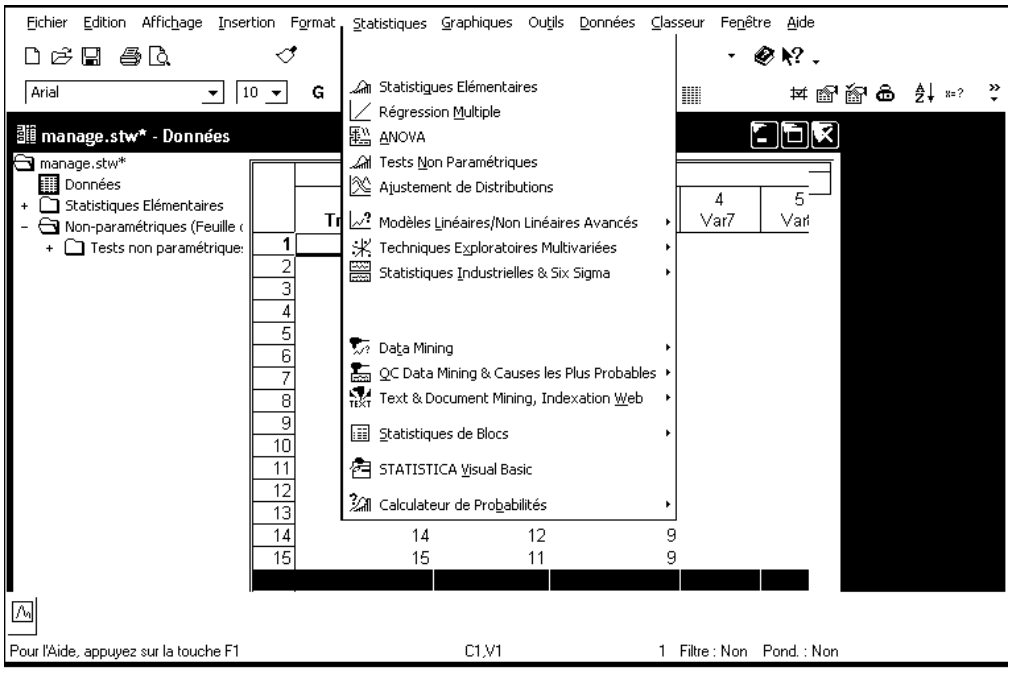
Analyse multidimensionnelle des données

1 Présentation de Statistica

1.1 . Statistica : l'interface utilisateur

1.1.1 L'écran de travail

Statistica 6.1 est un logiciel dédié aux traitements statistiques. C'est également la "brique" de base des logiciels proposés par Statsoft, et ses possibilités d'interaction avec d'autres logiciels (tableurs, systèmes de gestion de bases de données, traitements de textes, ...) sont nombreuses. En revanche, l'interface utilisateur pourra sembler un peu déconcertante au premier abord.



1.1.2 Les objets manipulés par Statistica

La **feuille de données** est organisée en variables et observations. Les colonnes sont les variables. Chaque ligne représente un individu statistique, appelé observation.

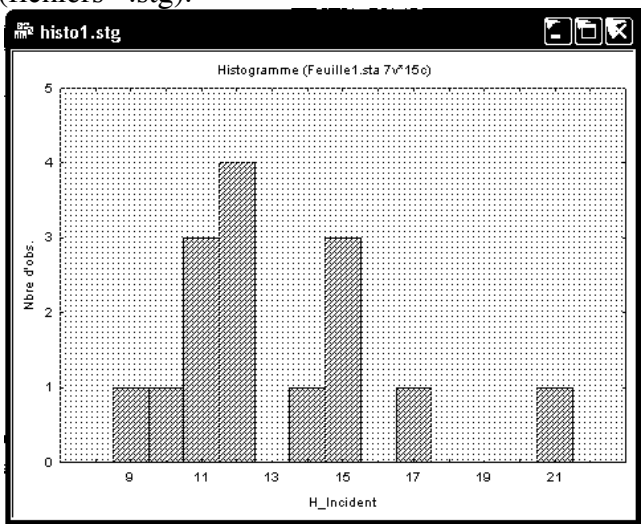
Données : Feuille1.sta (7v par 15c)

	1 Trimestre	2 H_Incident	3 D_Incident	4 Var7
1	1	11	8	
2	2	11	13	
3	3	14	12	
4	4	21	17	
5	5	12	14	
6	6	10	9	
7	7	15	10	
8	8	15	12	
9	9	17	13	
10	10	9	10	
11	11	12	8	
12	12	12	13	
13	13	15	12	
14	14	12	9	
15	15	11	9	

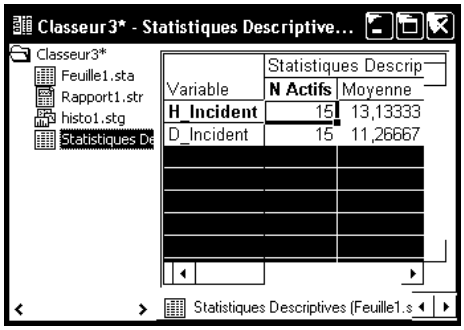
Les feuilles de données peuvent être enregistrées comme fichiers autonomes (fichiers *.sta). Elles contiennent les données d'entrée sur lesquelles s'effectuent les traitements statistiques. Les résultats de ces traitements s'affichent dans un document de sortie. Plusieurs possibilités sont offertes.

Fenêtre de rapport : C'est la méthode traditionnelle pour gérer les résultats produits par le logiciel. Un rapport se comporte plus ou moins comme un document produit par un traitement de textes. On peut insérer des commentaires, modifier la mise en forme, spécifier la mise en page, la numérotation des pages, l'en-tête et le pied de page en vue de l'impression. Les rapports peuvent être enregistrés comme fichiers autonomes (fichiers *.str).

Les résultats de sortie peuvent également être dirigés vers des fenêtres individuelles. Les résultats numériques sont alors affichés dans des fenêtres de données. Les graphiques sont affichés dans des **fenêtres de graphiques** (fichiers *.stg).

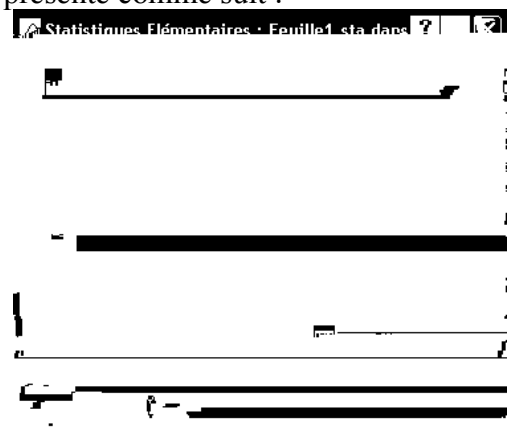


Les classeurs : les données d'entrée et de sortie peuvent également être stockées comme onglets dans un classeur. Un classeur est un "container" accueillant d'autres objets, organisés sous forme hiérarchique. Ils correspondent aux fichiers de type *.stw.



Traitements statistiques

Statistica est organisé en modules, accessibles à partir du menu Statistiques. Chaque module contient un groupe de procédures statistiques reliées entre elles. Par exemple, le module "Statistiques élémentaires" se présente comme suit :



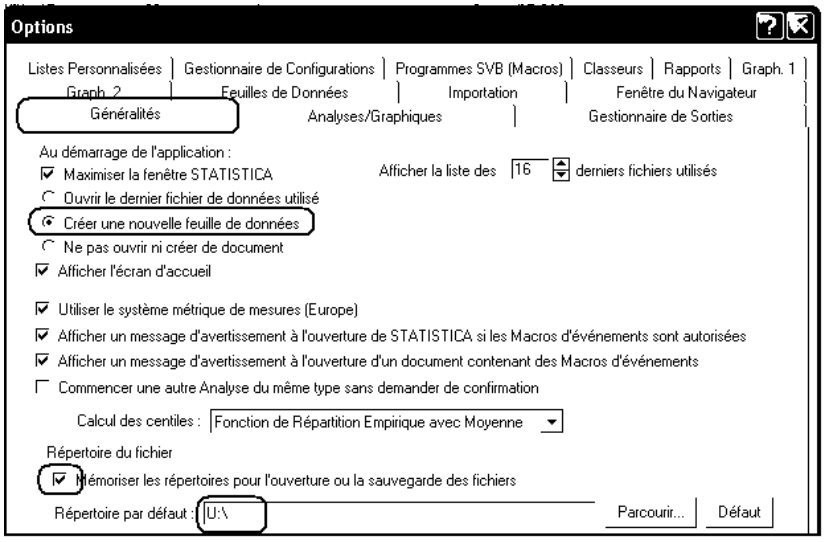
1.2 Gérer les sorties

1.2.1 Modifier le comportement de Statistica

Le comportement de Statistica peut être modifié en intervenant dans la fenêtre de dialogue affichée par le menu Outils - Options.

Par exemple, nous souhaitons :

- que Statistica n'ouvre plus systématiquement la dernière feuille de données utilisée lors du chargement du logiciel ;



1.2.2 Gérer les sorties

Lorsqu'on utilise Statistica sans se préoccuper des options de sortie des résultats, on se retrouve vite à la tête d'une quantité de fenêtres (classeurs, feuilles de données de résultats, fenêtres de graphiques...). Pour réaliser un travail que l'on souhaite conserver et reprendre au cours de plusieurs séances de travail, il paraît indispensable d'organiser correctement son espace de travail et ses sauvegardes.

En fait, plusieurs méthodes de travail sont envisageables avec Statistica :

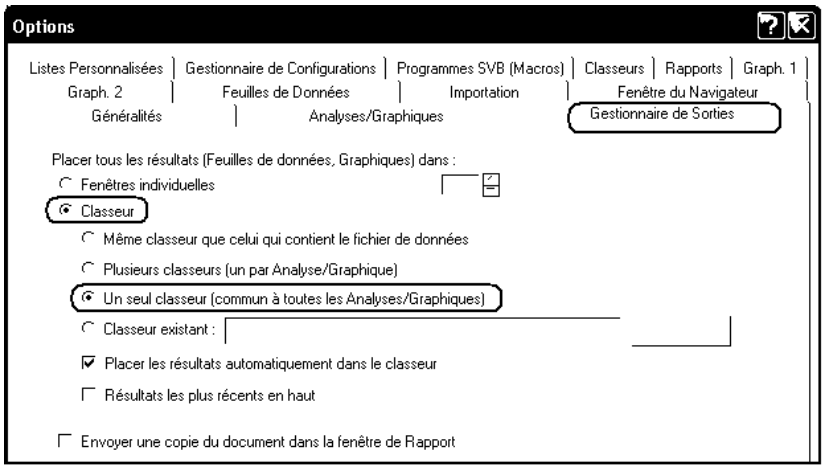
1.2.2.1 Première méthode : utiliser un fichier de données et un classeur de résultats

C'est la méthode que nous avons utilisée jusqu'à présent, pour la plupart des traitements que nous avons effectués :

- Les données se trouvaient dans une feuille de données séparée (fichier *.sta)

- Les résultats des traitements étaient produits dans un classeur (fichier *.stw) et Statistica produisait un seul classeur pour l'ensemble d'une session de travail.

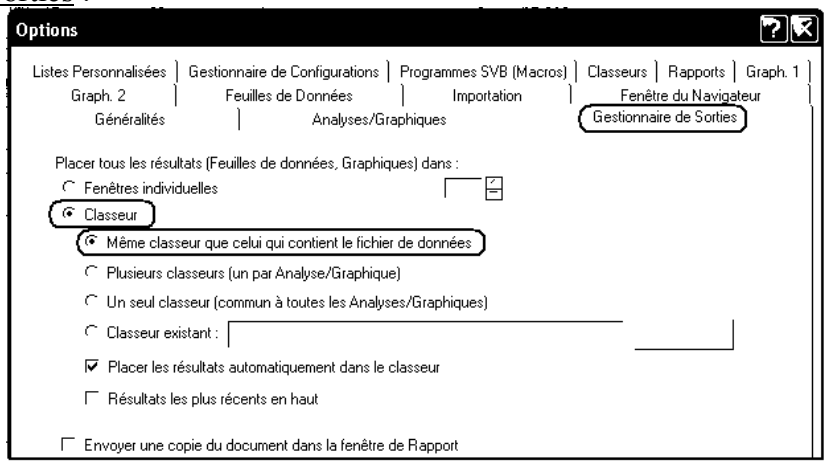
Ce comportement correspond aux options "par défaut" de Statistica. Mais ces options ne sont pas toujours adaptées au travail à réaliser. Ces options correspondent aux réglages suivants dans le menu Outils - Options - Onglet Gestionnaire de Sorties :



1.2.2.2 Deuxième méthode : enregistrer données et résultats dans un seul classeur

Cette méthode consiste à enregistrer les données, les résultats de traitements, et les commentaires éventuels comme objets d'un même classeur. Ainsi, un unique fichier du disque rassemble l'ensemble de notre travail sur un cas donné.

Ce comportement correspond aux réglages suivants dans le menu Outils - Options - Onglet Gestionnaire de Sorties :



Remarque : Le réglage ne sera actif que si la feuille de données se trouve effectivement dans un classeur. Or, ce ne sera pas le cas si la feuille de données a été ouverte à partir d'un fichier *.sta, ou importée à partir d'une feuille Excel. Dans ce cas, vous devez insérer la feuille de données dans le classeur comme il a été indiqué au paragraphe précédent.

1.2.2.3 Indiquer quelle est la feuille de données active

Lors des premières manipulations avec Statistica, nous n'avons pas eu besoin de nous préoccuper de la notion de "feuille de données active", les choix par défaut faits par Statistica nous convenant parfaitement. Cependant, cette notion permet de résoudre plusieurs problèmes :

- Ouvrir plusieurs fichiers .sta et effectuer un travail sur l'un d'eux (pas nécessairement le dernier ouvert)

- Utiliser une feuille de résultats comme feuille de données pour des traitements ultérieurs.

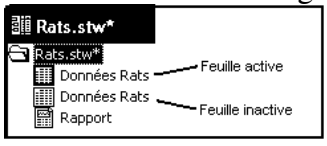
- Lorsque l'on travaille avec une feuille de données insérée dans un classeur, il arrive couramment que Statistica ne retrouve pas la feuille à partir de laquelle les traitements doivent être effectués. Mais on peut éviter ce comportement en spécifiant la propriété "feuille de données active" pour l'objet du classeur qui contient nos données.

Pour spécifier comme feuille de données active une feuille d'un classeur :

- Cliquez avec le bouton droit de la souris sur l'icône de la feuille de données dans le volet gauche du classeur.

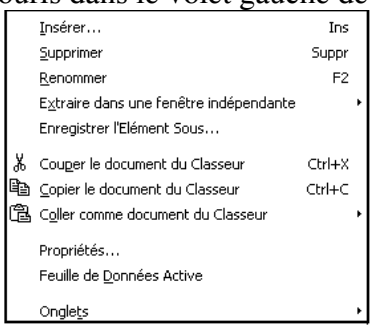
- Utilisez l'item Feuille de données active du menu local.

On peut également utiliser le menu Données - Feuille de données active.
Remarquez que le volet gauche d'un classeur indique si une feuille insérée dans le classeur est active ou non : l'icône d'un feuille active est encadrée en rouge :

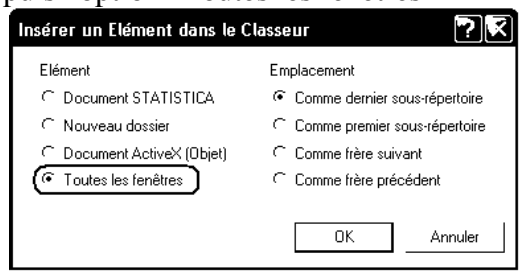


1.2.3 Enregistrer les données et l'ensemble des traitements réalisés dans un même classeur

Pour enregistrer données, traitements et rapport dans un seul classeur :
Affichez la fenêtre du classeur contenant les résultats.
Cliquez avec le bouton droit de la souris dans le volet gauche de la fenêtre du classeur.



Sélectionnez l'item Insérer..., puis l'option "Toutes les fenêtres" :



N'oubliez pas, ensuite, de spécifier la feuille Internat.sta du classeur comme feuille active.

Après avoir refermé toutes les fenêtres autres que celle du classeur, poursuivez le traitement en effectuant une comparaison de moyennes sur groupes appareillés. Rassemblez au besoin les fenêtres de résultats dans le classeur et enregistrez-le.

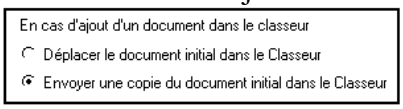
1.2.4 Manipuler les objets contenus dans un classeur

1.2.4.1 Copier - coller entre classeurs, entre un classeur et un objet Statistica

Pour déplacer un objet d'un classeur à un autre, il suffit de déplacer son icône depuis le volet gauche du premier classeur dans le volet gauche du second. On peut également utiliser les menus locaux Copier et Coller obtenus à l'aide d'un clic droit dans le volet gauche de chaque classeur.

Le menu local "Insérer" du volet gauche d'un classeur permet également d'insérer dans ce classeur un document contenu dans une fenêtre indépendante. Il suffit de choisir les options : Document Statistica - Créer à partir d'une fenêtre.

L'opération faite par Statistica est soit une copie (l'original de l'objet est conservé) soit un déplacement (l'original de l'objet n'est pas conservé) selon le paramétrage choisi dans le menu Outils - Options - Onglet Classeurs - Item "En cas d'ajout d'un document dans le classeur".



1.2.4.2 Supprimer un objet d'un classeur

Il est également possible de supprimer un objet d'un classeur, à l'aide d'un clic droit et de l'item de menu Supprimer. Cela permet notamment de ne garder, pour un traitement donné, que le résultat le plus abouti. Attention cependant : lorsque l'on supprime un objet qui n'est pas une feuille de la hiérarchie, on supprime en même temps tous les objets qui en dépendent.

1.2.5 Travail avec un rapport

Les rapports sont des documents "texte" contenant les résultats des traitements. Pour un certain nombre d'usages, ils sont préférables aux autres objets de Statistica.

- En vue d'une impression : lorsqu'il imprime un classeur, Statistica imprime chaque objet sur une page séparée. Au contraire, le contenu du rapport pourra être imprimé séquentiellement, et en indiquant des en-têtes, pieds de page, numéros de page, etc.

- Pour insérer des commentaires, ou des titres, entres les différents traitements. En effet, un rapport est fondamentalement un objet de type "texte" dans lequel l'utilisateur peut insérer du texte libre et le mettre en forme.

- En vue d'une importation des objets Statistica dans Word, à l'aide des menus Copier et Coller. En effet, lorsqu'un objet est copié à partir d'un rapport, sa taille est mieux ajustée.

- En vue d'une exploitation des résultats de traitement sous Word. En effet, un rapport peut être enregistré au format *.rtf, puis ouvert à l'aide de Word.

Remarque 1 : Dans le menu Outils - Options, l'onglet Gestionnaire de sorties permet d'obtenir une copie des résultats des traitements dans un rapport. Mais, même si l'option "Placer tous les résultats dans le même classeur que celui qui contient les données" est active, le rapport n'est pas automatiquement inséré dans le classeur des données et traitements. Il faut donc d'utiliser la méthode du paragraphe précédent pour insérer le rapport dans le classeur à un moment quelconque de la session. C'est ce rapport qui continuera à être utilisé pour les traitements ultérieurs.

2 Analyse en composantes principales ou ACP

2.1 Introduction

On a observé p variables sur n individus. On dit qu'il s'agit d'un protocole multivarié.
On cherche à remplacer ces p variables par q nouvelles variables résumant au mieux le protocole, avec $q \leq p$ et si possible q=2.
L'une des solutions à ce problème est l'ACP, méthode qui a l'avantage de résumer un ensemble de variables corrélées en un nombre réduit de facteurs non corrélés.

2.2 Analyse en composantes principales avec Statistica

Ouvrez le fichier Factor.sta.
Source : Exemple fourni avec le logiciel Statistica.

Cet exemple est basé sur un fichier de données fictives décrivant une étude de satisfaction dans la vie. Supposez qu'un questionnaire a été soumis à un échantillon aléatoire de 100 adultes. Le questionnaire comportait 10 questions créées pour mesurer la satisfaction au travail, la satisfaction dans les loisirs, la satisfaction au domicile et la satisfaction générale dans d'autres domaines. Les réponses à toutes les questions ont été enregistrées via un ordinateur et échelonnées pour que la moyenne de toutes les questions soit d'environ 100.

TRAV_1	Satisfaction professionnelle, première dimension
TRAV_2	Satisfaction professionnelle, seconde dimension
TRAV_3	Satisfaction professionnelle, troisième dimension
OCCUP_1	Satisfaction par rapport aux loisirs, première dimension

OCCUP_2	Satisfaction par rapport aux loisirs, seconde dimension
DOMI_1	Satisfaction au domicile, première dimension
DOMI_2	Satisfaction au domicile, seconde dimension
DOMI_3	Satisfaction au domicile, troisième dimension
DIVERS_1	Satisfaction générale, première dimension
DIVERS_2	Satisfaction générale, seconde dimension

Extrait des données :

Ce fichier contient des variables aléatoires basées sur deux facteurs

	1 TRAV_1	2 TRAV_2	3 TRAV_3	4)CCUP_	5)CCUP_	6 DOMI_1	7 DOMI_2	8 DOMI_3	9 IVERS_	10 IVERS_
1										
2										
3										
4										
5										

2.2.1 Statistiques descriptives - Matrice des corrélations

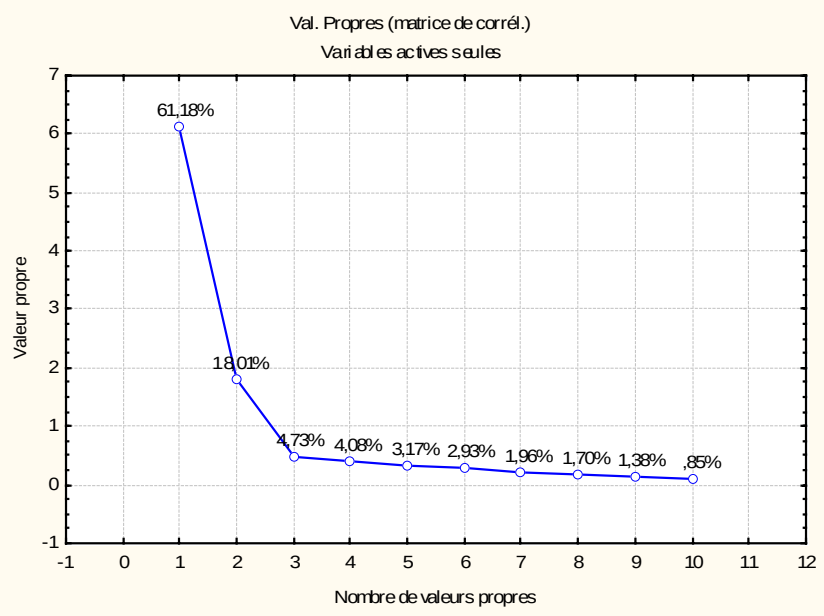
Ces résultats peuvent être obtenus à l'aide de l'onglet "Descriptives".

Variable	Corrélations (Factor.sta)									
	TRAV_1	TRAV_2	TRAV_3	)CCUP_	)CCUP_	DOMI_1	DOMI_2	DOMI_3	)IVERS_	)IVERS_
TRAV_1	1,0000	0,6474	0,6526	0,5981	0,5211	0,1428	0,1451	0,1378	0,6113	0,5489
TRAV_2	0,6474	1,0000	0,7319	0,6885	0,6978	0,1434	0,1819	0,2360	0,7086	0,6848
TRAV_3	0,6526	0,7319	1,0000	0,6369	0,6300	0,1636	0,2383	0,2546	0,6979	0,6706
OCCUP_1	0,5981	0,6885	0,6369	1,0000	0,8047	0,5364	0,6343	0,5828	0,9045	0,8432
OCCUP_2	0,5211	0,6978	0,6300	0,8047	1,0000	0,5059	0,4959	0,4824	0,8110	0,7558
DOMI_1	0,1428	0,1434	0,1636	0,5364	0,5059	1,0000	0,6577	0,5900	0,4984	0,4247
DOMI_2	0,1451	0,1819	0,2383	0,6343	0,4959	0,6577	1,0000	0,7306	0,6436	0,5934
DOMI_3	0,1378	0,2360	0,2546	0,5828	0,4824	0,5900	0,7306	1,0000	0,5859	0,5177
DIVERS_1	0,6113	0,7086	0,6979	0,9045	0,8110	0,4984	0,6436	0,5859	1,0000	0,8414
DIVERS_2	0,5489	0,6848	0,6706	0,8432	0,7558	0,4247	0,5934	0,5177	0,8414	1,0000

2.2.2 Choix des valeurs propres

Affichez d'abord le tableau des valeurs propres et le diagramme correspondant.
Pour cela, cliquez sur les boutons "Valeurs propres" et "Tracé des valeurs propres" de l'onglet "Base".

Valeur numéro	Val. Propres (matrice de corrél.) & stat. associées (Factor.sta) Variables actives seules			
	Val. propr	% Total variance	Cumul Val.	Cumul %
1	6,1184	61,1837	6,1184	61,1837
2	1,8007	18,0068	7,9191	79,1905
3	0,4729	4,7289	8,3919	83,9194
4	0,4080	4,0800	8,7999	87,9993
5	0,3172	3,1722	9,1172	91,1716
6	0,2933	2,9330	9,4105	94,1046
7	0,1958	1,9581	9,6063	96,0626
8	0,1704	1,7043	9,7767	97,7670
9	0,1380	1,3797	9,9147	99,1467
10	0,0853	0,8533	10,0000	100,0000

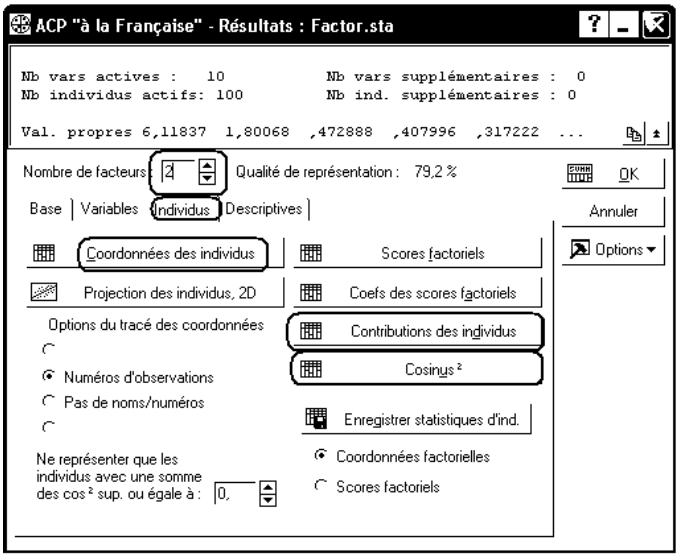


Dans notre cas, on peut choisir de retenir 2 composantes principales. Dans les manipulations qui suivent, on indiquera donc 2 dans la zone d'édition "nombre de facteurs".

Pour les résultats relatifs aux individus et aux variables, on utilisera de préférence les onglets correspondants.

2.2.3 Résultats relatifs aux individus

On pourra obtenir successivement les scores des individus, leurs contributions à la formation des composantes principales et leurs qualités de représentation en utilisant les boutons "Coordonnées des individus", "Contributions des individus", "Cosinus²".



	Coordonnées factorielles des ind., basées sur les corrélations (Factor.sta)		
Individus	Fact. 1	Fact. 2	Fact. 3
1	-0,6429	1,2723	0,3360
2	4,5343	-1,1247	0,2304
3	2,8225	-0,9212	-0,2351
4	0,7013	0,9090	0,5815
5	2,2856	1,8371	2,4521

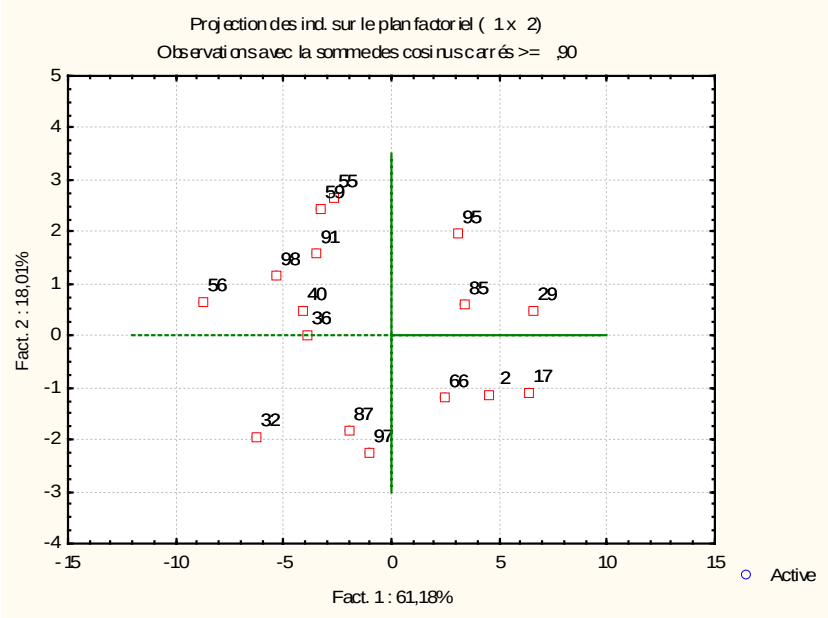
	Contributions des ind., basées sur les corrélations (Factor.sta)		
Individus	Fact. 1	Fact. 2	Fact. 3
1	0,07	0,90	0,24
2	3,36	0,70	0,11
3	1,30	0,47	0,12
4	0,08	0,46	0,72
5	0,85	1,87	12,72

	Cosinus carrés, basées sur les corrélations (Factor.sta)			
Individus	Fact. 1	Fact. 2	Fact. 3	Fact 1 & 2
1	0,0984	0,3854	0,0269	0,4839
2	0,9154	0,0563	0,0024	0,9717
3	0,6919	0,0737	0,0048	0,7656
4	0,1153	0,1937	0,0793	0,3090
5	0,3436	0,2220	0,3955	0,5656

Remarquez que les résultats ainsi obtenus sont présentés dans des feuilles de résultats sur lesquelles il est possible d'effectuer les mêmes transformations (tris, ajout ou suppression de colonne, etc) que sur les feuilles contenant les données de base. Ainsi, une colonne supplémentaire a été ajouté au tableau des cosinus-carrés pour indiquer la qualité de représentation des individus dans le premier plan factoriel.

On peut ensuite obtenir les projections du nuage des individus selon les premiers axes factoriels à l'aide du bouton "Projection de individus, 2D". Lorsque les individus ne sont pas anonymes (ce n'est pas le cas ici), il est utile d'étiqueter chaque point. Plusieurs méthodes sont possibles :

- Utiliser les identifiants d'individus figurant dans la première colonne du tableau de données (pour notre fichier de travail, ils n'ont pas été définis)
- Utiliser les numéros des observations
- Utiliser les étiquettes indiquées dans la variable "illustrative" : ces étiquettes peuvent être des identifiants des individus, mais peuvent également représenter un groupe d'appartenance, etc.



Dans certains cas, il pourra être utile de modifier les échelles sur les axes de manière à obtenir une représentation en axes orthonormés. L'importance de la part d'inertie expliquée par le premier axe principal apparaît ainsi plus clairement.

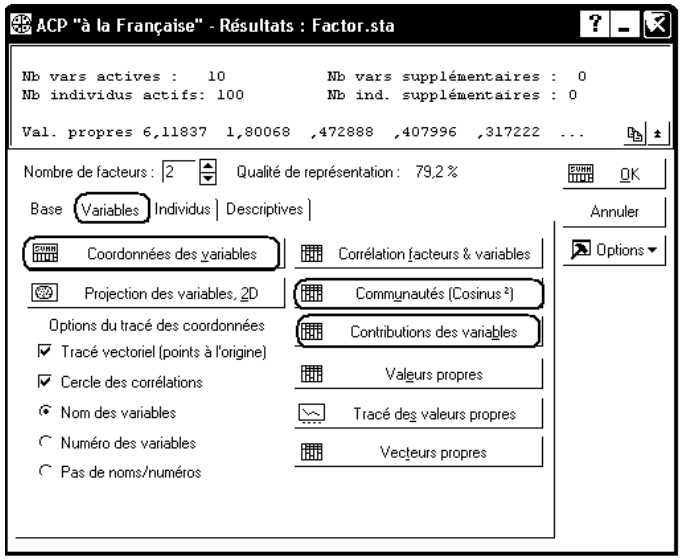
2.2.4 Résultats relatifs aux variables

Activons ensuite l'onglet "Variables".

On obtient les saturations des variables en cliquant sur le bouton "Coordonnées des variables" ou le bouton "Corrélation facteurs et variables" : dans le cas d'une ACP normée, ces deux traitements fournissent le même résultat.

On obtient leurs contributions à la formation des composantes principales en utilisant le bouton "Contributions des variables".

Les qualités de représentation sont calculées, de façon cumulative (qualité de la projection selon F1, puis selon le plan (F1,F2), puis selon l'espace (F1,F2,F3) en utilisant le bouton "Communautés (Cosinus²)".



Saturations des variables

	Coord. factorielles des var., basées sur les corrélations (Factor.sta)		
Variable	Fact. 1	Fact. 2	Fact. 3
TRAV_1	-0,6526	0,5142	0,3017
TRAV_2	-0,7570	0,4948	-0,0788
TRAV_3	-0,7457	0,4567	-0,1047
OCCUP_1	-0,9416	-0,0218	0,0127
OCCUP_2	-0,8756	0,0516	0,0997
DOMI_1	-0,5761	-0,6050	0,4910
DOMI_2	-0,6713	-0,6180	-0,1258
DOMI_3	-0,6415	-0,5739	-0,2686
DIVERS_1	-0,9515	0,0135	-0,0502
DIVERS_2	-0,9003	0,0482	-0,1518

Contributions des variables

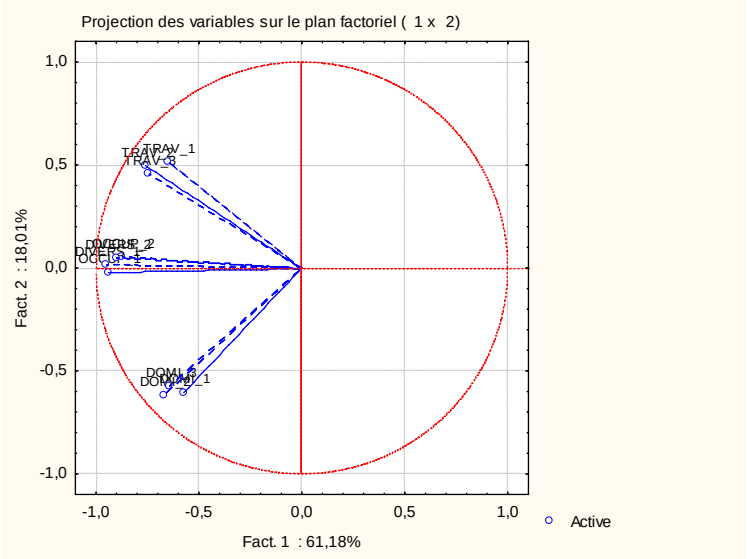
	Contributions des var., basées sur les corrélations (Factor.sta)		
Variable	Fact. 1	Fact. 2	Fact. 3
TRAV_1	0,0696	0,1468	0,1925
TRAV_2	0,0937	0,1359	0,0131
TRAV_3	0,0909	0,1158	0,0232
OCCUP_1	0,1449	0,0003	0,0003
OCCUP_2	0,1253	0,0015	0,0210
DOMI_1	0,0542	0,2033	0,5098
DOMI_2	0,0737	0,2121	0,0335
DOMI_3	0,0673	0,1829	0,1525
DIVERS_1	0,1480	0,0001	0,0053
DIVERS_2	0,1325	0,0013	0,0487

Qualités des représentations des variables

Variable	Communautés, basées sur les corrélations (Factor.sta)		
	Avec 1 facteur	Avec 2 facteurs	Avec 3 facteurs
TRAV_1	0,4259	0,6903	0,7813
TRAV_2	0,5730	0,8178	0,8240
TRAV_3	0,5561	0,7646	0,7756
OCCUP_1	0,8867	0,8871	0,8873
OCCUP_2	0,7667	0,7694	0,7793
DOMI_1	0,3318	0,6978	0,9389
DOMI_2	0,4506	0,8325	0,8483
DOMI_3	0,4116	0,7410	0,8131
DIVERS_1	0,9054	0,9056	0,9081
DIVERS_2	0,8106	0,8129	0,8360

Représentation des variables

Le bouton "Projection des variables, 2D" permet d'obtenir les diagrammes représentant les projections des variables selon les plans définis par deux axes principaux.



2.2.5 Coefficients des variables

Les coefficients des variables (c'est-à-dire la matrice permettant de passer des variables centrées réduites aux composantes principales et vice-versa) sont obtenus à l'aide du bouton "Vecteurs propres" de l'onglet "Variables".

Variable	Vecteurs propres de la matrice de corrélation (Factor.sta) Variables actives seules									
	Fact. 1	Fact. 2	Fact. 3	Fact. 4	Fact. 5	Fact. 6	Fact. 7	Fact. 8	Fact. 9	Fact.10
TRAV_1	-0,264	0,383	0,439	0,687	-0,024	0,235	0,116	-0,054	0,215	0,013
TRAV_2	-0,306	0,369	-0,115	-0,332	-0,161	0,318	-0,533	0,395	0,279	0,042
TRAV_3	-0,301	0,340	-0,152	0,048	-0,364	-0,779	0,076	0,047	-0,048	0,133
OCCUP_1	-0,381	-0,016	0,018	0,003	0,214	0,173	-0,054	0,004	-0,655	0,589
OCCUP_2	-0,354	0,038	0,145	-0,508	-0,028	0,168	0,707	-0,062	0,239	0,062
DOMI_1	-0,233	-0,451	0,714	-0,180	-0,200	-0,212	-0,328	-0,058	0,011	-0,067
DOMI_2	-0,271	-0,461	-0,183	0,250	0,400	-0,192	0,064	0,487	0,391	0,165
DOMI_3	-0,259	-0,428	-0,391	0,239	-0,644	0,295	0,026	-0,194	0,019	0,003
DIVERS_1	-0,385	0,010	-0,073	0,042	0,136	0,023	0,081	0,232	-0,422	-0,766
DIVERS_2	-0,364	0,036	-0,221	-0,055	0,402	-0,094	-0,272	-0,709	0,236	-0,104

2.2.6 Quelques remarques sur l'interprétation

Les variables sont toutes corrélées positivement entre elles. Le premier facteur est ici un facteur de "taille". Par contre, deux groupes de variables apparaissent relativement peu corrélés : TRAV_x d'une part et DOM_x d'autre part.

En fait : Le "Secret" de l'exemple parfait. L'exemple que vous avez étudié fournit en fait une solution à deux facteurs parfaite. Elle représente la plus grande partie de la variance, permet une interprétation directe, et reproduit la matrice de corrélations avec de faibles perturbations (corrélations résiduelles restantes). Bien sûr, la nature permet rarement une telle simplicité, et en réalité, ce fichier de données fictives a été généré via un générateur de nombres aléatoires. Plus précisément, deux facteurs orthogonaux (indépendants) ont été "placés" dans les données, à partir desquelles les corrélations entre les variables ont été générées. L'exemple sur l'analyse factorielle a récupéré ces deux facteurs prévus (c'est-à-dire, le facteur sur la satisfaction au travail et celui sur la satisfaction à domicile) ; en conséquence, si la nature avait placé les deux facteurs, vous auriez appris quelque chose sur la structure sous-jacente ou latente de la nature.

2.3 Interpréter les résultats d'une ACP

2.3.1 Examen des valeurs propres. Choix du nombre d'axes

On examine les résultats relatifs aux valeurs propres.
Plusieurs critères peuvent nous guider :

- "méthode du coude" on examine la courbe de décroissance des valeurs propres pour déterminer les points où la pente diminue de façon brutale ; seuls les axes qui précèdent ce changement de pente seront retenus.
- si l'analyse porte sur p variables et n > p individus, la variation totale est répartie sur p axes. On peut alors choisir de conserver les axes dont la contribution relative est supérieure à $\frac{100\%}{p}$.

2.3.2 Interpréter les résultats relatifs aux individus

Très souvent, les individus pris en compte pour une ACP sont en nombre très élevé et sont considérés comme anonymes. Les éléments qui suivent concernent évidemment les cas où ils ne le sont pas.

2.3.2.1 Contributions des individus à la formation d'un axe

On relève, pour chaque axe, quels sont les individus qui ont la plus forte contribution à la formation de l'axe. Par exemple, on retient (pour l'analyse) les individus dont la contribution relative est supérieure à $\frac{100\%}{n}$. On note également si cette contribution intervient dans la partie positive ou dans la partie négative de l'axe.

On peut ainsi caractériser l'axe en termes d'opposition entre individus. Il peut également être intéressant d'étudier comment l'axe classe les individus.

Si un individu a une contribution très forte à la formation d'un axe, on peut choisir de recommencer l'analyse en retirant cet individu, puis de l'introduire en tant qu'individu supplémentaire.

2.3.2.2 Projections des individus dans un plan factoriel

Même s'il s'agit du plan (F1, F2), les proximités entre individus doivent être interprétées avec prudence : deux points proches l'un de l'autre sur le graphique peuvent correspondrent à des

individus éloignés l'un de l'autre. Pour interpréter ces proximités, il est nécessaire de tenir compte des qualités de représentation des individus.
Se méfier également des individus proches de l'origine : mal représentés, ou proches de la moyenne, ils ont, de toutes façons, peu contribué à la formation des axes étudiés.

2.3.3 Interpréter les résultats relatifs aux variables

2.3.3.1 Contributions des variables

L'examen du tableau des contributions des variables peut permettre d'identifier des variables qui ont un rôle dominant dans la formation d'un axe factoriel.

2.3.3.2 Analyse des projections des variables sur les plans factoriels

Les diagrammes représentant les projections des variables sur les axes factoriels nous fournissent plusieurs types d'informations :

- La longueur du vecteur représentant la variable est liée à la qualité de la représentation de la variable par sa projection dans ce plan factoriel

- Pour les variables bien représentées, l'angle entre deux variables est lié au coefficient de corrélation entre ces variables (si la représentation est exacte, le coefficient de corrélation est le cosinus de cet angle). Ceci permet de dégager des "groupes de variables" de significations voisines, des groupes de variables qui "s'opposent", des groupes de variables relativement indépendantes entre eux.

- De même, pour les variables bien représentées, l'angle que fait la projection de la variable avec un axe factoriel est lié au coefficient de corrélation de cette variable et de l'axe factoriel.

2.4 ACP avec Individus et variables supplémentaires

Lorsque des individus ou des variables ont une influence trop importante sur les résultats d'une ACP, on peut essayer de recommencer les calculs en les déclarant comme individus ou variables supplémentaires.

Les données correspondantes n'interviennent plus dans le calcul de détermination des composantes principales. En revanche, on leur applique les mêmes transformations qu'aux autres données afin de les ré-introduire dans les tableaux et graphiques de résultats.

Avec Statistica, il est simple de déclarer une variable comme variable supplémentaire : le premier dialogue de l'ACP prévoit une zone d'édition pour cela. Pour déclarer des individus comme "inactifs", il est nécessaire de construire une variable supplémentaire, qui ne contiendra que deux modalités, et d'utiliser les zones d'édition "Variable avec individus actifs" et "Code des individus actifs".

2.5 ACP pondérée, ACP non normée

Dans certains cas, il peut être pertinent de pondérer les individus. Par exemple, il peut s'agir de regrouper les observations identiques. Ou encore, dans une ACP relative à des données socio-économiques sur des entités géographiques telles que des régions ou des départements, il peut être pertinent de pondérer chaque observation par une donnée démographique (nombre d'habitants).

Il est également possible de réaliser l'ACP sur les covariances des variables de départ, au lieu d'utiliser les corrélations. Le poids d'une variable dépend alors de son écart type, alors que dans l'ACP normée, toutes les variables ont le même poids.

2.5.1 Exemple d'ACP non normée

Ouvrez le fichier Protein.sta.
Source : Exemple fourni avec le logiciel Statistica.
Cet exemple particulier est présenté par Greenacre (1984) dans le cadre d'une comparaison entre l'analyse en composantes principales (voir l'Analyse Factorielle) et l'analyse des correspondances.

Les données du fichier d'exemple Protein.sta représentent des estimations de la consommation protéique issue de 9 sources différentes, par habitant dans 25 pays (les données ont initialement été reportées par Weber, 1973, dans un polycopié publié à l'Université de Kiel, Institut für Agrarpolitik und Marktlehre, intitulé "Agrarpolitik im Spannungsfeld der Internationalen Ernährungspolitik").

Extrait des données :

	Evaluation des consommations de protéïnes, en grammes/habitant/jour								
	1 VIANDE	2 ORC_VO	3 OEUFS	4 LAIT	5 'OISSON	6 EREALE	7 ECULEN	8 NOIX	9 RUI_LE
Belgique/Lux.	13,5	9,3	4,1	17,5	4,5	26,6	5,7	2,1	4,0
Bulgarie	7,8	6,0	1,6	8,3	1,2	56,7	1,1	3,7	4,2
Tchécoslovaquie	9,7	11,4	2,8	12,5	2,0	34,3	5,0	1,1	4,0
Danemark	10,6	10,8	3,7	25,0	9,9	21,9	4,8	0,7	2,4
R.D.A.	8,4	11,6	3,7	11,1	5,4	24,6	6,5	0,8	3,6
Finlande	9,5	4,9	2,7	33,7	5,8	26,3	5,1	1,0	1,4

Toutes les variables s'expriment ici avec la même unité (g.hab/jour). Pour réaliser une ACP, deux possibilités s'offrent à nous :

- Faire une ACP sur les valeurs non réduites. Ainsi, une information telle que "l'apport protéique des viandes, porc et volailles est, dans tous les cas, supérieur à celui des fruits et légumes" est prise en compte dans l'étude.

- Faire une ACP sur les valeurs réduites (ACP calculée à partir du tableau des corrélations). Dans ce cas, l'étude "gomme" les inégalités des apports protéiques des différentes sources.

Réalisons une ACP sur les covariances. Interprétons les résultats.

Affichez les tableaux des covariances et des corrélations. On voit déjà apparaître une opposition entre protéines d'origine animale et protéines d'origine végétale.

Combien de valeurs propres faut-il ici retenir ? Leur décroissance semble indiquer que l'essentiel de l'information est contenue dans les deux premières valeurs propres.

Interprétation du nuage des individus

Affichez en particulier les contributions des individus à la formation du premier axe, classées par valeurs décroissantes :

	Contributions des ind., basées sur les covariances (Protein.sta) Var. illustrative : Code Pays				
Individus	Fact. 1	Fact. 2	Fact. 3	Fact. 4	Code Pays
Bulgarie	18,31	1,50	0,18	0,25	BU
Yougoslavie	17,79	1,71	0,00	2,60	YU
Roumanie	9,76	0,91	0,13	0,72	RO
Suède	5,91	0,07	1,77	2,42	SU
Albanie	5,34	0,24	1,48	9,19	AL
Danemark	5,16	0,26	0,56	4,87	DA
R.F.A.	4,90	1,53	5,77	0,25	RFA
Finlande	4,04	17,30	10,17	7,66	FI

On constate que les pays qui ont le plus contribué à la formation du premier axe factoriel sont les la Bulgarie, l Yougoslavie et la Roumanie, qui correspondent à des valeurs négatives de la première composante principale. La suite de la liste indique ensuite des pays d'Europe de l'Ouest et du Nord (Suède, RFA, Danemark, Finlande) qui correspondent à des scores positifs. L'examen du tableau des cosinus carrés montre en outre que ces pays sont bien représentés par la première composante principale :

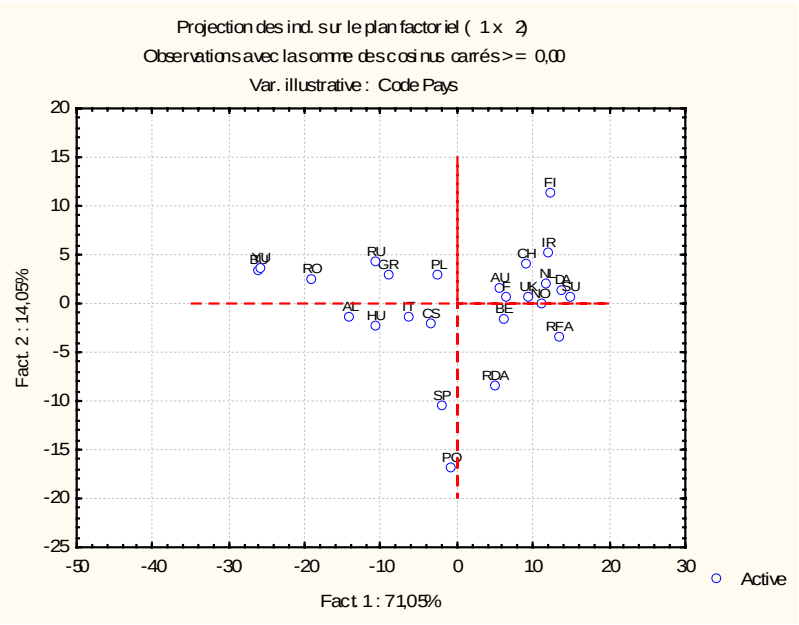
	Cosinus carrés, basés sur les covariances (Protein.sta) Var. illustrative : Code Pays				
Individus	Fact. 1	Fact. 2	Fact. 3	Fact. 4	Code Pays
Yougoslavie	0,97	0,02	0,00	0,01	YU
Roumanie	0,97	0,02	0,00	0,00	RO
Bulgarie	0,97	0,02	0,00	0,00	BU
Suède	0,92	0,00	0,03	0,02	SU
Danemark	0,88	0,01	0,01	0,04	DA
R.F.A.	0,84	0,05	0,10	0,00	RFA

Le même travail sur le deuxième facteur conduit au résultat suivant :

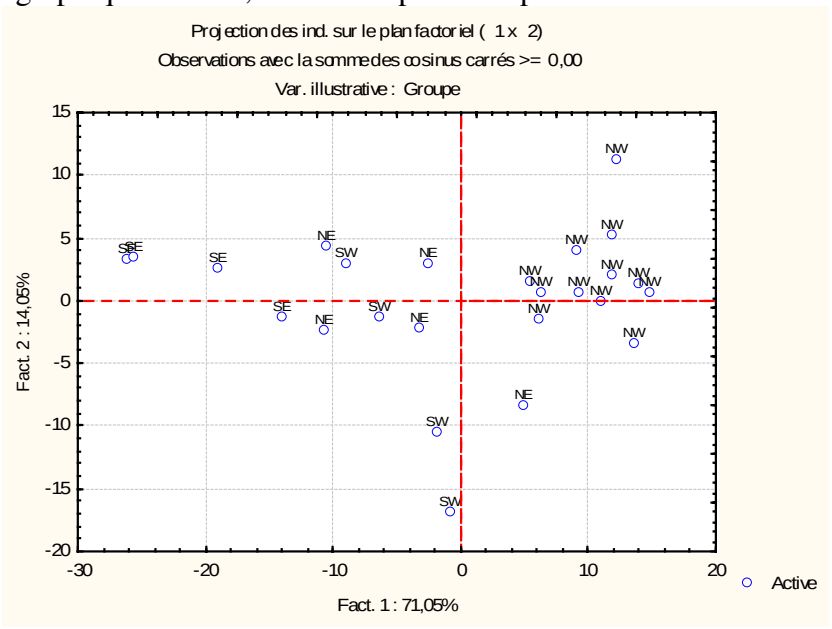
	Contributions des ind., basées sur les covariances (Protein.sta) Var. illustrative : Code Pays				
Individus	Fact. 1	Fact. 2	Fact. 3	Fact. 4	Code Pays
Portugal	0,02	38,09	9,62	3,07	PO
Finlande	4,04	17,30	10,17	7,66	FI
Espagne	0,10	14,92	3,10	0,01	SP
R.D.A.	0,65	9,48	5,57	0,85	RDA

Cet axe montre clairement une opposition entre la Finlande d'une part, et des pays tels que le Portugal et l'Espagne d'autre part.

La représentation des individus dans le premier plan factoriel est la suivante :



On voit ainsi se dessiner une double opposition : pays à économie de marché / pays à économie dirigée et pays du nord / pays du sud. Il pourrait donc être intéressant de créer une variable contenant les étiquettes "nord-ouest" (NW), "sud-ouest" (SW), "nord-est" (NE) et "sud-est" (SE).
On obtient ainsi le graphique suivant, dont l'interprétation peut être intéressante :



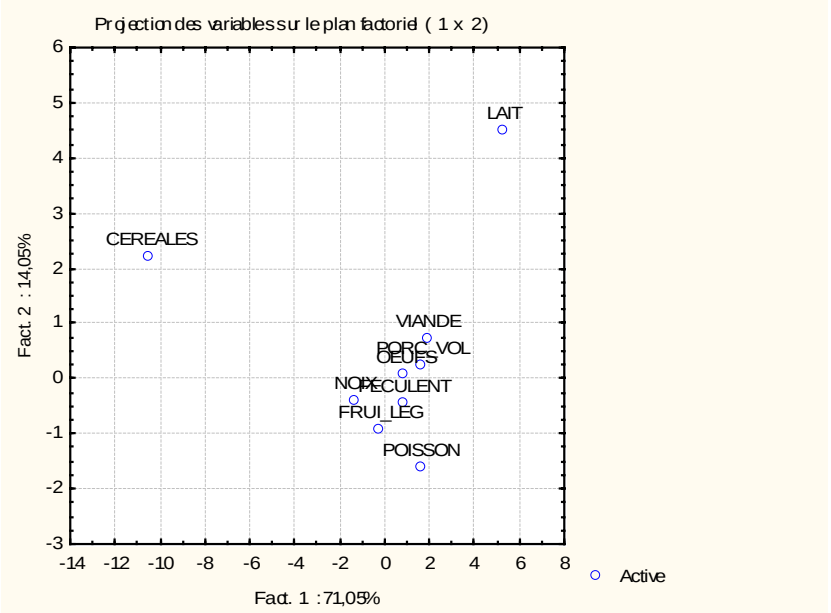
Interprétation des résultats relatifs aux variables

Variable	Cor. facteur-var. (poids fact.), basées sur les covariances (Protein.sta			
	Fact. 1	Fact. 2	Fact. 3	Fact. 4
VIANDE	0,56	0,22	0,04	0,77
PORC_VOL	0,44	0,07	-0,85	-0,14
OEUFS	0,75	0,10	-0,35	0,20
LAIT	0,75	0,65	0,12	-0,08
POISSON	0,46	-0,48	0,61	-0,24
CEREALES	-0,98	0,21	0,01	-0,01
FECULENT	0,51	-0,26	-0,08	-0,20
NOIX	-0,71	-0,20	0,33	0,18
FRUI_LEG	-0,14	-0,52	0,05	0,12

L'examen de saturations, c'est-à-dire des corrélations entre les variables et les composantes principales montre que la première composante principale est très fortement corrélée (négativement) à "céréales". Plus généralement, elle est corrélée négativement avec la plupart des sources de protéines végétales et positivement avec les sources de protéines animales. L'interprétation de la seconde composante principale est moins évidente. On peut cependant s'appuyer sur le tableau des contributions des variables pour faire apparaître l'importance prise par la variable 'lait" dans cette deuxième composante :

	Contributions des var., basées sur les covariances (Protein.sta)			
Variable	Fact. 1	Fact. 2	Fact. 3	Fact. 4
VIANDE	0,02	0,02	0,00	0,80
PORC_VOL	0,02	0,00	0,64	0,03
OEUFS	0,00	0,00	0,01	0,01
LAIT	0,18	0,69	0,05	0,04
POISSON	0,02	0,09	0,27	0,08
CEREALES	0,74	0,16	0,00	0,00
FECULENT	0,00	0,01	0,00	0,01
NOIX	0,01	0,00	0,03	0,02
FRUI_LEG	0,00	0,03	0,00	0,01

Le graphique relatif aux variables montre les rôles particuliers joués par les variables "Céréales" et "Lait", pendant que les autres variables sont assez bien regroupées.



Remarque : il peut également être intéressant d'étudier quels sont les pays les plus mal représentés par les deux premiers axes (Tchécoslovaquie, Pologne, France...) et quels sont les axes qui ont été fortement influencés par ces pays (le facteur 4, et la variable "viande" pour la France par exemple.

2.5.2 Exemple d' étude avec des individus supplémentaires

Dans l'étude précédente, il pourrait être intéressante de placer comme individus supplémentaires les moyennes de consommation de protéines pour chacun des 4 groupes de pays qui ont été définis, ce qui permettrait de faire figurer ces éléments sur les graphiques relatifs aux individus.

On peut aussi choisir de placer en individus supplémentaires certains individus atypiques qui ont une contribution trop importante à la formation d'un axe donné. Par exemple, reprenez l'étude en plaçant la Yougoslavie, la Bulgarie et la Roumanie en individus supplémentaires (inactifs).

De même, reprenez l'étude en plaçant en outre les variables "Lait" et "Céréales" en variables supplémentaires.

Etudiez ensuite les mêmes données à l'aide d'une ACP normée. De même, il peut être intéressant de rendre inactifs certains individus (pays) ou de placer certaines variables en variables supplémentaires.
(Par exemple, l'Albanie, la Roumanie, la Bulgarie et la Yougoslavie, ainsi que la variable "Poisson").

2.6 ACP avec rotation

Par construction, les composantes principales sont des abstractions mathématiques et ne possèdent pas nécessairement de signification intuitive. Après avoir réalisé l'ACP, il peut parfois être intéressant de définir d'autres variables en effectuant une combinaison linéaire des composantes principales retenues, à l'aide d'une "rotation". L'objectif est généralement d'augmenter les saturations, c'est-à-dire les corrélations entre ces nouveaux "facteurs" et certaines variables de départ. Les nouveaux "facteurs" ainsi obtenus perdent les propriétés des facteurs principaux. Par exemple, le premier d'entre eux ne correspond plus à la direction de plus grande dispersion du nuage des individus. En revanche, la part de variance expliquée par les facteurs retenus reste identique. Il existe différents critères (varimax, quartimax, equamax, etc) permettant d'obtenir une rotation conduisant à des saturations proches de 1 ou -1, ou au contraire proches de 0.

Cette possibilité n'est pas disponible dans la méthode "ACP à la française" de Statistica. En revanche, on peut l'utiliser en utilisant le module "Analyse factorielle" convenablement paramétré.

2.7 Une ACP fournit-elle toujours des informations interprétables ?

Tout tableau de données peut être soumis à une ACP, et les méthodes d'analyse qui ont été développées permettent de "trouver des résultats". Mais ces résultats correspondent-ils à une réalité plus ou moins cachée ou ne constituent-ils qu'un artefact de la méthode ?
Pour étudier cet aspect, réalisons une ACP sur des données ... où il n'y a rien à dire (il s'agit de données produites à l'aide d'un générateur de nombres aléatoires).

Ouvrez le fichier aleatoire-20sujets.stw et réalisez une ACP normée sur ces données. La représentation graphique des valeurs propres nous indique déjà l'absence d'intérêt des données traitées :