliser l'ACPS symbolique.

La visualisation à l'aide de rectangles permet d'une part de localiser le champ de dispersion de chaque concept quand les valeurs observées varient dans leurs intervalles respectifs et d'autre part de comparer l'amplitude de la dispersion des différents concepts.

Exemple 2 : Fichier HUILES (données individus x variables)

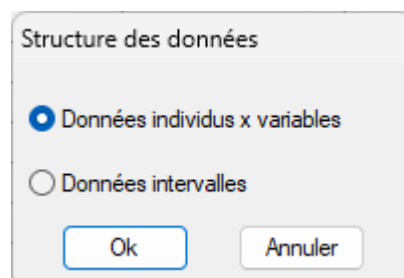
Nous utiliserons à nouveau le fichier HUILES pour illustrer ce deuxième exemple.

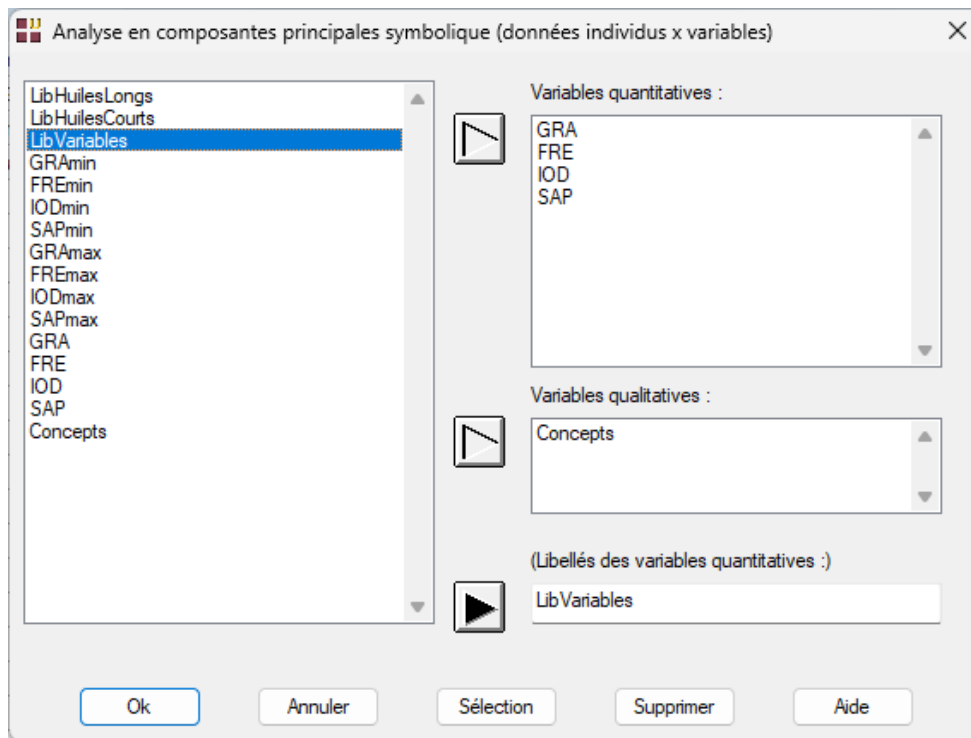
Dans cet exemple les données pour chacune des 4 variables ont été simulées pour 200 individus de façon à définir des intervalles identiques pour chacune des variables à ceux de l'exemple 1.

Une variable qualitative est utilisée pour créer les concepts à étudier.

Les quatre variables quantitatives sont GRA, FRE, IOD et SAP et la variable qualitative est 'Concepts'. Cette variable précise les concepts B, Ca, Co, H, L, O, P et S auxquels appartiennent chacun des individus.

Renseignons la boîte de dialogue comme montré ci-dessous après avoir choisi le type des données « individus x variables »





Cliquons sur le bouton Ok pour exécuter l'analyse.

Cette analyse débute par la génération des données au niveau des concepts en utilisant la variable qualitative 'Concepts' indiquée :

- Calcul des intervalles pour chacun des concepts
- Calcul des centres, minimums et maximums des concepts

Une fois ces éléments calculés (les données ont été agrégées), la procédure se déroule comme dans l'exemple 1 et les résultats obtenus sont identiques.

Exemple 3 : Fichier BATS (données intervalles)

L'ensemble de données sur les chauves-souris affiché dans le tableau ci-dessous est un exemple de données naturelles collectées sous la forme d'intervalles.

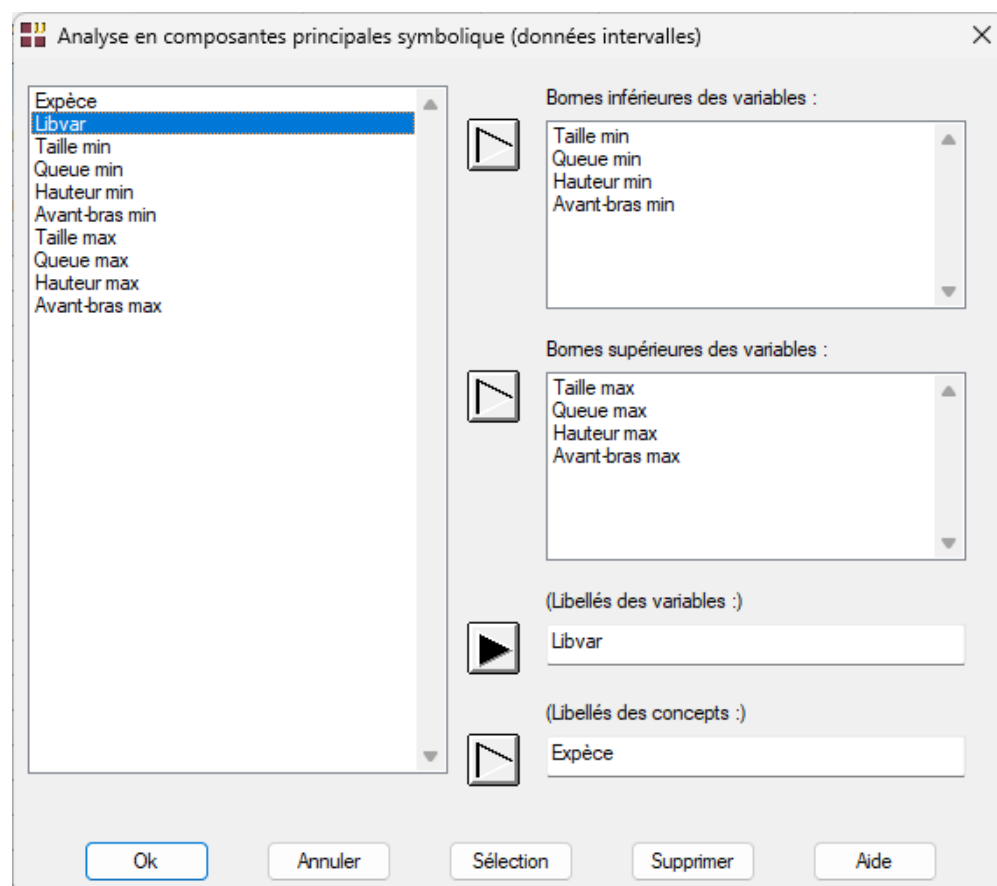
Il y a quatre variables 'taille de la tête', 'longueur de la queue', 'hauteur', 'longueur de l'avant-bras' et il y a 21 espèces (de PIPC à MGES). L'identifiant de l'espèce est une abréviation du descripteur latin biologique plus long, par exemple, 'BARB' est l'espèce 'Barbastella barbastellus'.

La question scientifique porte sur la ressemblance ou non de certaines espèces. Puisque les données sont naturellement des intervalles, une analyse en composantes principales symbolique est requise.

Expèce	Libvar	Taille min	Queue min	Hauteur min	Avant-bras min	Taille max	Queue max	Hauteur max	Avant-bras max
Type = Caractère	Type = Caractère	Type = Numérique	Type = Numérique	Type = Numérique	Type = Numérique	Type = Numérique	Type = Numérique	Type = Numérique	Type = Numérique
PIPC	Taille	33	26	4	27	52	33	7	32
PRH	Queue	35	24	8	34	43	30	11	41
MOUS	Hauteur	38	30	7	32	50	40	8	37
PIPS	Avant-bras	43	34	6	31	48	39	7	38
PIPN		44	34	7	31	48	44	8	36
MDAUB		41	30	8	33	51	39	11	41
MNAT		42	32	8	36	50	43	9	42
MDEC		40	39	9	36	45	44	9	42
MGP		45	35	10	39	53	38	12	44
OCOM		41	34	9	34	51	50	10	50
MBEC		46	34	9	39	53	44	11	44
SBOR		48	38	9	37	54	47	11	42
BARB		44	41	6	35	58	54	8	41
OGRIS		47	43	7	37	53	53	9	41
SBIC		50	40	8	40	63	45	10	47
FCHEV		50	30	11	51	69	43	13	61
MSCH		52	50	10	42	60	60	11	48
SCOM		62	46	9	48	80	57	12	56
NOCT		69	41	10	45	82	59	12	55
GMUR		65	48	12	55	80	60	16	68
MGES		82	46	11	58	87	57	12	63

Après avoir sélectionné la structure « données intervalles », la boîte de dialogue montrée ci-après s'affiche.

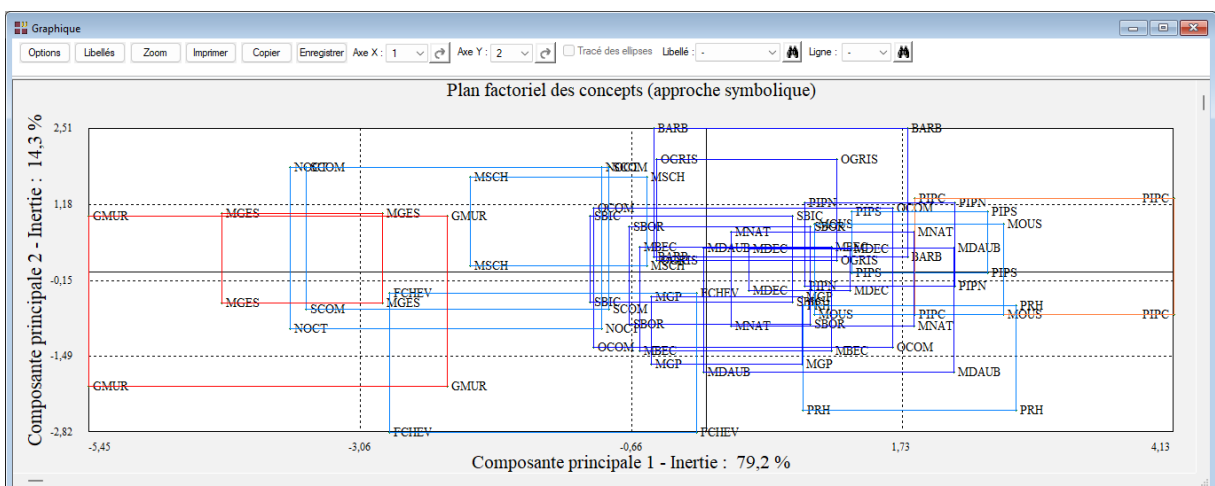
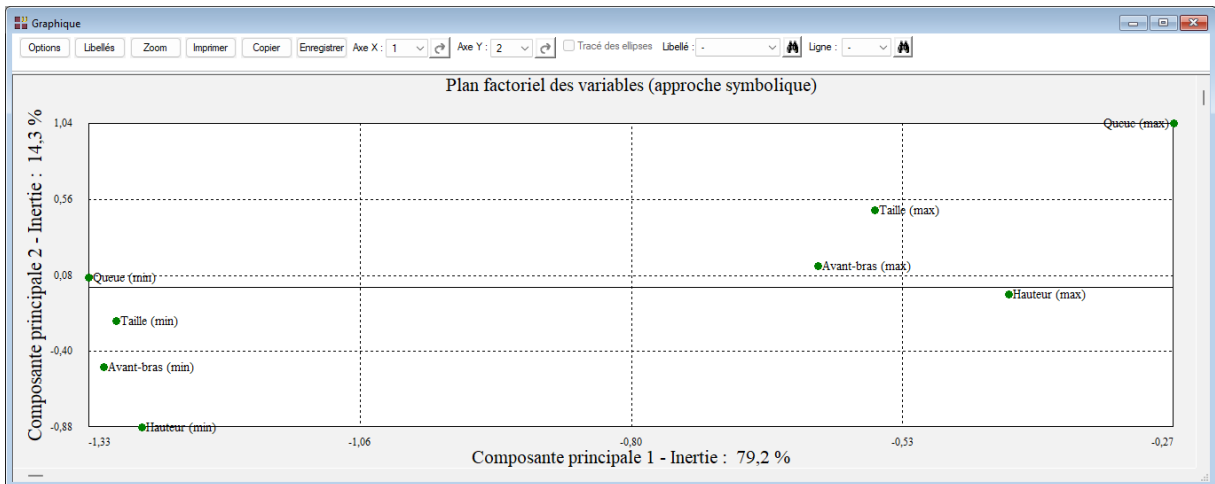
Sélectionnons 'Taille min', 'Queue min', 'Hauteur min', 'Avant-bras min' comme variables définissant les bornes inférieures des intervalles et 'Taille max', 'Queue max', 'Hauteur max', 'Avant-bras max' comme variables définissant les bornes supérieures des intervalles. Sélectionnons 'Libvar' comme variable définissant les libellés des variables et 'Espèce' comme variable définissant les libellés des concepts.



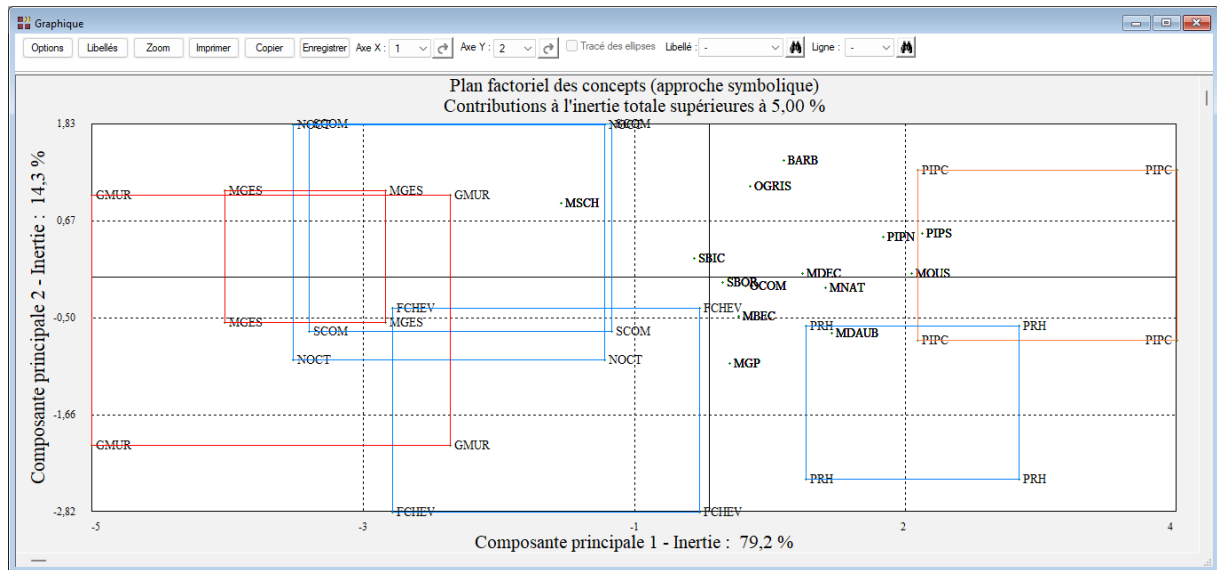
Cliquons enfin sur Ok pour exécuter le traitement de l'analyse.

Après quelques instants, les résultats de l'analyse s'affichent.

	1	2	3	4	5	6	7	8
1								
2	COMPOSANTES PRINCIPALES (SYMBOLIQUE)							
3								
4								
5		Composante 1 (min)	Composante 1 (max)	Composante 2 (min)	Composante 2 (max)	Composante 3 (min)	Composante 3 (max)	Composante 4 (min)
6	PIP	1,84226	4,13052	-0,76009	1,20660	-0,33420	2,06152	-0,9067
7	PRH	0,85340	2,73329	-2,42682	-0,59469	-0,84589	0,90427	-0,8059
8	MOUS	0,95598	2,62514	-0,75837	0,83344	-0,85325	1,05504	-0,5842
9	PIPS	1,28288	2,48341	-0,01703	1,04855	-0,07948	0,94199	-0,6824
10	PIPN	0,87264	2,19465	-0,25665	1,21084	-0,60380	0,85123	-0,2794
11	MDAUB	-0,01985	2,18589	-1,76385	0,41287	-1,13121	0,95576	-0,6202
12	MNAT	0,21673	1,83448	-0,96209	0,69117	-0,78346	0,78756	-0,6946
13	MDEC	0,37525	1,26671	-0,32849	0,38594	-0,96382	-0,21064	-0,5706
14	MGP	-0,48465	0,84710	-1,62539	-0,44794	-0,94851	0,30990	-0,3174
15	OCOM	-0,99083	1,64301	-1,32950	1,11623	-1,59986	0,64416	-1,2980
16	MBEC	-0,59003	1,10752	-1,39232	0,43427	-1,04797	0,62272	-0,5190
17	SBOR	-0,68138	0,91432	-0,92410	0,79200	-1,18053	0,36650	-0,2737
18	BARB	-0,46437	1,77601	0,26472	2,51362	-1,18368	1,10725	-1,0006
19	OGRIS	-0,43430	1,15615	0,16876	1,97116	-1,14753	0,44402	-0,6688
20	SBIC	-1,02817	0,76487	-0,53394	0,96844	-0,61995	1,10080	-0,8508
21	FCHEV	-2,79954	-0,08738	-2,82219	-0,38096	-0,97631	1,68789	-1,5001



Demandons uniquement l'affichage des concepts dont les contributions à l'inertie totale sont supérieures ou égales à 5% :



Les variables internes créées par la procédure

Voici la liste des variables internes créées par la procédure. A noter que certaines des variables mentionnées ci-dessous peuvent ne pas apparaître en fonction des options choisies.

<i>Variable</i>	<i>Contenu</i>
libvarquant	Libellés des variables quantitatives
libvaragr	Libellés des variables qualitatives
libconcept	Libellés des concepts
intervalles	Intervalles des concepts
centres	Centres des concepts
minimums	Minimums des concepts
maximums	Maximums des concepts
vcoor	Coordonnées des variables (classique)
corr symb	Coordonnées des variables (symbolique)
compclas	Composantes principales (classique)
compsymb	Composantes principales (symbolique)
vcos	Cosinus carrés des variables
vcon	Contributions des variables
ccos	Cosinus carrés des concepts
ccon	Contributions des concepts
distorig	Distances carrées des concepts à l'origine
cciner	Contributions des concepts à l'inertie totale

Références

Billard L. et Diday E. (2006) : Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester

Bock H-H. et Diday E. (eds.) (2000) : Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data. Springer, Germany

Cazes P., Chouakria A., Diday E. et Schektman Y. (1997) : Extension de l'analyse en composantes principales à des données de type intervalle, Revue de Statistique Appliquée, Vol. XLV, Num. 3, pages. 5-24, France

Chouakria A. (1998) : Extension des méthodes d'analyse factorielle à des données de type intervalle, Thèse Université Paris IX Dauphine

Ichino M (1994) : Generalized Minkowski metrics for mixed feature type data analysis. IEEE, transactions on systems, man and cybernetics, vol 24, n° 4.

Lynn Billard, Ahlame Douzal-Chouakria, E. Diday (2008) : Symbolic principal component for interval valued observations.