UNIWIN VERSION 10.7.0

CLASSIFICATION ASCENDANTE HIERARCHIQUE

Révision : 26/08/2026

Définition 1
Entrée des données 2
Données manquantes 3
Exemple 1 : Fichier VEHICULE 3
L'option Rapports 7
L'option Graphiques 8
Exemple 2 : Fichier SUICIDE 13
Exemple 3 : Fichier PHYTOS 17
Exemple 4 : Fichier CHIENS 19
Les variables internes créées par la procédure 21

Définition

La méthode de Classification Ascendante Hiérarchique (CAH) permet de construire une typologie (ou partition) d'un ensemble d'individus en classes telles que les individus appartenant à une même classe sont proches alors que les individus appartenant à des classes différentes sont éloignés.

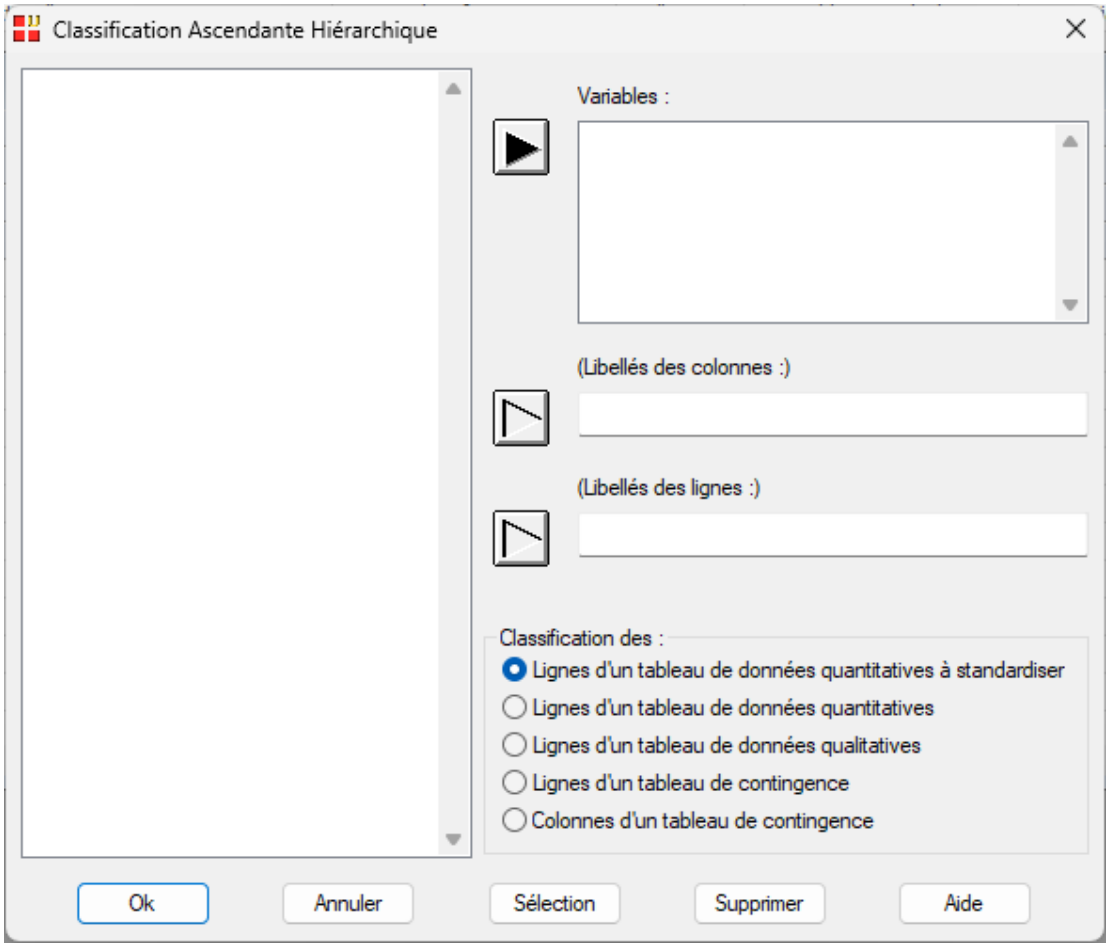
La méthode utilisée est celle de Ward, méthode dite des voisins réciproques. C'est une méthode ascendante partant d'un nombre de classes égal au nombre des individus et regroupant à chaque itération un (ou plusieurs) individus ou classes en fonction de la diminution de l'inertie inter-classes. L'algorithme s'applique à des données quantitatives individus x variables (standardisées ou non en utilisant la distance euclidienne au carré), à un tableau de contingence ou à un tableau de variables qualitatives (en utilisant la distance du khi2). Dans le cas d'un tableau de contingence, la classification peut se faire suivant les lignes ou les colonnes.

Un tableau résumé de la classification et le tracé de l'arbre de classification (avec zoom et troncature) et de nuages codifiés sont proposés ainsi que des outils d'analyse de la partition et des nœuds.

Cette méthode peut notamment s'associer à une Analyse en Composantes Principales (ACP) et à une Analyse des Correspondances Multiples (ACM).

Entrée des données

Cliquons sur l'icône CAH dans le ruban Décrire. La boîte de dialogue montrée ci-dessous s'affiche :



Cette boîte de dialogue permet de choisir les variables qui serviront à la classification, la variable contenant les libellés des lignes et la variable contenant les libellés des colonnes.

Elle permet également de préciser si la méthode est utilisée sur des variables quantitatives (métriques), sur des variables qualitatives ou sur un tableau de contingence

Dans le premier cas, il est proposé de faire la classification sur les données standardisées (centrées et réduites) ou sur les données brutes.

Dans le troisième, il est proposé de faire la classification par rapport aux lignes ou par rapport aux colonnes du tableau de contingence.

Données manquantes

- Les données manquantes sont autorisées si l'analyse porte sur des données quantitatives (métriques).
- Les données manquantes ne sont pas autorisées si l'analyse se fait sur un tableau de contingence ou de données qualitatives.

Exemple 1 : Fichier VEHICULE

Pour illustrer cet exemple, nous utiliserons le fichier VEHICULE. Ce fichier contient 7 informations caractérisant 24 véhicules :

- *Modèle*
- *Cylindrée*
- *Puissance*
- *Vitesse*
- *Poids*
- *Longueur*
- *Largeur*

Les libellés des variables quantitatives sont dans la variable *Mesures.*

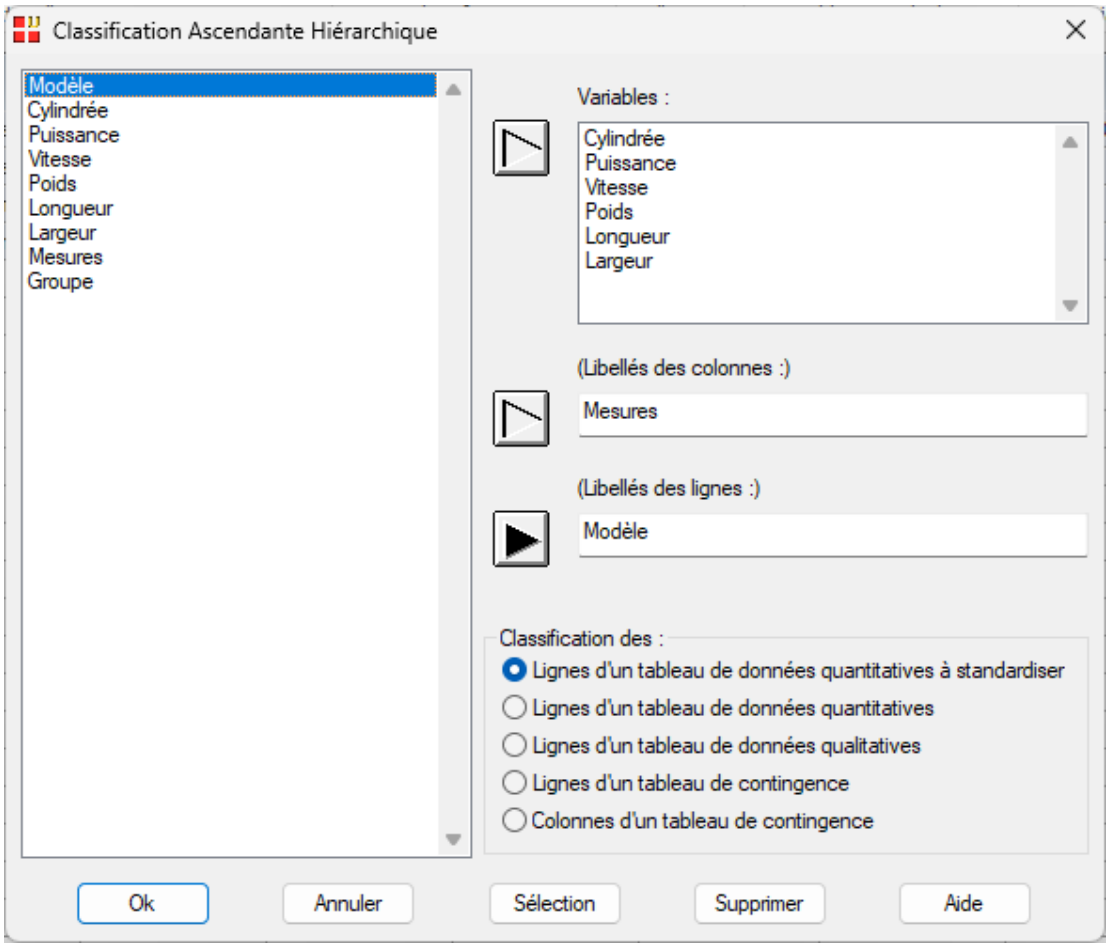
Les 24 modèles d'automobiles sont :

Honda Civic	R19	Fiat Tipo	405
R21	BX	BMW 530i	Rover 827i
R25	Opel Omega	405 Break	Ford Sierra
BMW 325ix	Audi 90 Quattro	Ford Scorpio	Espace
Nissan Vanette	VW Caravelle	Ford Fiesta	Fiat Uno
205	205 Rallye	Seat Ibiza SXI	AX Sport

Cliquons sur l'icône CAH dans le ruban Décrire.

La boîte de dialogue montrée ci-après s'affiche.

Nous choisissons toutes les variables de *Cylindrée* à *Largeur* comme variables quantitatives, la variable *Mesures* comme variable contenant les libellés associés et la variable *Modèle* comme variable contenant les libellés des individus.



Nous choisissons de faire une classification sur un tableau de données quantitatives à standardiser.

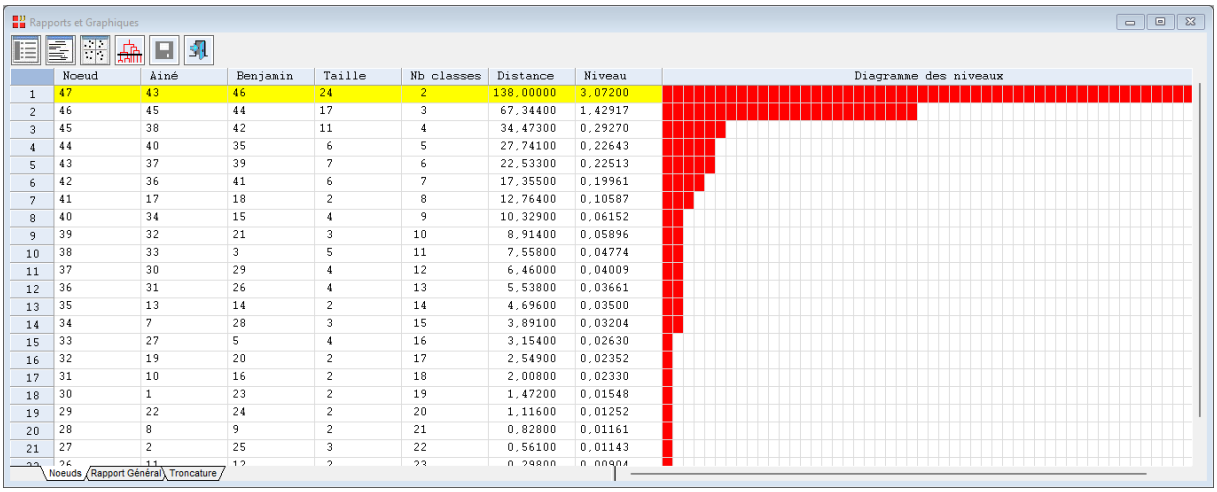
Cliquons sur Ok. UNIWIN débute le calcul de la classification ascendante hiérarchique.

Une jauge s'affiche montrant la progression des calculs. Ils peuvent en effet prendre quelques instants si les données sont nombreuses.

Après quelques instants, UNIWIN affiche le tableau résumé de la classification (onglet 'Nœuds' du tableur).

Cet écran nous affiche les informations suivantes :

- Noeud
- Aîné
- Benjamin
- Taille
- Distance
- Indice
- Diagramme des indices



	Noeud	Aîné	Benjamin	Taille	Nb classes	Distance	Niveau	Diagramme des niveaux
1	47	43	46	24	2	138,00000	3,07200	
2	46	45	44	17	3	67,34400	1,42917	
3	45	38	42	11	4	34,47300	0,29270	
4	44	40	35	6	5	27,74100	0,22643	
5	43	37	39	7	6	22,53300	0,22513	
6	42	36	41	6	7	17,35500	0,19961	
7	41	17	18	2	8	12,76400	0,10587	
8	40	34	15	4	9	10,32900	0,06152	
9	39	32	21	3	10	8,91400	0,05896	
10	38	33	3	5	11	7,55800	0,04774	
11	37	30	29	4	12	6,46000	0,04009	
12	36	31	26	4	13	5,53800	0,03661	
13	35	13	14	2	14	4,69600	0,03500	
14	34	7	28	3	15	3,89100	0,03204	
15	33	27	5	4	16	3,15400	0,02630	
16	32	19	20	2	17	2,54900	0,02352	
17	31	10	16	2	18	2,00800	0,02330	
18	30	1	23	2	19	1,47200	0,01548	
19	29	22	24	2	20	1,11600	0,01252	
20	28	8	9	2	21	0,82800	0,01161	
21	27	2	25	3	22	0,56100	0,01143	

Par exemple le nœud 46 est formé de la réunion des véhicules des nœuds 45 et 44, il regroupe 17 des 24 véhicules à une distance d'environ 67,344, un niveau de 1,42917.

L'indice de niveau est calculé de la façon suivante :

(Distance niveau i – Distance niveau i-1) / (n - 1) où n = nombre d'individus

La ligne éclairée dans le tableau 'Nœuds' indique le nœud sélectionné.

Par défaut le nœud sélectionné est celui qui regroupe la totalité des individus.

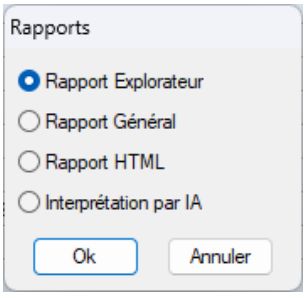
Vous pouvez bien évidemment modifier cette sélection en cliquant sur la ligne désirée, soit dans la partie droite de cet écran, soit dans la partie gauche.

Cette sélection est importante pour la suite de l'analyse.

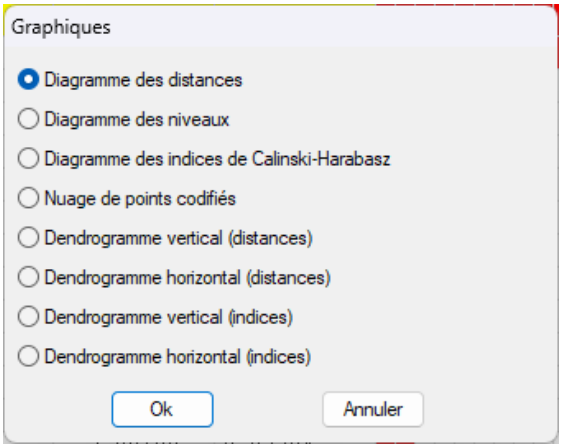
En effet, elle vous permettra de visualiser le dendrogramme, les nuages codifiés et de faire des troncatures à partir du nœud que vous désirez.

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

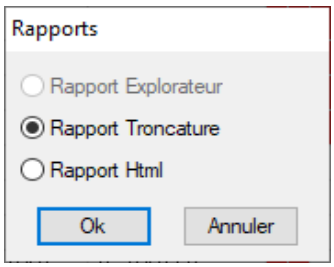
L'icône 'Rapports' affiche la boîte de dialogue des options pour les rapports :



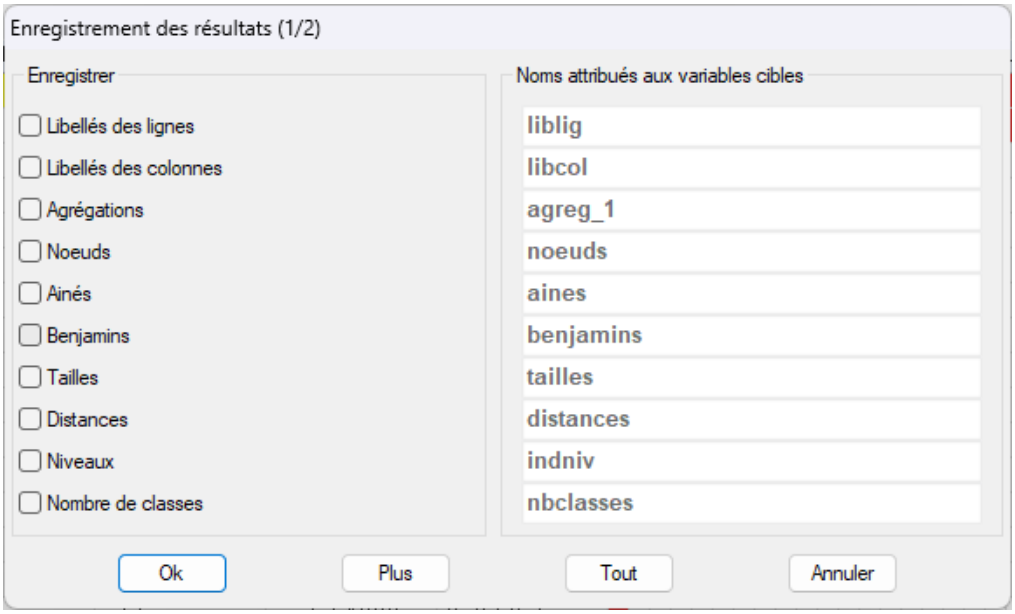
et l'icône 'Graphiques' affiche la boîte de dialogue des options pour les graphiques.



L'icône 'Troncature' affiche la boîte de dialogue des options pour les rapports associés à la troncature :



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



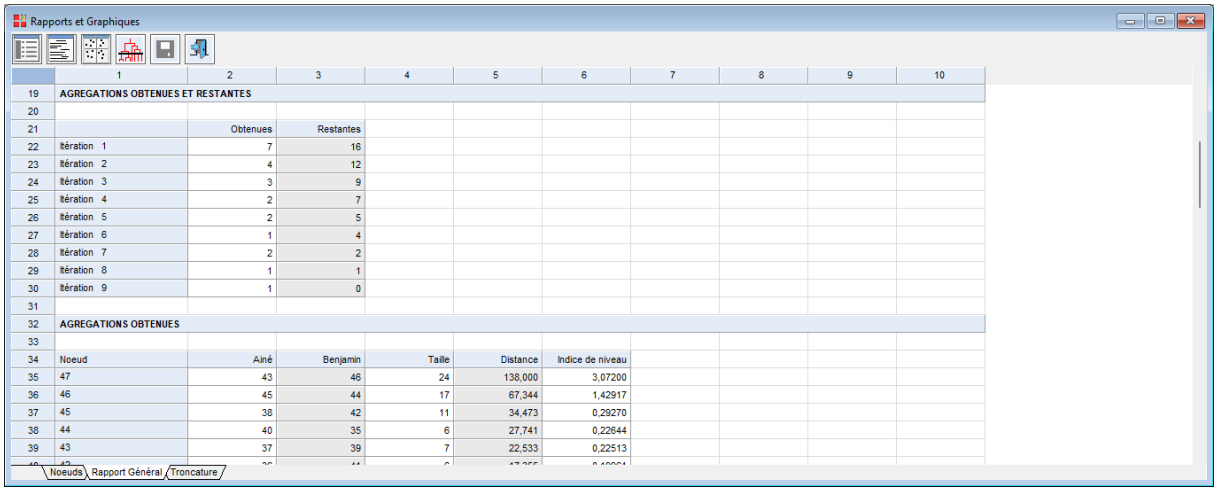
L'icône 'Quitter' permet de quitter l'analyse.

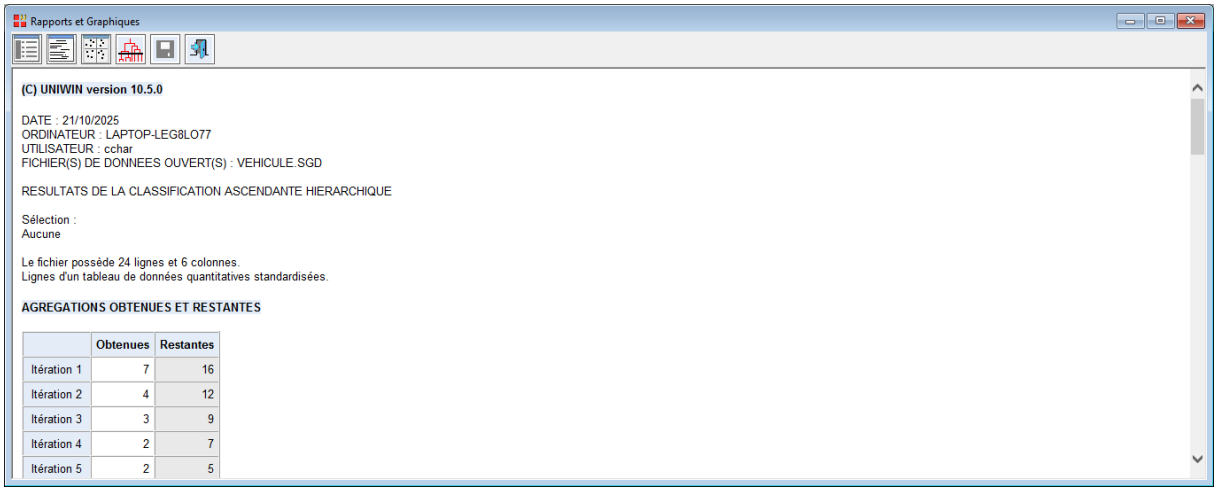
L'option Rapports

Cette option permet d'obtenir le rapport à l'écran sous la forme Explorateur, Général ou HTML et d'accéder à l'aide à l'interprétation par Intelligence Artificielle.

L'impression des rapports fait appel à la procédure 'Aperçu avant impression'. Pour des informations sur cette procédure, voir le 'Manuel de l'Utilisateur'.

Voici deux exemples du rapport pour notre CAH : Général, HTML.





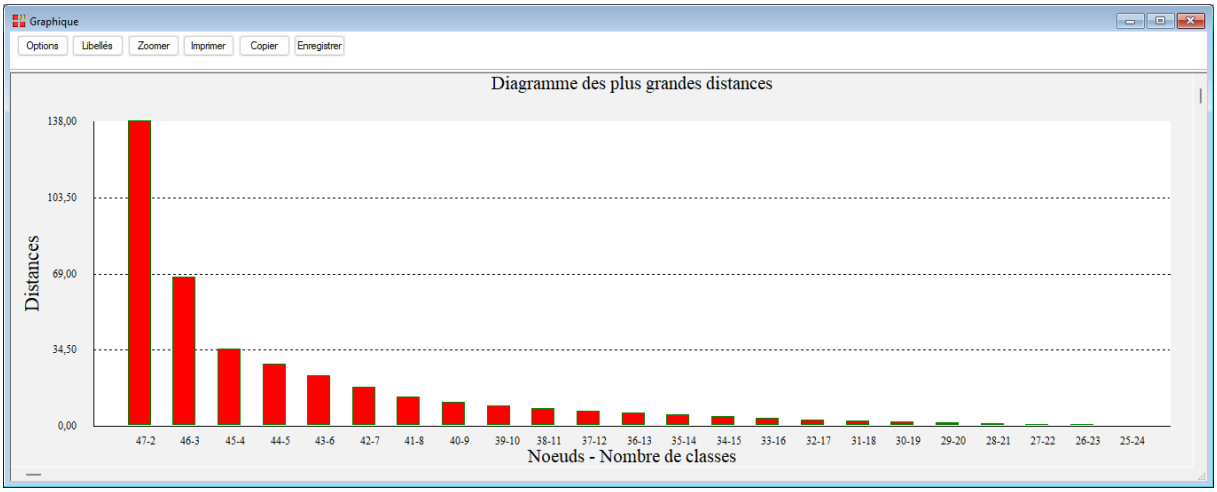
Ce rapport affiche les différentes itérations de la classification (agrégations obtenues et restantes à chaque itération), les agrégations finales (nœud, aîné, benjamin, taille, distance d'agrégation, indice de niveau, indice de Calinski-Harabasz), les compositions de chacun des nœuds (nœud, distance d'agrégation, fils du nœud).

L'option Graphiques

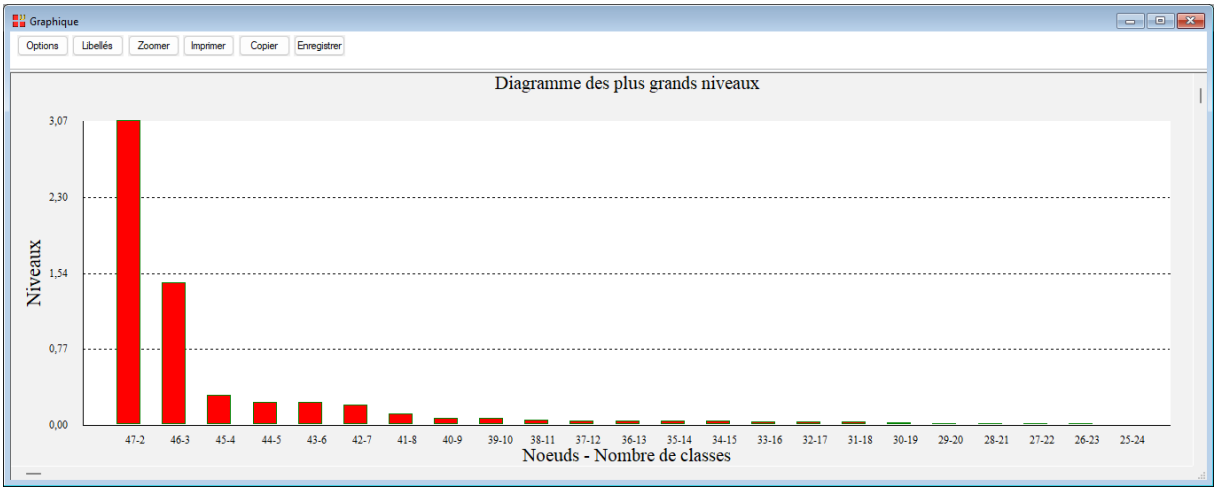
Cette option permet de visualiser le diagramme des distances, le diagramme des indices de niveau, le diagramme des indices de Calinski-Harabasz, des nuages de points codifiés, des dendrogrammes horizontaux ou verticaux des distances ou des indices.

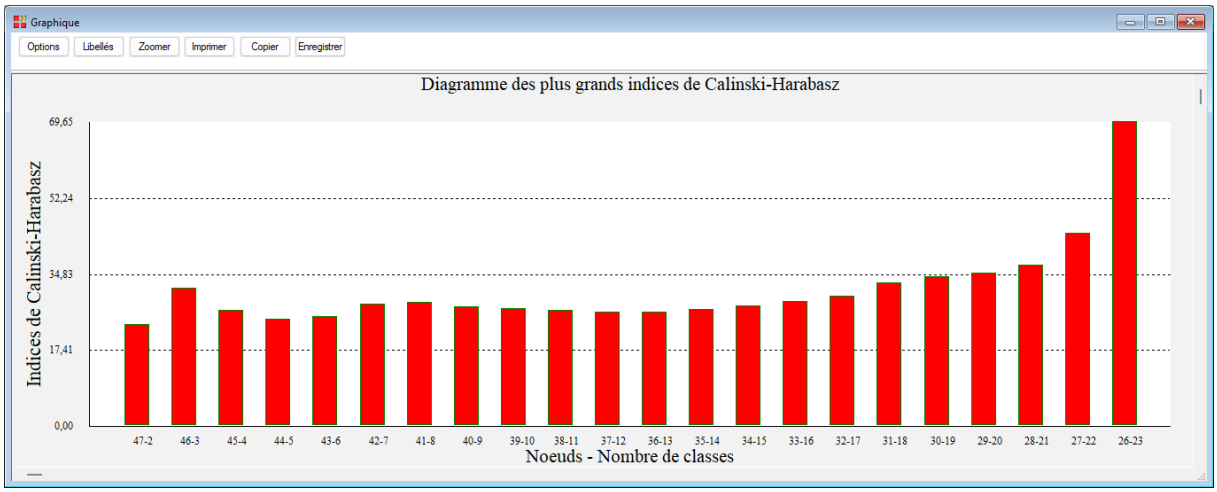
- Les options Diagramme des distances, des indices de niveau et des indices de Calinski-Harabasz

Le diagramme des distances permet de repérer le ou les meilleurs nœuds pour effectuer des troncatures.



Le diagramme des indices de niveau permet également de repérer le ou les meilleurs nœuds pour effectuer des troncatures.





Ce diagramme affiche les comparaisons entre les sommes des carrés entre les classes et les sommes des carrés dans les classes. Les plus grandes valeurs de cet indice sont généralement préférables

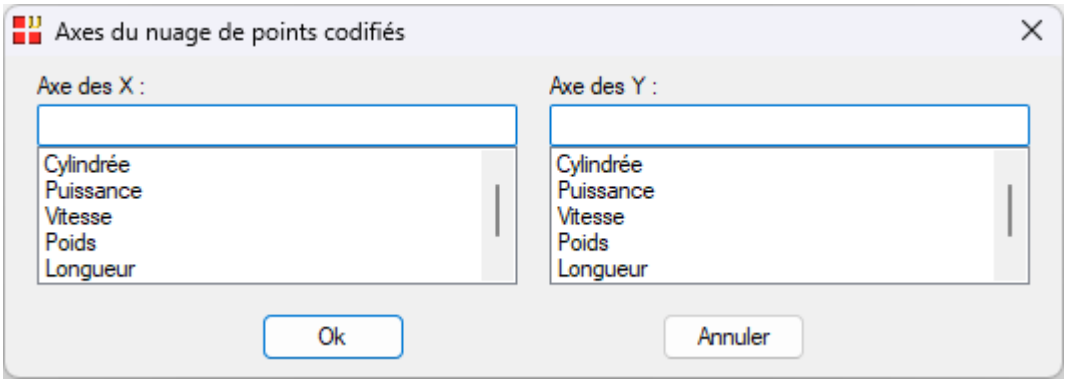
- L'option Nuage de points codifiés

Ils permettent de visualiser les individus et les groupes auxquels ils appartiennent par rapport à deux des colonnes d'origine.

Le graphique créé par cette option dépend du nœud sélectionné c'est-à-dire de la ligne éclairée dans le tableau 'Nœuds'.

Choisissons le nœud 47 dans ce tableau et cliquons sur 'Nuage de points codifiés'.

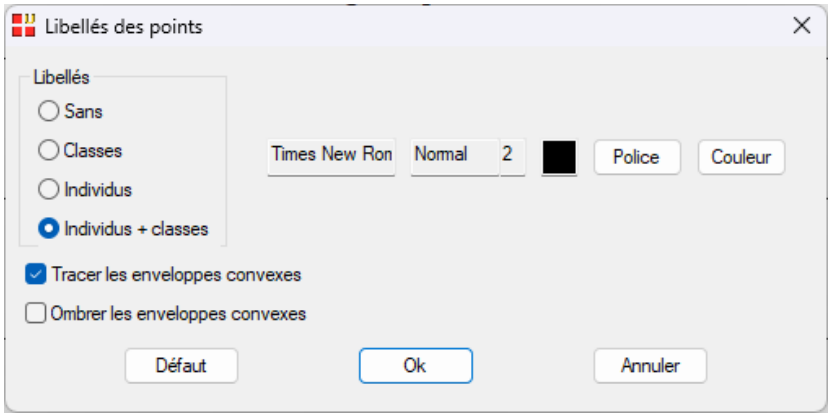
Après avoir choisi cette option la boîte de dialogue suivante s'affiche :



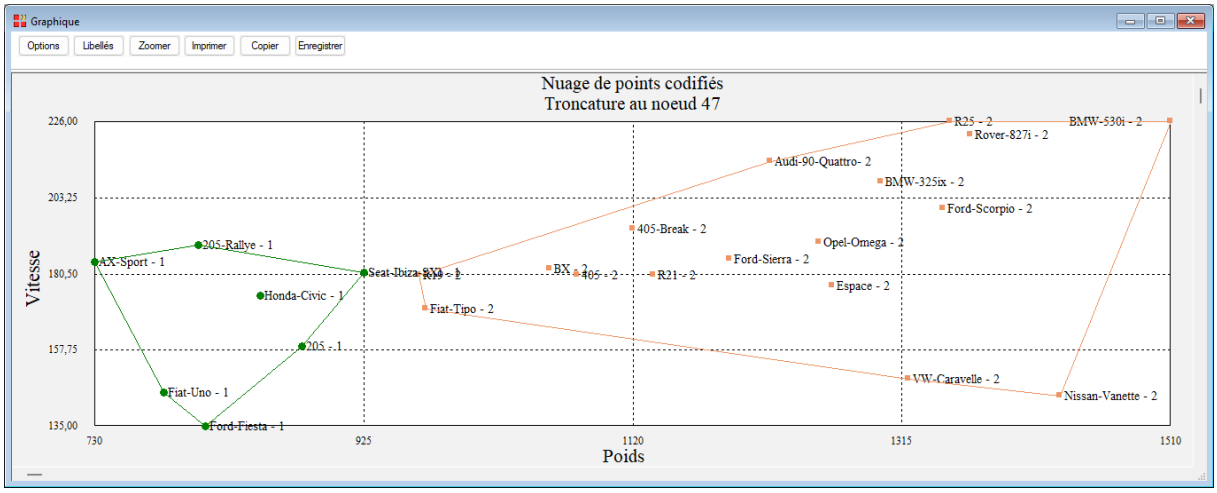
Cette boite de dialogue permet de choisir les données à utiliser pour les axes X et Y.

Choisissons 'Poids' pour l'axe des X et 'Vitesse' pour l'axe des Y. Le graphique s'affiche.

Cliquons sur 'Libellés' et choisissons 'Individus + classes' et demandons le tracé des enveloppes convexes.



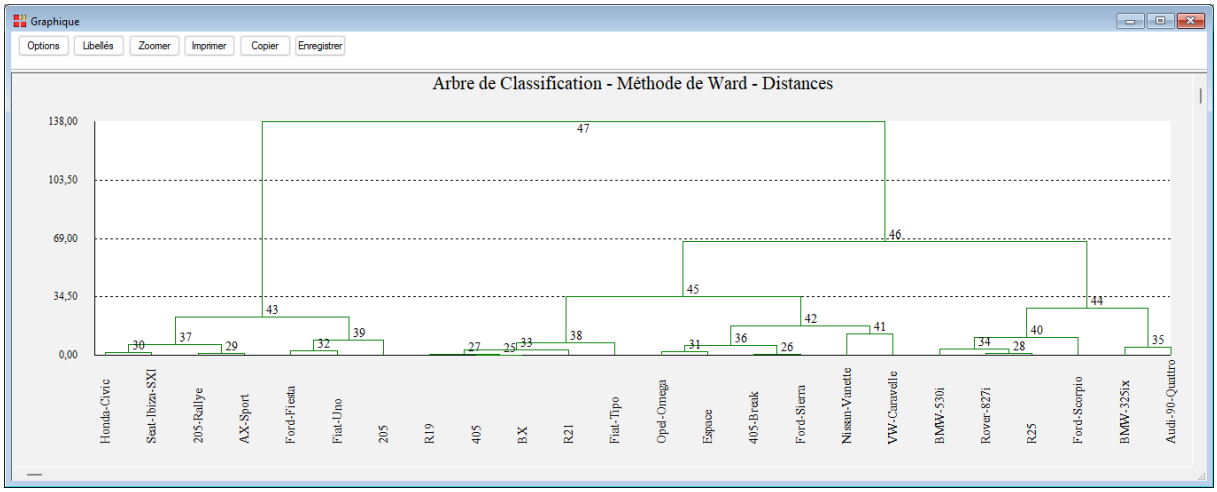
Au maximum huit classes peuvent être représentées. Si la sélection du nœud forme plus de huit classes, un message apparaît indiquant qu'un unique symbole sera utilisé.

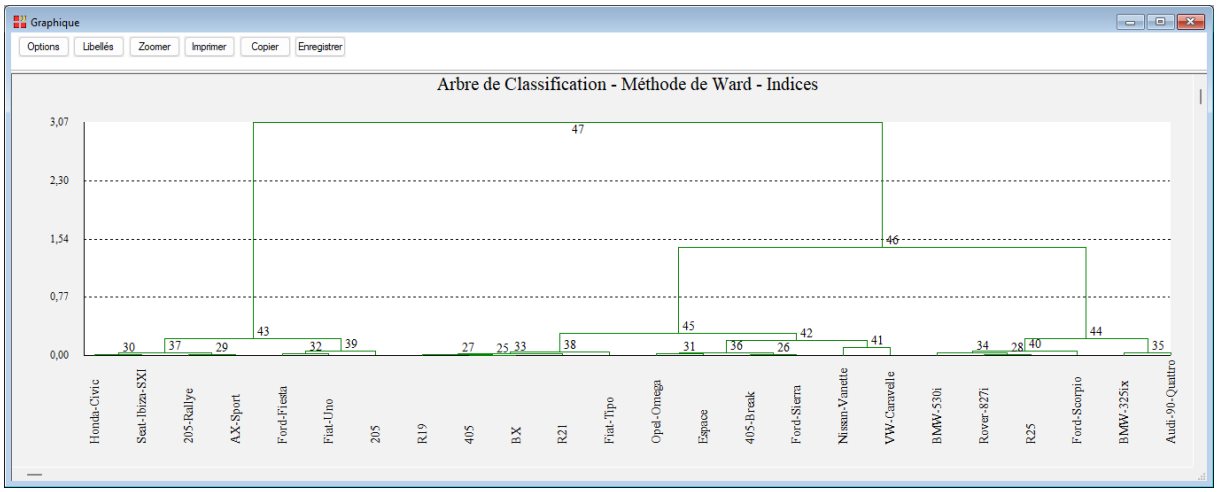


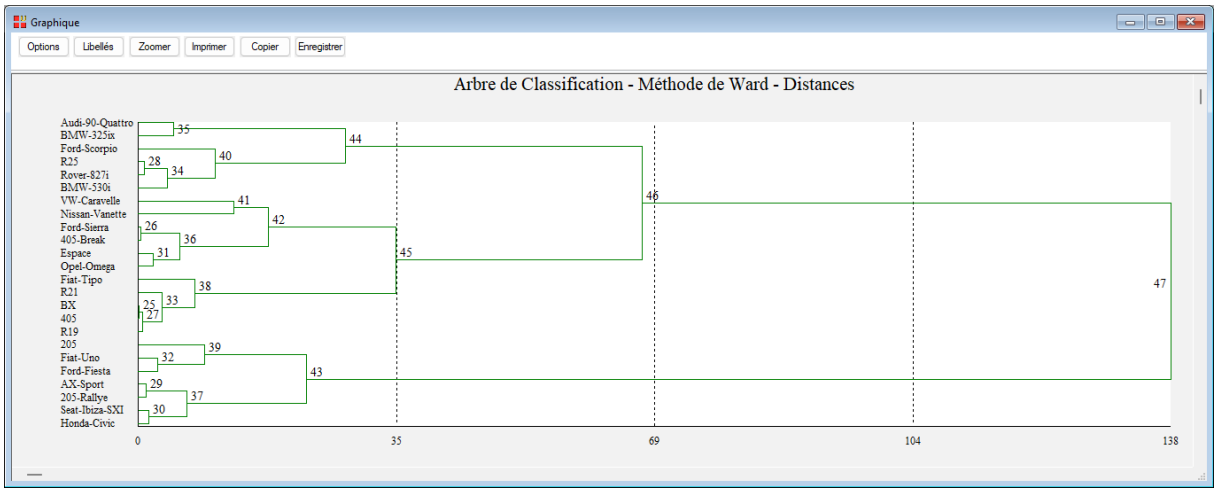
- Les options Dendrogramme

Ces options permettent de tracer le dendrogramme ou arbre de classification.

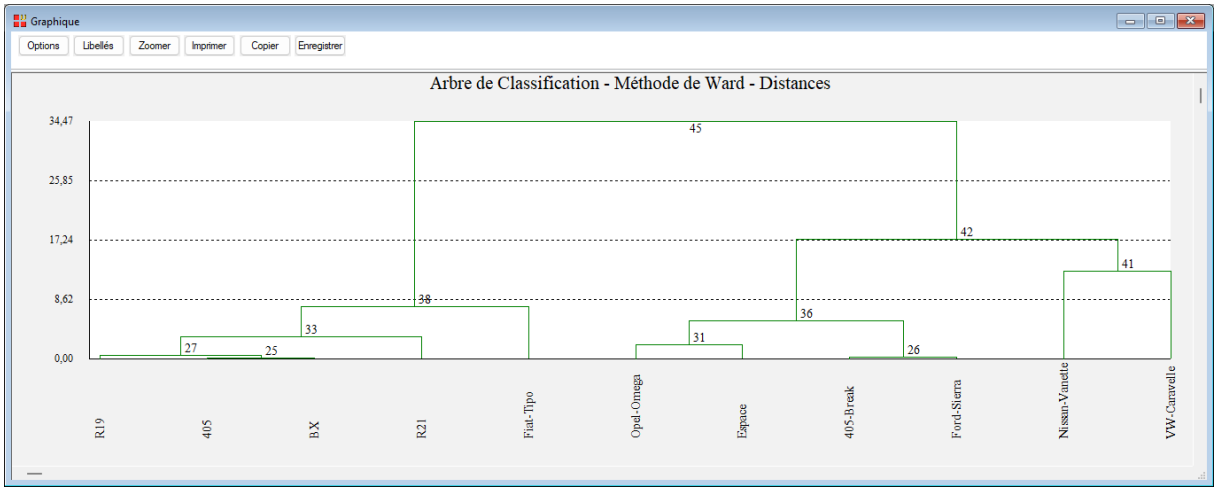
Ce tracé peut être vertical ou horizontal, pour les distances ou les indices de niveau.







Le graphique créé par cette option dépend du nœud sélectionné c'est-à-dire de la ligne éclairée dans le tableau 'Nœuds'. Revenons au tableur 'Nœuds', cliquons sur le nœud 45 puis demandons à nouveau le tracé du dendrogramme. Par cette approche, nous pouvons visualiser de façon plus précise toute sous-partie de l'arbre et ci-dessous les fils du nœud 45.



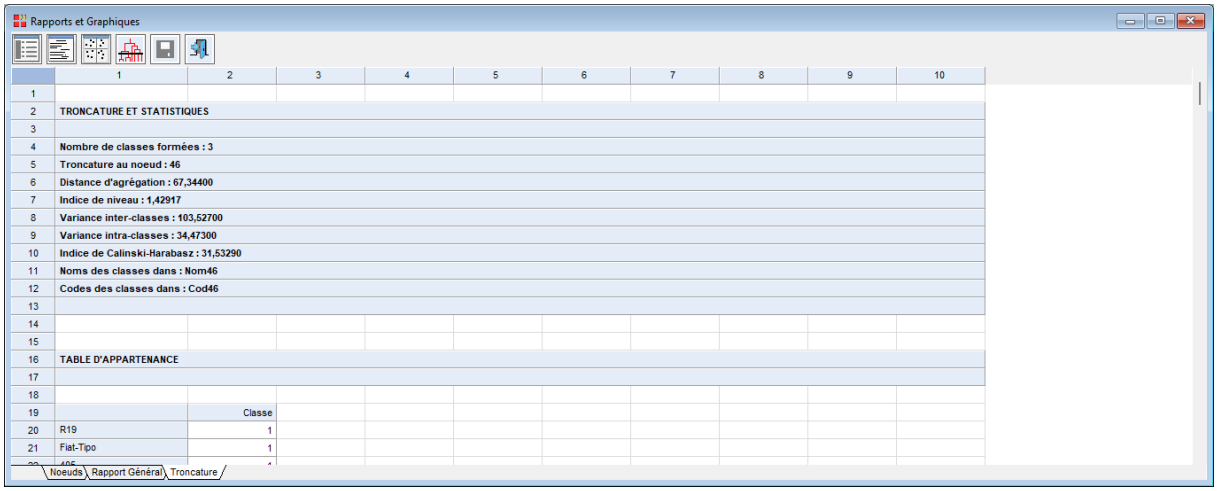
- L’option Troncature

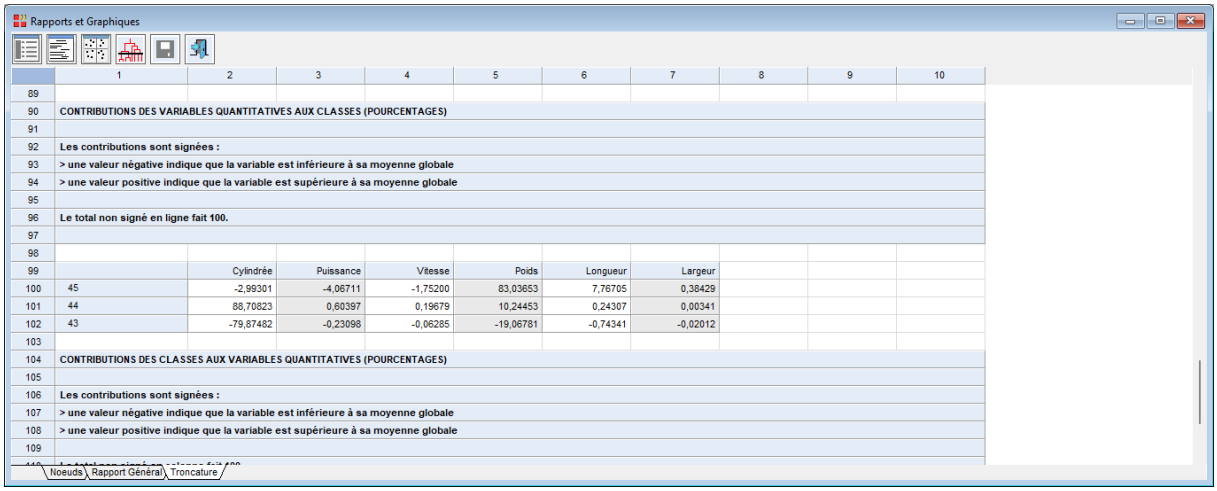
Cette option permet de faire une troncature c’est-à-dire à partir d’un niveau donné de la distance d’agrégation, de créer des classes d’individus repérées par les valeurs d’une variable codée indiquant pour chaque individu sa classe d’appartenance.

La troncature créée par cette option dépend du nœud sélectionné c’est-à-dire de la ligne éclairée dans le tableau 'Nœuds'. Cette option permet un ensemble de statistiques descriptives sur les classes formées par cette troncature.

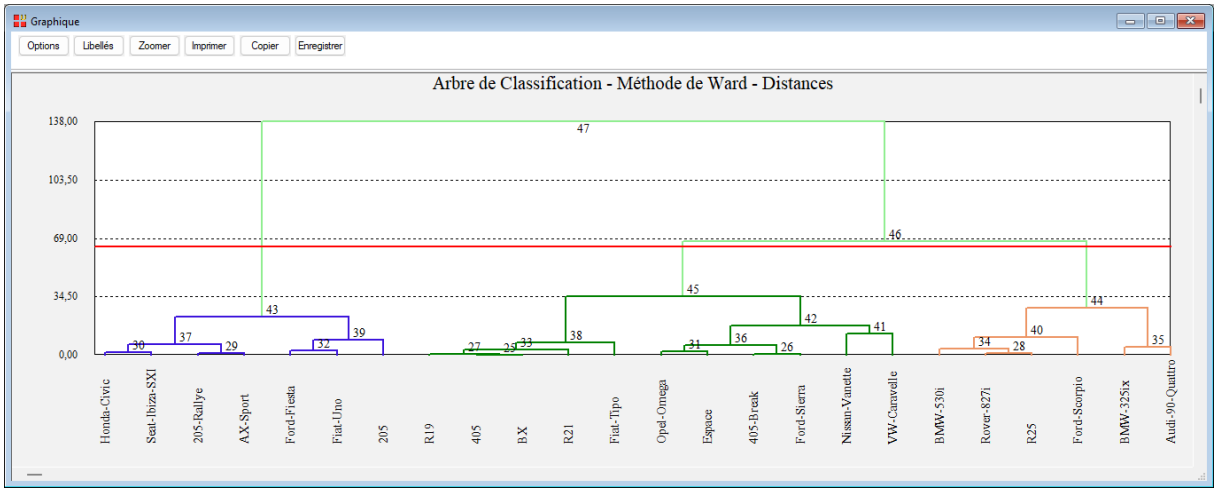
Sélectionnons le nœud 46 dans le tableau 'Nœuds' et demandons la troncature. Un rapport s’affiche indiquant le nœud sélectionné pour la troncature, la distance d’agrégation associée, le nombre de classes formées, la variable contenant les noms des classes et la variable codée indiquant pour chaque individu sa classe d’appartenance. Ici les variables s’appellent Nom46 et Code46.

Le rapport donne pour chaque classe formée et pour chaque variable différentes statistiques descriptives. Ces statistiques permettent de mieux caractériser les groupes issus de la classification. De plus les contributions, signées et exprimées en pourcentages, des variables aux classes et des classes aux variables sont données.





Demandons à nouveau le dendrogramme complet à partir du nœud 47.



Chaque classe apparaît avec une couleur spécifique et une ligne horizontale rappelle le niveau de la troncature.

Exemple 2 : Fichier SUICIDE

Pour illustrer ce deuxième exemple, nous utiliserons le fichier SUICIDE.

Ce fichier contient les nombres de suicides (pour 100 000 habitants) dans 19 pays pour 6 causes différentes.

Il s'agit donc ici d'un tableau de contingence dans lequel les lignes sont les pays et les colonnes les différentes causes.

Les pays sont :

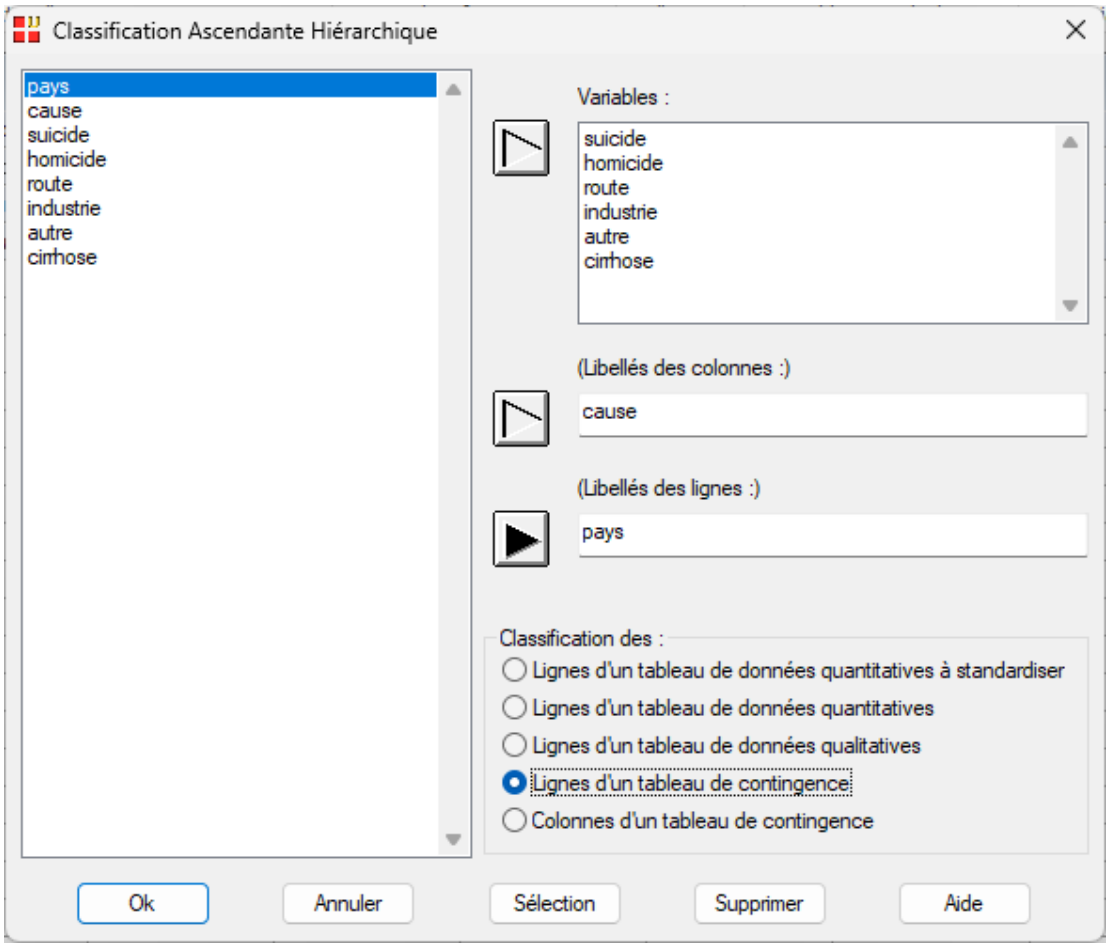
Autriche	France	Portugal	Allemagne
Belgique	Finlande	Suède	Suisse
Italie	Irlande nord	Danemark	Islande
Ecosse	Espagne	Norvège	Irlande sud
Pays Bas	Angleterre	Usa	

Les six causes sont :

Suicide	Homicide	Route	Industrie
Autres	Cirrhose		

Cliquons sur l'icône CAH dans le ruban Décrire.

La boîte de dialogue montrée ci-après s'affiche.



Nous choisissons toutes les variables de *suicide* à *cirrhose* comme variables définissant le tableau de contingence.

Nous choisissons la variable *cause* comme variable contenant les libellés associés et la variable *pays* comme variable contenant les libellés des individus.

Nous choisissons de faire une analyse sur les lignes d'un tableau de contingence.

Cliquons sur Ok.

Voici quelques exemples des résultats obtenus : classification, dendrogrammes vertical et horizontal, nuage codifié, statistiques et contributions des variables aux classes (coefficient de Cramer).

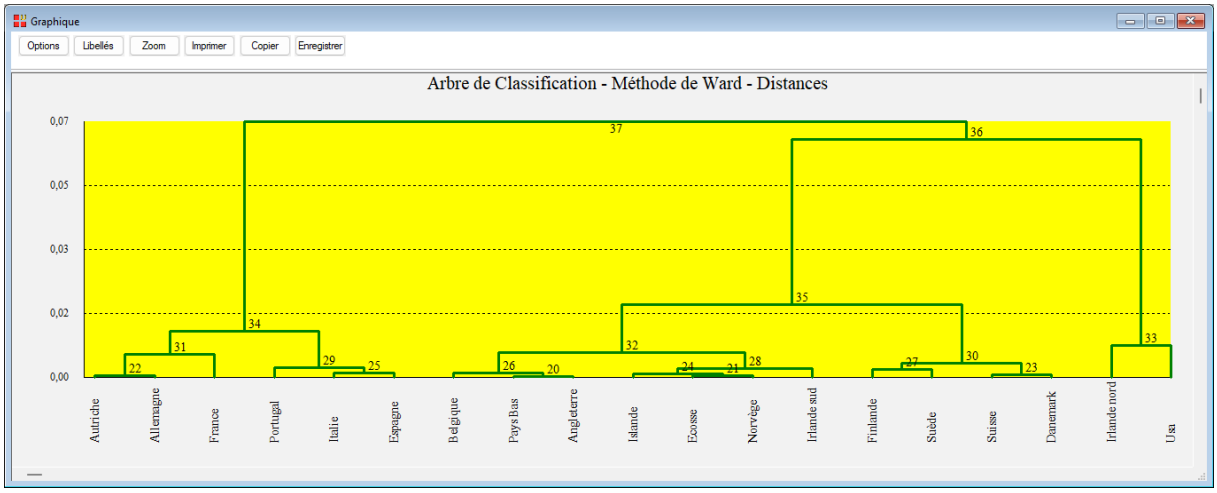
L'indice de niveau est calculé de la façon suivante :

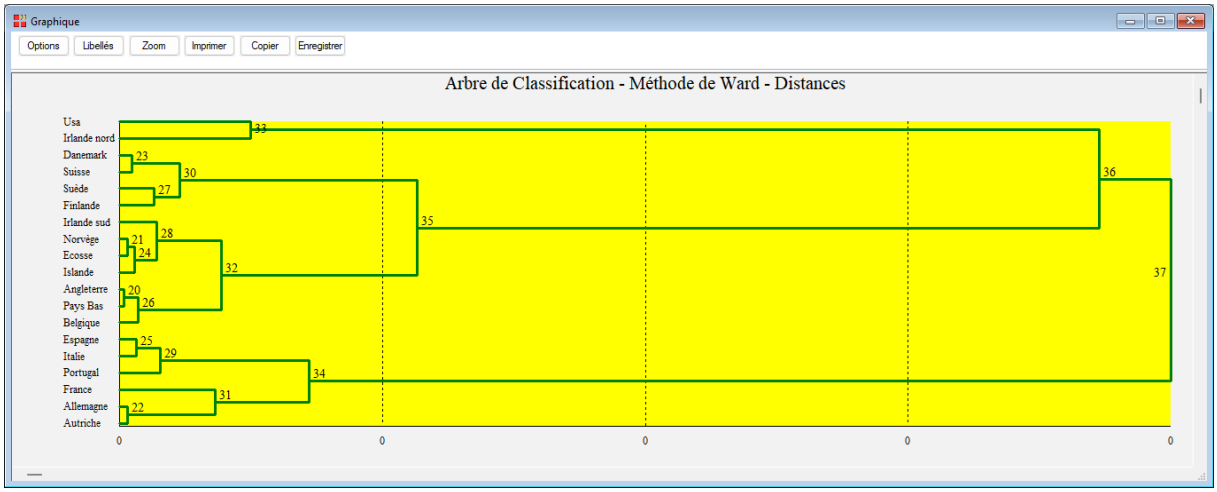
100 x (Distance niveau i / Somme des distances)

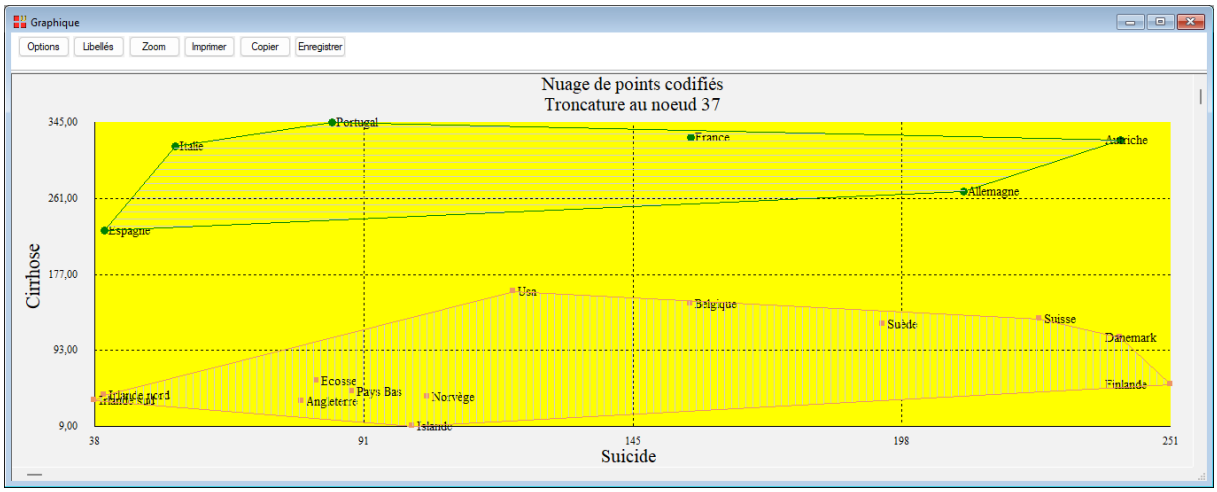
Rapports et Graphiques

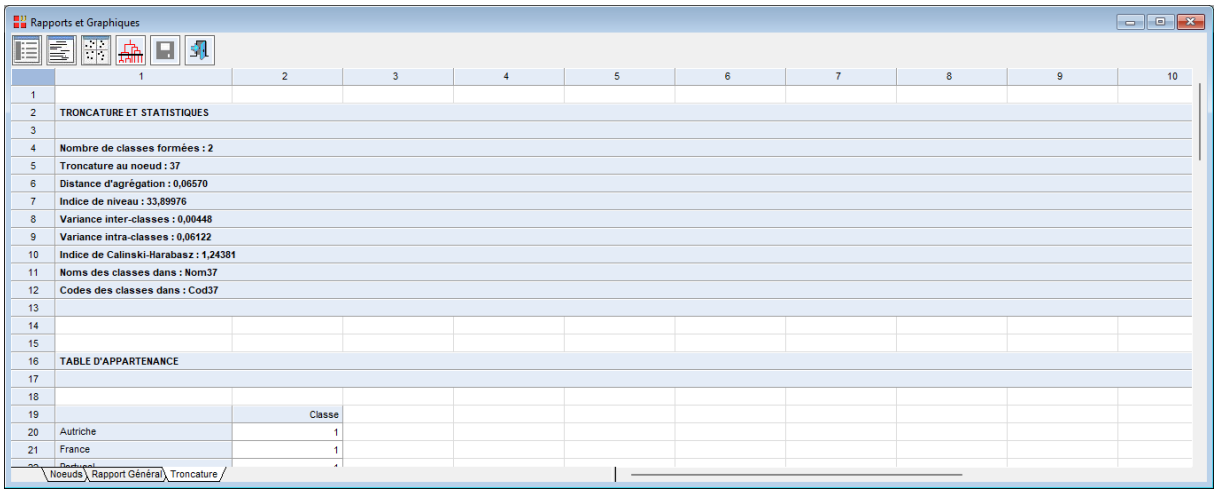
	Noeud	Aîné	Benjamin	Taille	Nb classes	Distance	Indice	Diagramme des indices
1	37	34	36	19	2	0,06570	33,89976	
2	36	35	33	13	3	0,06122	31,58857	
3	35	32	30	11	4	0,01861	9,60355	
4	34	31	29	6	5	0,01184	6,11181	
5	33	10	19	2	6	0,00818	4,22205	
6	32	26	28	7	7	0,00638	3,29096	
7	31	22	2	3	8	0,00596	3,07780	
8	30	27	23	4	9	0,00375	1,93700	
9	29	3	25	3	10	0,00254	1,31053	
10	28	24	16	4	11	0,00230	1,18570	
11	27	6	7	2	12	0,00213	1,10009	
12	26	5	20	3	13	0,00116	0,59835	
13	25	9	14	2	14	0,00105	0,54031	
14	24	12	21	3	15	0,00094	0,48711	
15	23	8	11	2	16	0,00079	0,40809	
16	22	1	4	2	17	0,00050	0,25712	
17	21	13	15	2	18	0,00048	0,24829	
18	20	17	18	2	19	0,00026	0,13291	

Noeuds | Rapport Général | Troncature

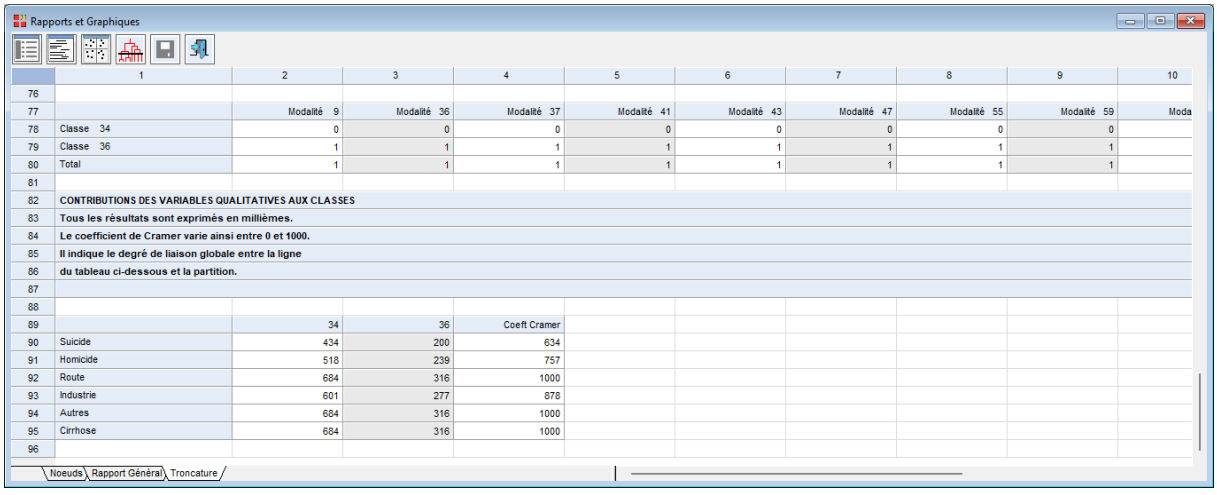








	1	2	3	4	5	6	7	8	9	10
1										
2	TRONCATURE ET STATISTIQUES									
3										
4	Nombre de classes formées : 2									
5	Troncature au noeud : 37									
6	Distance d'agrégation : 0,06570									
7	Indice de niveau : 33,89976									
8	Variance inter-classes : 0,00448									
9	Variance intra-classes : 0,06122									
10	Indice de Calinski-Harabasz : 1,24381									
11	Noms des classes dans : Nom37									
12	Codes des classes dans : Cod37									
13										
14										
15										
16	TABLE D'APPARTENANCE									
17										
18										
19		Classe								
20	Autriche	1								
21	France	1								



	1	2	3	4	5	6	7	8	9	10
76										
77		Modalité 9	Modalité 36	Modalité 37	Modalité 41	Modalité 43	Modalité 47	Modalité 55	Modalité 59	Moda
78	Classe 34	0	0	0	0	0	0	0	0	
79	Classe 36	1	1	1	1	1	1	1	1	
80	Total	1	1	1	1	1	1	1	1	
81										
82	CONTRIBUTIONS DES VARIABLES QUALITATIVES AUX CLASSES									
83	Tous les résultats sont exprimés en millièmes.									
84	Le coefficient de Cramer varie ainsi entre 0 et 1000.									
85	Il indique le degré de liaison globale entre la ligne									
86	du tableau ci-dessous et la partition.									
87										
88										
89		34	36	Coeft Cramer						
90	Suicide	434	200	634						
91	Homicide	518	239	757						
92	Route	684	316	1000						
93	Industrie	601	277	878						
94	Autres	684	316	1000						
95	Cirrhose	684	316	1000						
96										

Exemple 3 : Fichier PHYTOS

Pour illustrer ce troisième exemple, nous utiliserons le fichier PHYTOS.

Ce fichier contient les données d'une étude des affinités de terrain entre espèces végétales. Des relevés ont été effectués sur 16 lieux. Chaque relevé indique pour un lieu donné la présence (1) ou l'absence (0) d'une des 66 espèces végétales.

L'objectif de l'étude est de vérifier le bien-fondé de la classification des pelouses "à nard" (du nom de l'espèce dominante), c'est-à-dire des lieux.

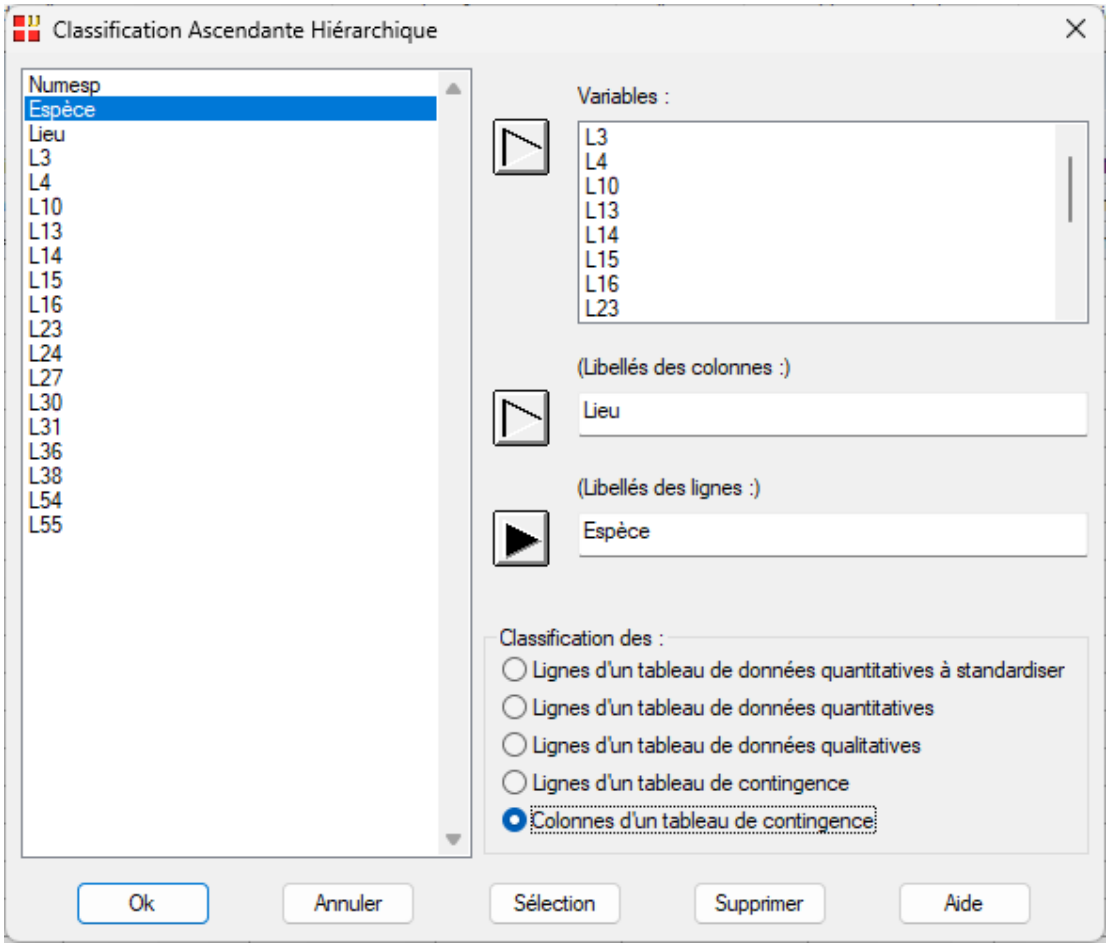
Cliquons sur l'icône CAH dans le ruban Décrire.

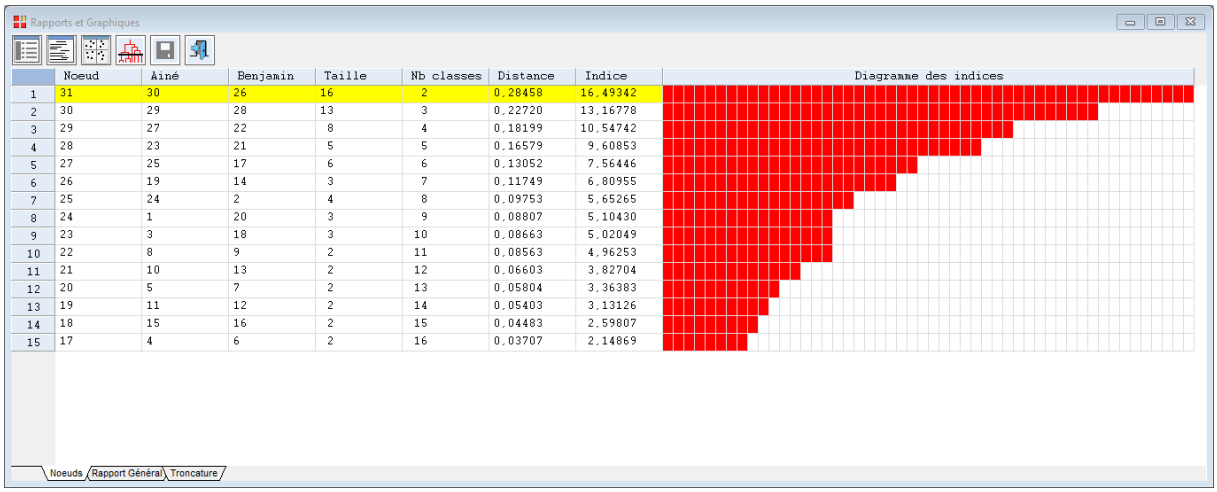
Nous choisissons toutes les variables de *L3* à *L55* comme variables définissant le tableau de contingence.

Nous choisissons la variable *Lieu* comme variable contenant les libellés associés et la variable *Espèce* comme variable contenant les libellés des lignes.

Nous choisissons de faire une analyse sur les colonnes d'un tableau de contingence.

La boîte de dialogue montrée ci-après s'affiche.

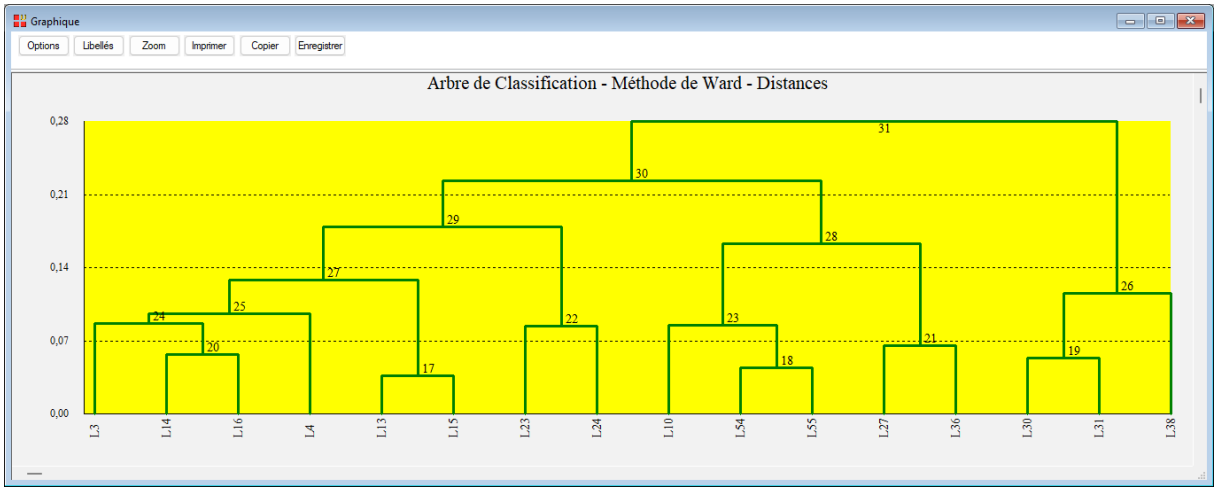


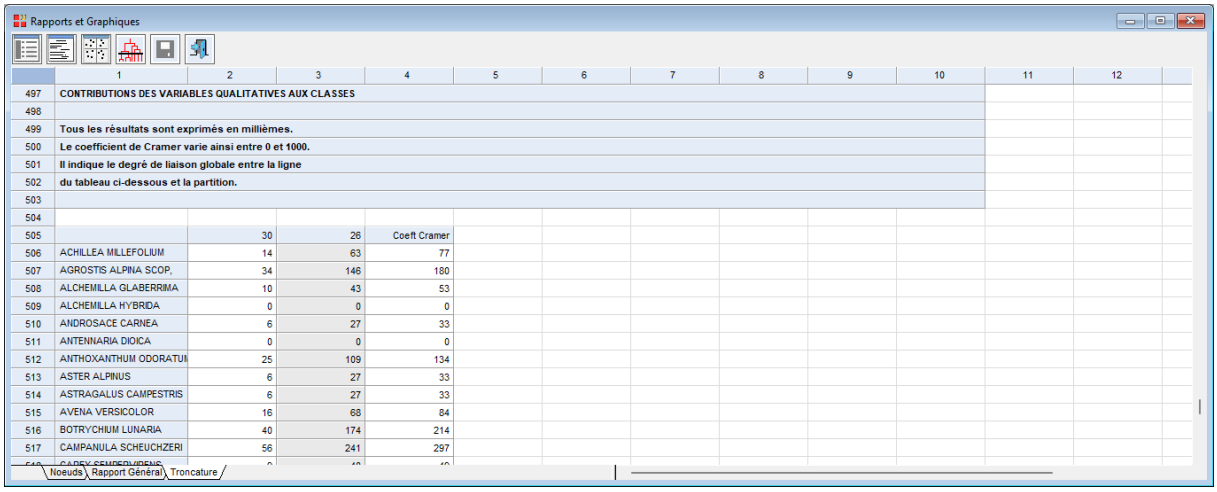


	Noeud	Aîné	Benjamin	Taille	Nb classes	Distance	Indice	Diagramme des indices
1	31	30	26	16	2	0,28458	16,49342	
2	30	29	28	13	3	0,22720	13,16778	
3	29	27	22	8	4	0,18199	10,54742	
4	28	23	21	5	5	0,16579	9,60853	
5	27	25	17	6	6	0,13052	7,56446	
6	26	19	14	3	7	0,11749	6,80955	
7	25	24	2	4	8	0,09753	5,65265	
8	24	1	20	3	9	0,08807	5,10430	
9	23	3	18	3	10	0,08663	5,02049	
10	22	8	9	2	11	0,08563	4,96253	
11	21	10	13	2	12	0,06603	3,82704	
12	20	5	7	2	13	0,05804	3,36383	
13	19	11	12	2	14	0,05403	3,13126	
14	18	15	16	2	15	0,04483	2,59807	
15	17	4	6	2	16	0,03707	2,14869	

L'indice de niveau est calculé de la façon suivante :

100 x (Distance niveau i / Somme des distances)





	1	2	3	4
497	CONTRIBUTIONS DES VARIABLES QUALITATIVES AUX CLASSES			
498				
499	Tous les résultats sont exprimés en milièmes.			
500	Le coefficient de Cramer varie ainsi entre 0 et 1000.			
501	Il indique le degré de liaison globale entre la ligne			
502	du tableau ci-dessous et la partition.			
503				
504				
505		30	26	Coeft Cramer
506	ACHILLEA MILLEFOLIUM	14	63	77
507	AGROSTIS ALPINA SCOP.	34	146	180
508	ALCHEMILLA GLABERRIMA	10	43	53
509	ALCHEMILLA HYBRIDA	0	0	0
510	ANDROSACE CARNEA	6	27	33
511	ANTENNARIA DIOICA	0	0	0
512	ANTHOXANTHUM ODORATUI	25	109	134
513	ASTER ALPINUS	6	27	33
514	ASTRAGALUS CAMPESTRIS	6	27	33
515	AVENA VERSICOLOR	16	68	84
516	BOTRYCHIUM LUNARIA	40	174	214
517	CAMPANULA SCHEUCHZERI	56	241	297

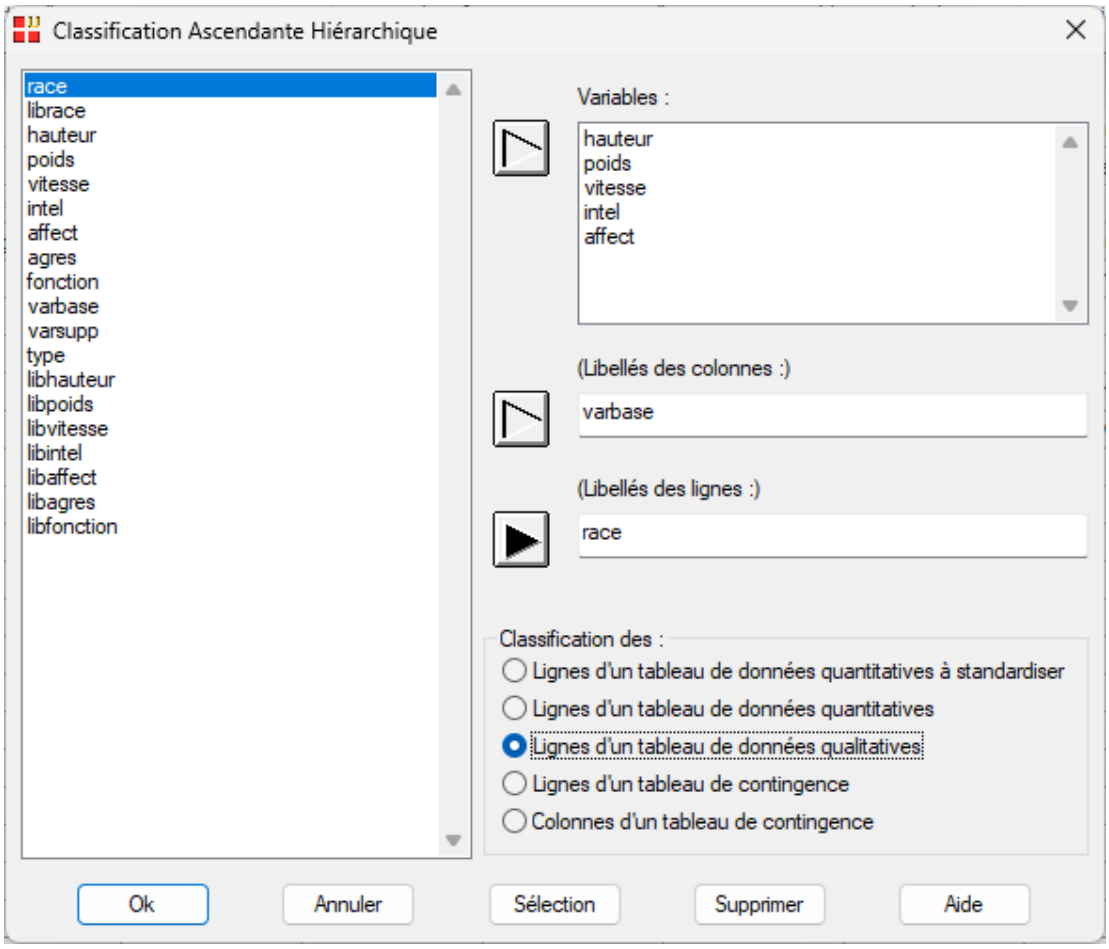
Exemple 4 : Fichier CHIENS

Pour illustrer ce quatrième exemple, nous utiliserons le fichier CHIENS.

Pour une description du contenu de ce fichier, voir le document 'Analyse des Correspondances Multiples'.

Cliquons sur l'icône CAH dans le ruban Décrire.

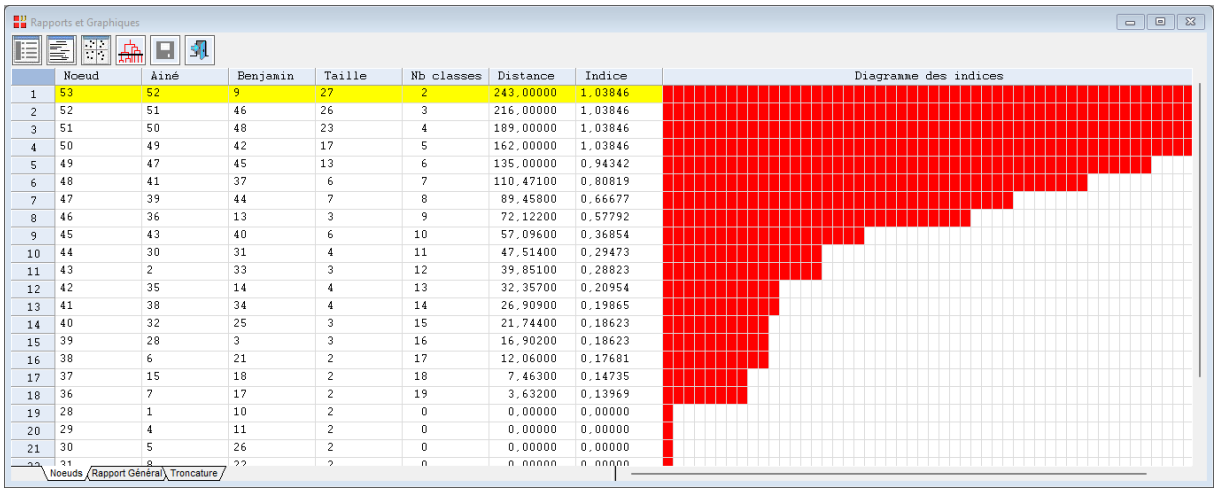
La boîte de dialogue montrée ci-dessous s'affiche :



Nous choisissons les variables de *hauteur* à *affect* comme variables définissant le tableau des données.

Nous choisissons la variable *varbase* comme variable contenant les libellés des colonnes et la variable *race* comme variable contenant les libellés des individus.

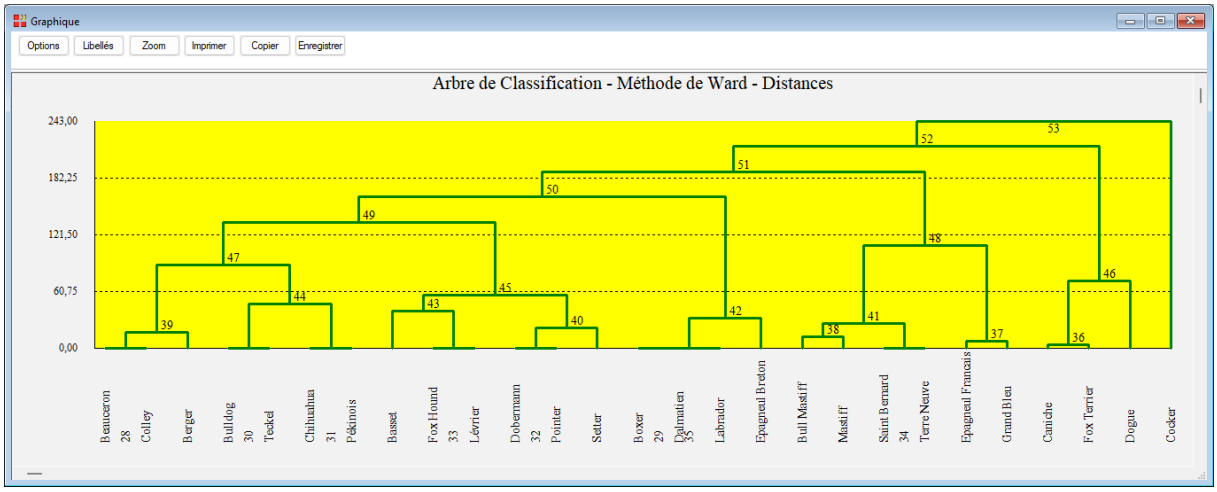
Nous choisissons de faire une analyse sur les lignes d'un tableau de données qualitatives.

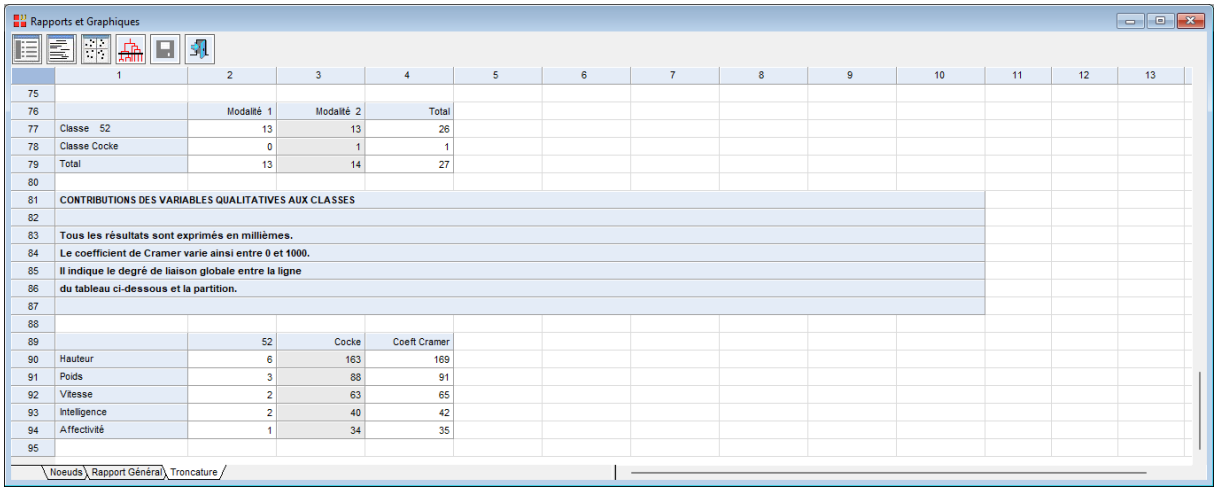


	Noeud	Aîné	Benjamin	Taille	Nb classes	Distance	Indice
1	53	52	9	27	2	243,00000	1,03846
2	52	51	46	26	3	216,00000	1,03846
3	51	50	48	23	4	189,00000	1,03846
4	50	49	42	17	5	162,00000	1,03846
5	49	47	45	13	6	135,00000	0,94342
6	48	41	37	6	7	110,47100	0,80819
7	47	39	44	7	8	89,45800	0,66677
8	46	36	13	3	9	72,12200	0,57792
9	45	43	40	6	10	57,09600	0,36854
10	44	30	31	4	11	47,51400	0,29473
11	43	2	33	3	12	39,85100	0,28823
12	42	35	14	4	13	32,35700	0,20954
13	41	38	34	4	14	26,90900	0,19865
14	40	32	25	3	15	21,74400	0,18623
15	39	28	3	3	16	16,90200	0,18623
16	38	6	21	2	17	12,06000	0,17681
17	37	15	18	2	18	7,46300	0,14735
18	36	7	17	2	19	3,63200	0,13969
19	28	1	10	2	0	0,00000	0,00000
20	29	4	11	2	0	0,00000	0,00000
21	30	5	26	2	0	0,00000	0,00000

L'indice de niveau est calculé de la façon suivante :

(Distance niveau i – Distance niveau i-1) / (n - 1) où n = nombre de lignes





	Modalité 1	Modalité 2	Total
Classe 52	13	13	26
Classe Cocke	0	1	1
Total	13	14	27

CONTRIBUTIONS DES VARIABLES QUALITATIVES AUX CLASSES

Tous les résultats sont exprimés en millièmes.
Le coefficient de Cramer varie ainsi entre 0 et 1000.
Il indique le degré de liaison globale entre la ligne
du tableau ci-dessous et la partition.

	52	Cocke	Coeft Cramer
Hauteur	6	163	169
Poids	3	88	91
Vitesse	2	63	65
Intelligence	2	40	42
Affectivité	1	34	35

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.

Variable	*Contenu*
agreg	Agrégations effectuées et restantes
nœuds	Numéros des nœuds
aines	Aînés associés aux nœuds
benjamins	Benjamins associés aux nœuds
tailles	Tailles des classes associées aux nœuds
distances	Valeurs des distances d'agrégation
indniv	Indices de niveau
nbclasses	Nombres de classes
indch	Indices de Calinski-Harabasz
liblig	Libellés des lignes
libcol	Libellés des colonnes
Nomxxx	Noms des classes (libellés des nœuds)
Codxxx	Codes des classes (codage des individus de 1 à n dans l'ordre de Nomxxx, avec n = nombre de classes formées)

Il y a autant de couples (Nomxxx, Codxxx) que de troncatures effectuées (xxx = numéro du nœud où la troncature a été effectuée).