UNIWIN VERSION 10.7.0

CLASSIFICATION PAR LA METHODE DES K-MOYENNES

Révision : 26/08/2026

Définition 1
Entrée des données 2
Données manquantes 3
Exemple 1 : Fichier VEHICULE 3
L'option Rapports 6
L'option Graphiques 9
Exemple 2 : Fichier IRIS3 12
Exemple 3 : Fichier IRIS - Classification mixte 15
Les variables internes créées par la procédure 19
Références 19

Définition

La classification par la méthode des K-moyennes (KM) permet de construire une partition d'un ensemble d'observations en classes telles que les observations appartenant à une même classe sont proches alors que les observations appartenant à des classes différentes sont éloignées. C'est une méthode de classification non hiérarchique complémentaire à la méthode de classification ascendante hiérarchique (CAH).

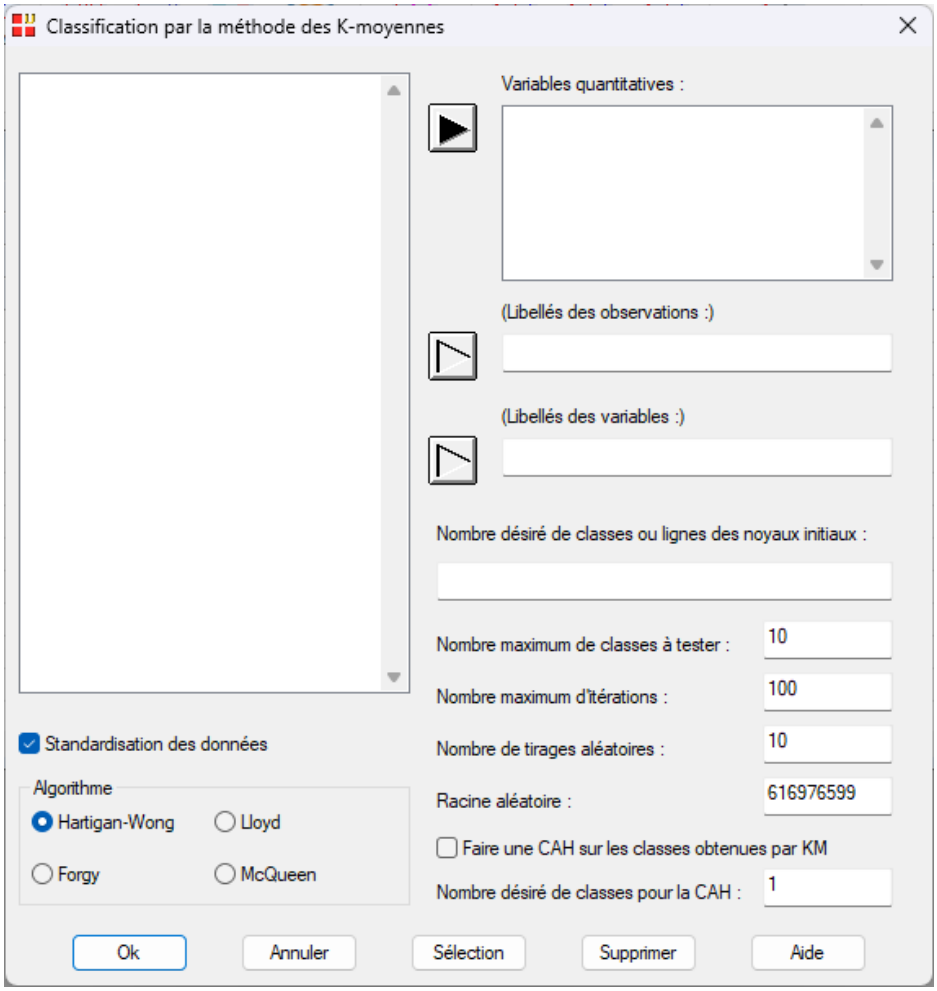
La procédure affiche un rapport indiquant notamment les centroïdes des classes formées, la classification des observations, des statistiques descriptives pour les classes formées et les contributions des variables aux classes. Si une classification mixte a été mise en œuvre, les résultats de la CAH sont également fournis.

Les graphiques des inerties intra-classes, des inerties expliquées, des indices de Calinski-Harabasz, des coefficients moyens et individuels de silhouette et des nuages de points des classes formées sont proposés. Si une classification mixte a été mise en œuvre, le diagramme des indices de la classification et l'affichage de l'arbre sont proposés.

Cette procédure est basée sur les packages R 'stats' et 'cluster'.

Entrée des données

Cliquons sur l'icône KM dans le ruban Décrire et choisissons K-moyennes pour afficher la boîte de dialogue montrée ci-dessous :



Cette boîte de dialogue permet de choisir les variables quantitatives à utiliser pour la classification, la variable contenant les libellés des observations et la variable contenant les libellés des variables quantitatives utilisées pour la classification.

Par défaut les données sont standardisées.

Le nombre désiré de classes peut être précisé de deux façons, soit en entrant ce nombre, soit en entrant les numéros des lignes du fichier des données définissant les noyaux initiaux de ces classes.

Le champ 'Nombre maximum de classes à tester' permet de calculer diverses statistiques pouvant aider à valider le nombre adéquat de classes à former.

L'algorithme à utiliser peut être sélectionné. La méthode de Hartigan-Wong (1979) est la méthode par défaut. Elle est généralement plus performante que les méthodes de Lloyd (1957), Forgy (1965) et McQueen (1967).

Le nombre maximum d'itérations de l'algorithme peut être précisé.

Le nombre de tirages aléatoires des noyaux de départ et la racine aléatoire associée peuvent être indiqués dans le cas où le nombre de classes à former est précisé. Ces deux options ne sont pas utilisées si les noyaux initiaux des classes à former sont entrés.

Enfin, il est possible de compléter l'analyse par une classification ascendante hiérarchique (CAH) sur les classes obtenues par les K-moyennes.

Données manquantes

Les données manquantes ne sont pas autorisées par cette procédure.

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.*

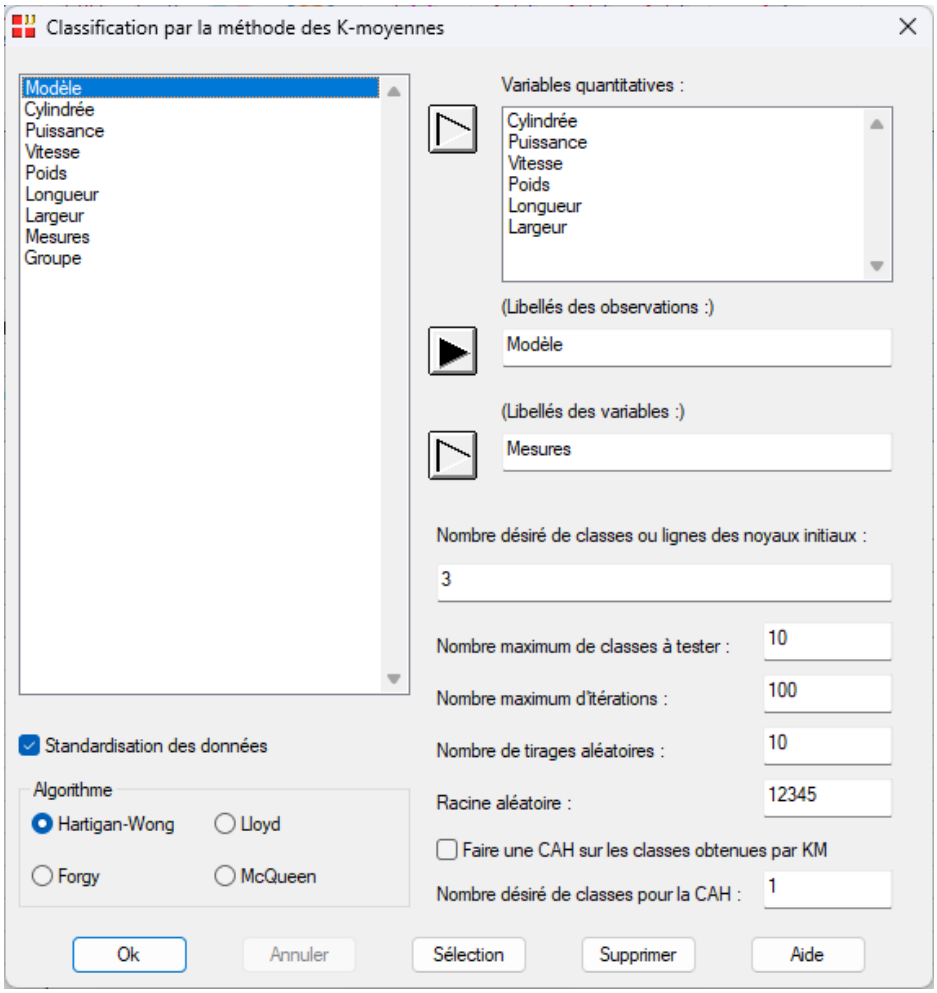
Les 24 modèles d'automobiles sont :

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

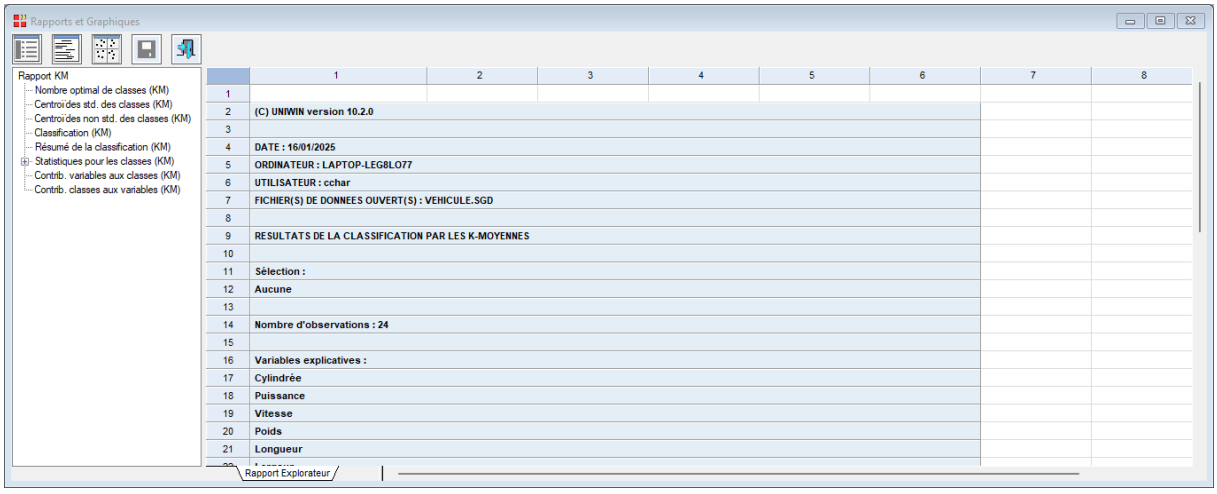
Cliquons sur l'icône KM dans le ruban Décrire et choisissons K-moyennes. La boîte de dialogue montrée ci-après s'affiche.

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 observations.

Nous demandons de former 3 classes et choisissons de faire une classification sur les données standardisées.

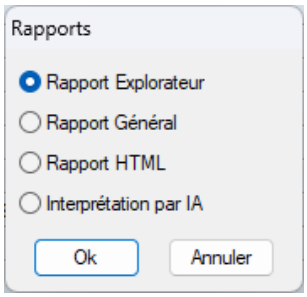


Cliquons sur Ok. UNIWIN débute le calcul de la classification. Après quelques instants, l'écran suivant s'affiche :

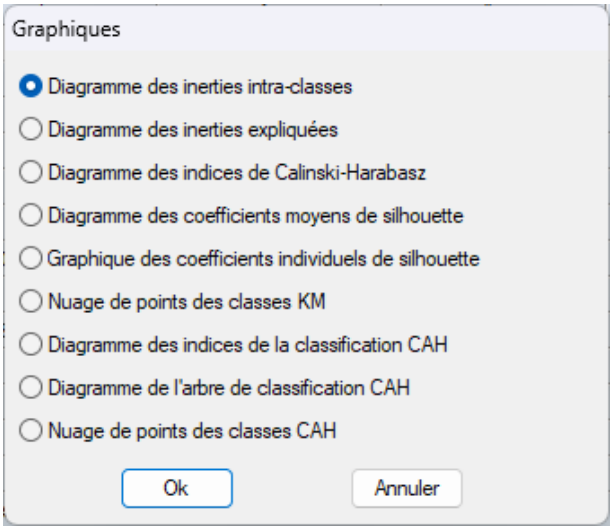


La barre d'outils 'Rapports et Graphiques' permet par l'icône 'Données' de rappeler la boîte de dialogue d'entrée des données.

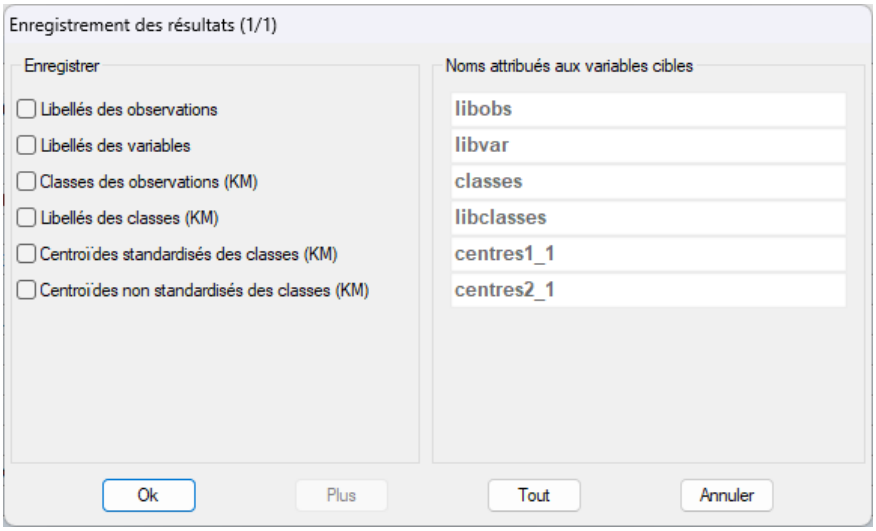
L'icône 'Rapports' affiche la boîte de dialogue des options pour les rapports :



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



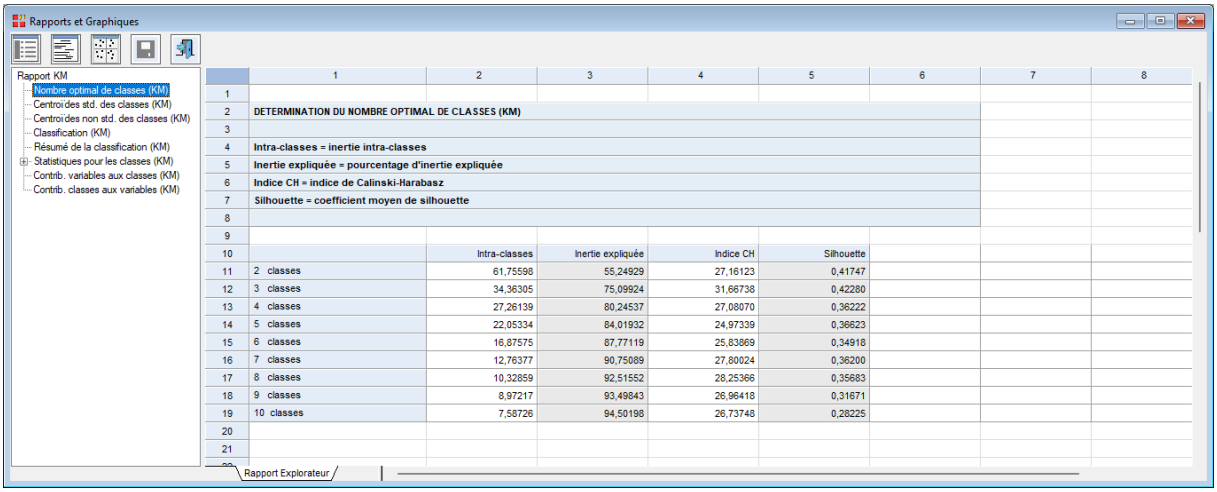
L'icône 'Enregistrer' permet de sélectionner les résultats de l'analyse à enregistrer dans un fichier.



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

L'option Rapports

Le premier tableau affiche pour les nombres de classes variant de 2 au nombre maximum de classes indiqué, les inerties expliquées, les indices de Calinski-Harabasz, les inerties intra-classes et les coefficients moyens de silhouette. Ces quatre indices sont également représentés de façon graphique. Ils aident à déterminer le nombre adéquat de classes à former.



DETERMINATION DU NOMBRE OPTIMAL DE CLASSES (KM)

Intra-classes = inertie intra-classes
Inertie expliquée = pourcentage d'inertie expliquée
Indice CH = indice de Calinski-Harabasz
Silhouette = coefficient moyen de silhouette

	Intra-classes	Inertie expliquée	Indice CH	Silhouette
2 classes	61,75598	55,24929	27,16123	0,41747
3 classes	34,36305	75,09924	31,66738	0,42280
4 classes	27,26139	80,24537	27,08070	0,36222
5 classes	22,05334	84,01932	24,97339	0,36623
6 classes	16,87575	87,77119	25,83869	0,34918
7 classes	12,76377	90,75089	27,80024	0,36200
8 classes	10,32859	92,51552	28,25366	0,35683
9 classes	8,97217	93,49843	26,96418	0,31671
10 classes	7,58726	94,50198	26,73748	0,28225

L'indice CH varie entre 0 et l'infini. Une valeur élevée de l'indice CH signifie que les classes formées sont denses et bien séparées. Il faut choisir la solution qui donne un pic ou au moins un coude abrupt dans le graphique des indices CH.

Les coefficients de silhouette varient entre -1 et +1. Structure forte de la classe si [0,71 ; 1,00], raisonnable si [0,51 ; 0,70], faible si [0,25 ; 0,50] et pas de structure si [-1,00 ; 0,25].

Les deuxième et troisième tableaux affichent les centroïdes standardisés et non standardisés des classes formées.

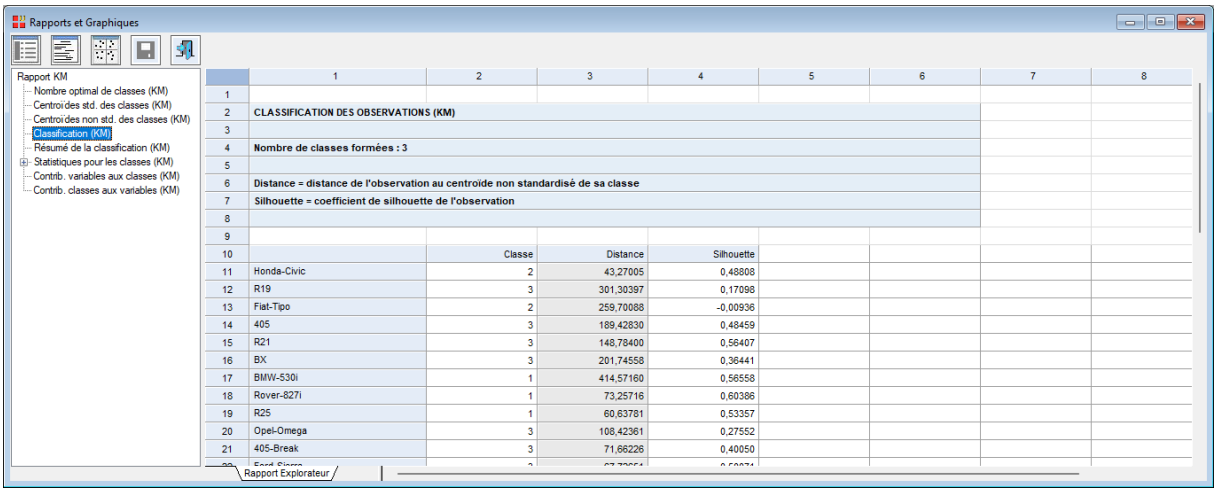
CENTROIDES STANDARDISES DES CLASSES

	Cylindrée	Puissance	Vitesse	Poids	Longueur	Largeur
Classe 1	1,32386	1,48685	1,30542	1,03130	0,88493	0,56617
Classe 2	-1,04445	-0,80681	-0,63288	-1,15976	-1,27196	-1,03980
Classe 3	0,04125	-0,24666	-0,27695	0,30903	0,48661	0,49213

CENTROIDES NON STANDARDISES DES CLASSES

	Cylindrée	Puissance	Vitesse	Poids	Longueur	Largeur
Classe 1	2605,00	171,33333	216,000	1348,33333	458,16667	173,16667
Classe 2	1354,75	82,37500	167,125	843,75000	369,00000	160,87500
Classe 3	1927,90	104,10000	176,100	1182,00000	441,70000	172,60000

Le quatrième tableau affiche les résultats de la classification.



Pour chaque observation, il indique sa classe d'affectation, sa distance au centroïde non standardisé de cette classe et son coefficient de silhouette.

Le coefficient de silhouette est la différence entre la distance moyenne avec les observations de sa classe et la distance moyenne avec les observations des autres classes. Il varie entre -1 et +1.

Si cette différence est négative, l'observation est en moyenne plus proche de la classe voisine que de la sienne et donc elle est donc mal classée.

À l'inverse, si cette différence est positive, l'observation est en moyenne plus proche de sa classe que de la classe voisine et donc elle est donc bien classée.

Le cinquième tableau affiche un résumé de la classification : nombre d'observations dans chaque classe, pourcentages associés ainsi que les coefficients de silhouette.

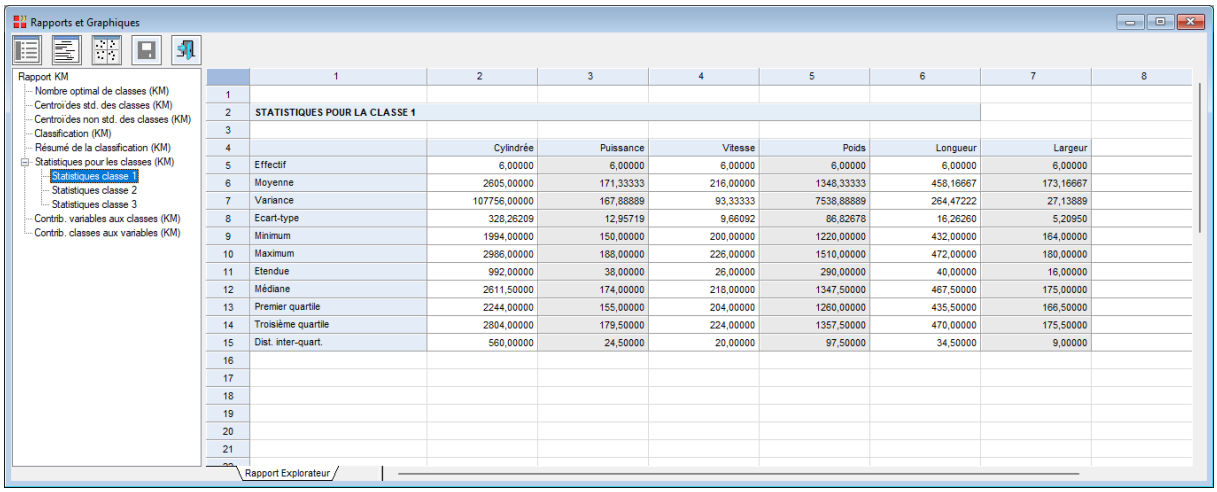
RESUME DE LA CLASSIFICATION (KM)

Silhouette : coefficient moyen de silhouette pour la classe

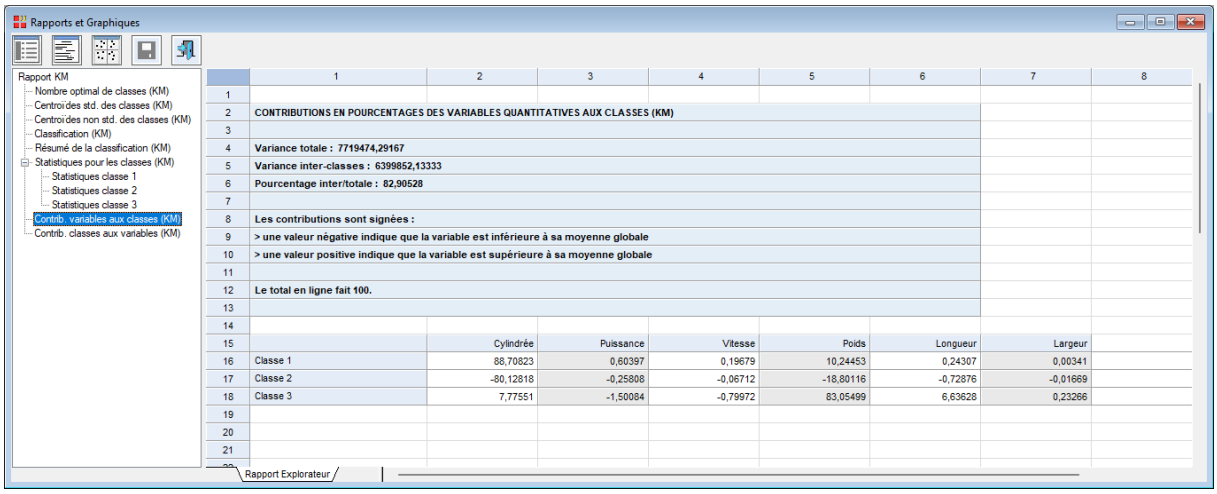
	Nombre d'observations	Pourcentage	Silhouette
Classe 1	6	25,00000	0,42423
Classe 2	8	33,33333	0,45865
Classe 3	10	41,66667	0,39327
Total	24	100,00000	0,42280

Le tableau suivant affiche pour chaque classe un ensemble de statistiques descriptives pour chacune des variables utilisées pour la classification.

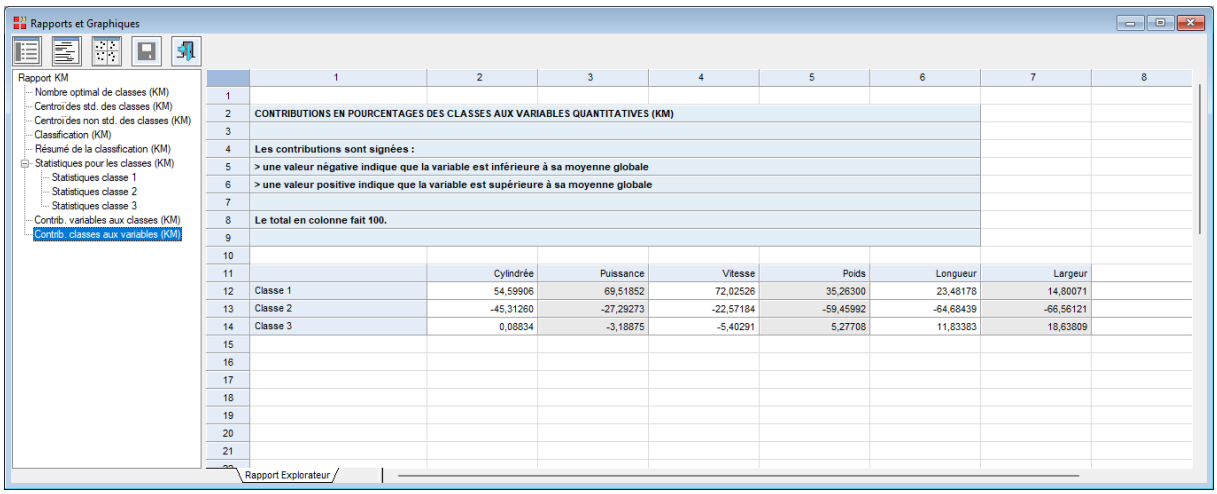
Les deux tableaux suivants affichent les contributions signées en pourcentages des variables aux classes et les contributions signées en pourcentages des classes aux variables.



	1	2	3	4	5	6	7	8
1								
2	**STATISTIQUES POUR LA CLASSE 1**							
3								
4		Cylindrée	Puissance	Vitesse	Poids	Longueur	Largeur	
5	Effectif	6,00000	6,00000	6,00000	6,00000	6,00000	6,00000	
6	Moyenne	2605,00000	171,33333	216,00000	1348,33333	458,16667	173,16667	
7	Variance	107756,00000	167,88889	93,33333	7538,88889	264,47222	27,13889	
8	Ecart-type	328,26209	12,95719	9,66092	86,82678	16,26260	5,20950	
9	Minimum	1994,00000	150,00000	200,00000	1220,00000	432,00000	164,00000	
10	Maximum	2986,00000	188,00000	226,00000	1510,00000	472,00000	180,00000	
11	Etendue	992,00000	38,00000	26,00000	290,00000	40,00000	16,00000	
12	Médiane	2611,50000	174,00000	218,00000	1347,50000	467,50000	175,00000	
13	Premier quartile	2244,00000	155,00000	204,00000	1260,00000	435,50000	166,50000	
14	Troisième quartile	2804,00000	179,50000	224,00000	1357,50000	470,00000	175,50000	
15	Dist. inter-quart.	560,00000	24,50000	20,00000	97,50000	34,50000	9,00000	
16								
17								
18								
19								
20								
21								



	1	2	3	4	5	6	7	8
1								
2	**CONTRIBUTIONS EN POURCENTAGES DES VARIABLES QUANTITATIVES AUX CLASSES (KM)**							
3								
4	**Variance totale : 7719474,29167**							
5	**Variance inter-classes : 6399852,13333**							
6	**Pourcentage inter/totale : 82,90528**							
7								
8	**Les contributions sont signées :**							
9	**> une valeur négative indique que la variable est inférieure à sa moyenne globale**							
10	**> une valeur positive indique que la variable est supérieure à sa moyenne globale**							
11								
12	**Le total en ligne fait 100.**							
13								
14								
15		Cylindrée	Puissance	Vitesse	Poids	Longueur	Largeur	
16	Classe 1	88,70823	0,60397	0,19679	10,24453	0,24307	0,00341	
17	Classe 2	-80,12818	-0,25808	-0,06712	-18,80116	-0,72876	-0,01669	
18	Classe 3	7,77551	-1,50084	-0,79972	83,05499	6,63628	0,23266	
19								
20								
21								

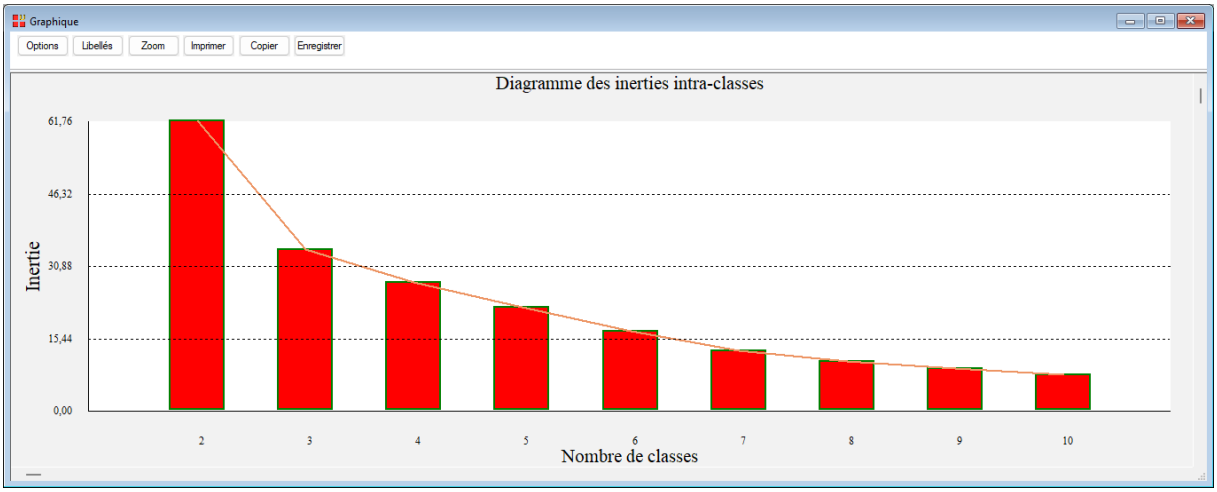


	1	2	3	4	5	6	7	8
1								
2	**CONTRIBUTIONS EN POURCENTAGES DES CLASSES AUX VARIABLES QUANTITATIVES (KM)**							
3								
4	**Les contributions sont signées :**							
5	**> une valeur négative indique que la variable est inférieure à sa moyenne globale**							
6	**> une valeur positive indique que la variable est supérieure à sa moyenne globale**							
7								
8	**Le total en colonne fait 100.**							
9								
10								
11		Cylindrée	Puissance	Vitesse	Poids	Longueur	Largeur	
12	Classe 1	54,59906	69,51852	72,02526	35,26300	23,48178	14,80071	
13	Classe 2	-45,31260	-27,29273	-22,57184	-59,45992	-64,68439	-66,56121	
14	Classe 3	0,08834	-3,18875	-5,40291	5,27708	11,83383	18,63809	
15								
16								
17								
18								
19								
20								
21								

L'option Graphiques

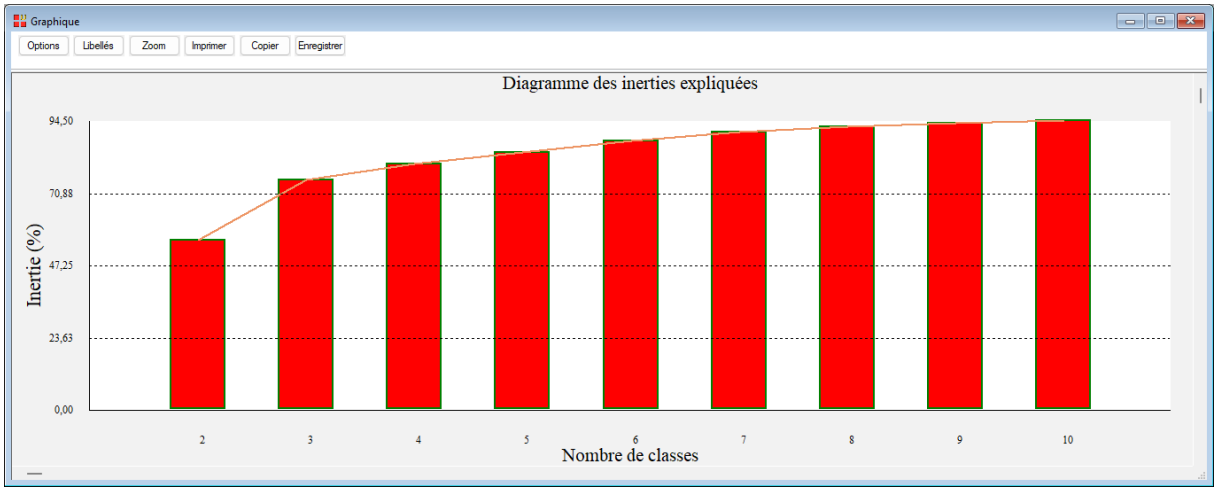
- Diagramme des inerties intra-classes

Le diagramme des inerties intra-classes permet de visualiser l'évolution de l'inertie intra-classes en fonction du nombre de classes formées.



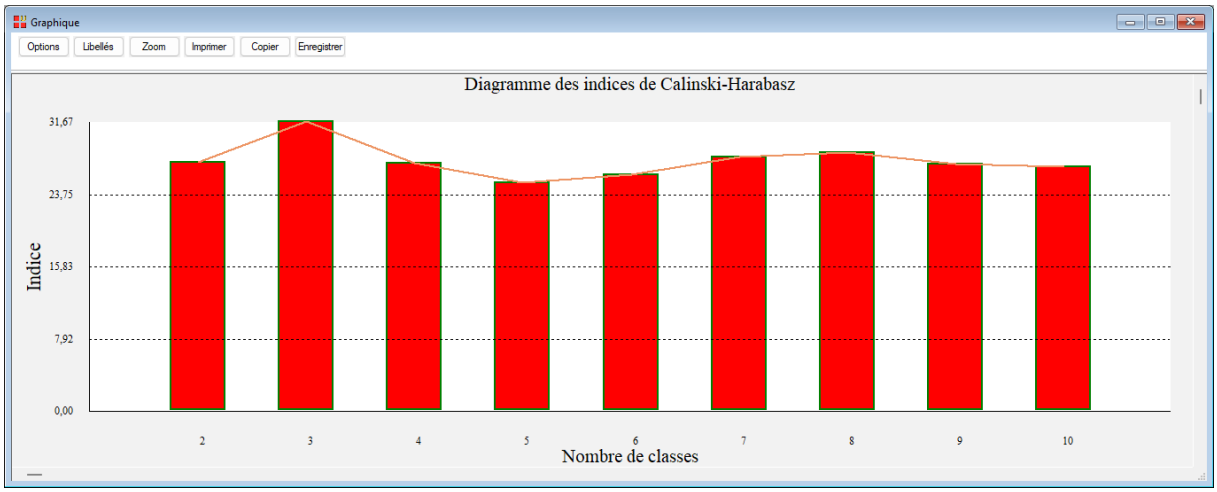
- Diagramme des inerties expliquées

Le diagramme des inerties expliquées permet de visualiser l'évolution de l'inertie expliquée en fonction du nombre de classes formées.



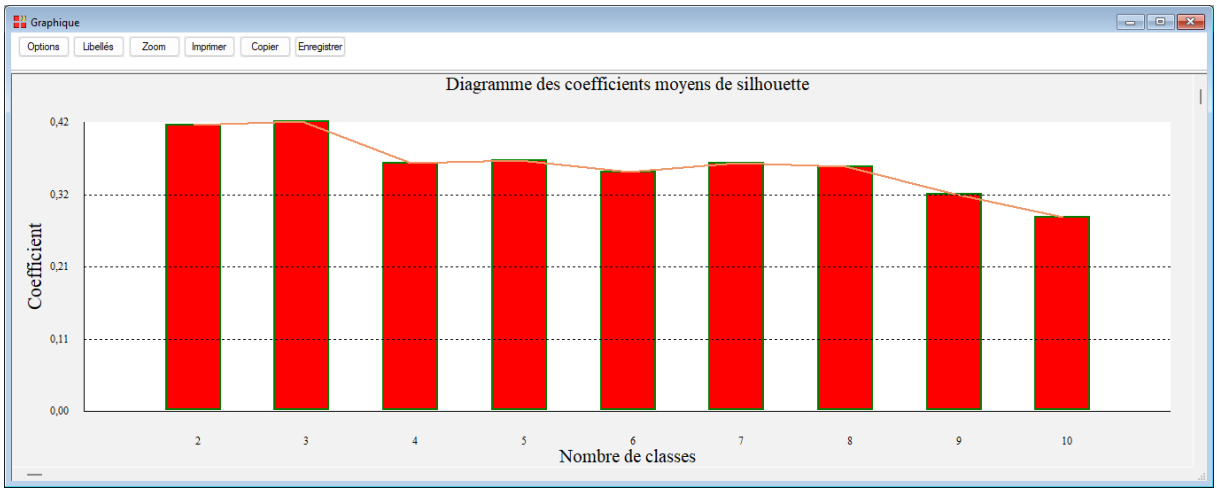
- Diagramme des indices de Calinski-Harabasz

Le diagramme des indices de Calinski-Harabasz (CH) permet de visualiser l'évolution de cet indice en fonction du nombre de classes formées. L'indice CH varie entre 0 et l'infini. Une valeur élevée de l'indice CH signifie que les classes formées sont denses et bien séparées. Il faut choisir la solution qui donne un pic ou au moins un coude abrupt dans le graphique des indices CH.



- Diagramme des coefficients moyens de silhouette

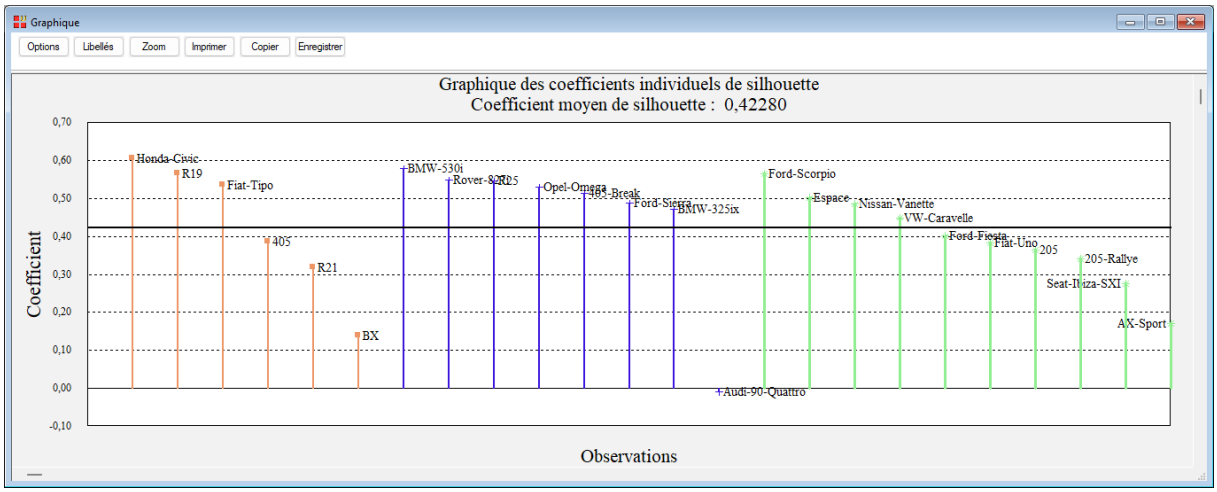
Le diagramme des coefficients moyens de silhouette permet de visualiser l'évolution de ces coefficients en fonction du nombre de classes formées.



- Graphique des coefficients individuels de silhouette

Le graphique des coefficients individuels de silhouette permet de visualiser ces coefficients pour chacune des observations. Pour chaque observation, le coefficient de silhouette est la différence entre la distance moyenne avec les observations de sa classe et la distance moyenne avec les observations des autres classes. Il varie entre -1 et +1. Si négatif, l'observation est en moyenne plus proche de la classe voisine que de la sienne et donc elle est donc mal classée. Si positif, l'observation est en moyenne plus proche de sa classe que de la classe voisine et donc elle est donc bien classée.

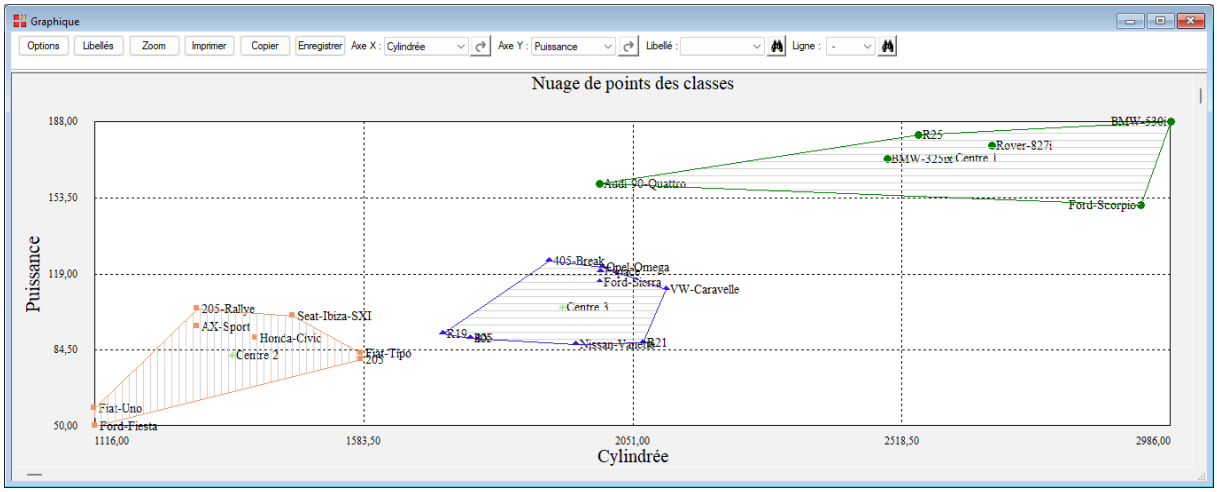
La ligne horizontale indique la position du coefficient moyen de silhouette. Si la plupart des observations d'une classe ont des coefficients inférieurs à ce coefficient moyen, cela indique probablement que la partition obtenue n'est pas bonne.

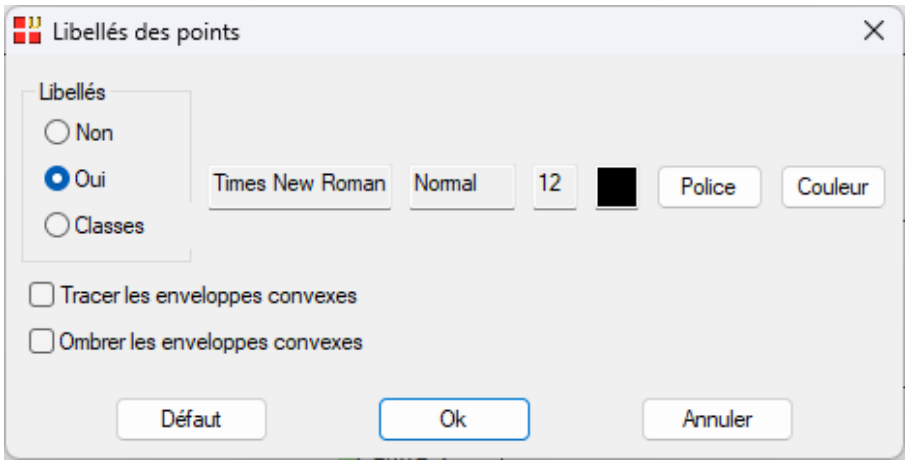


- Graphique des nuages de points des classes

Le graphique des nuages de points des classes permet de visualiser les classes formées par rapport à deux variables quantitatives sélectionnées.

Le bouton ‘Libellés’ permet de préciser les libellés affichés et si les enveloppes convexes des classes sont tracées ou non.



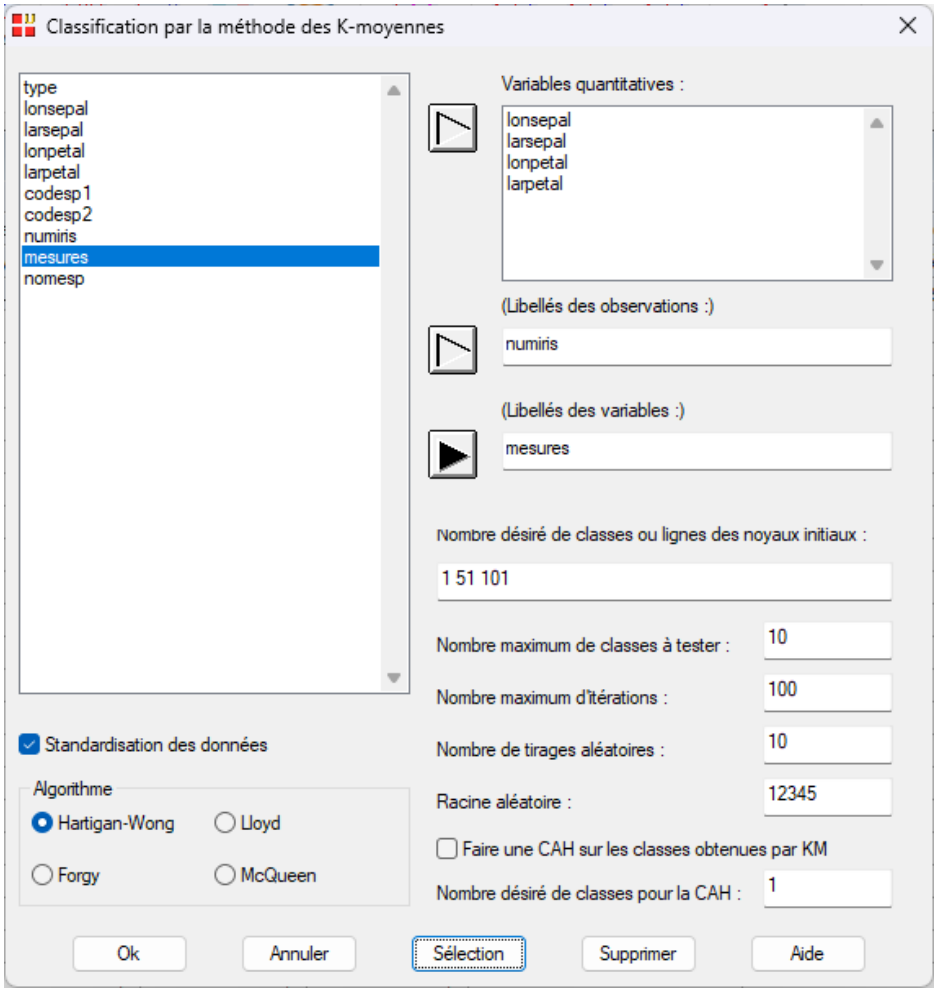


Exemple 2 : Fichier IRIS3

Nous utiliserons le fichier IRIS3 pour illustrer ce deuxième exemple.

Ce fichier contient pour 150 iris les mesures des quatre caractéristiques suivantes exprimées en millimètres : longueur du sépale, largeur du sépale, longueur du pétale et largeur du pétale

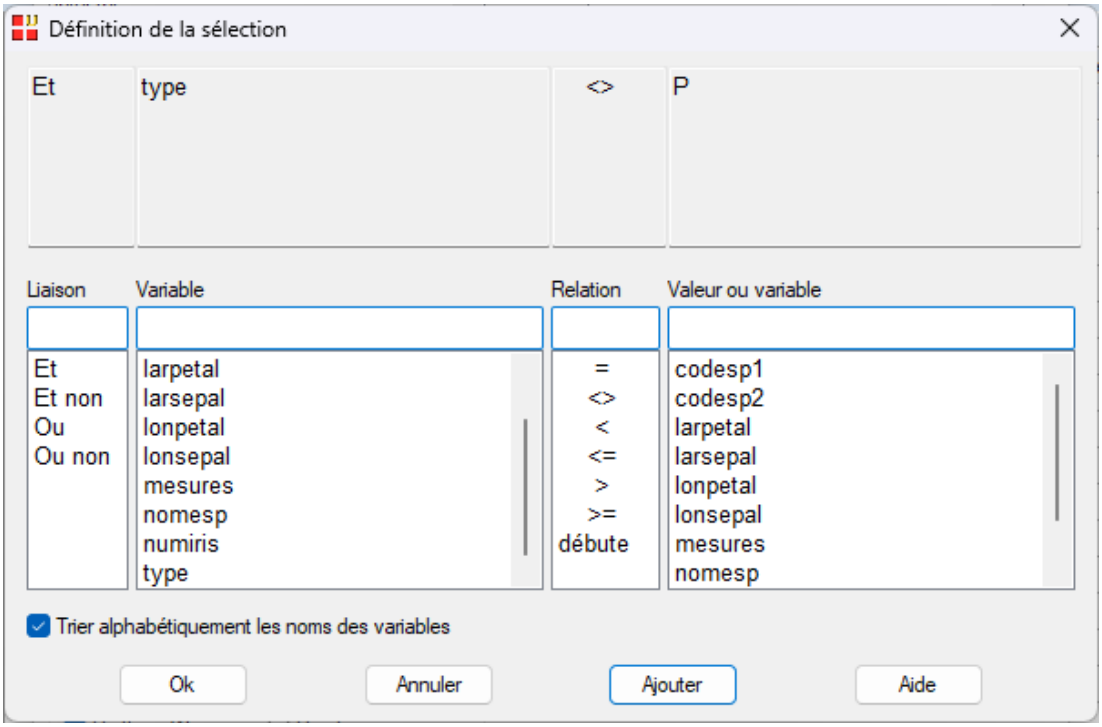
Renseignons la boîte de dialogue comme montré ci-dessous en précisant les noyaux initiaux à utiliser : les lignes 1, 51 et 101 du fichier des données.



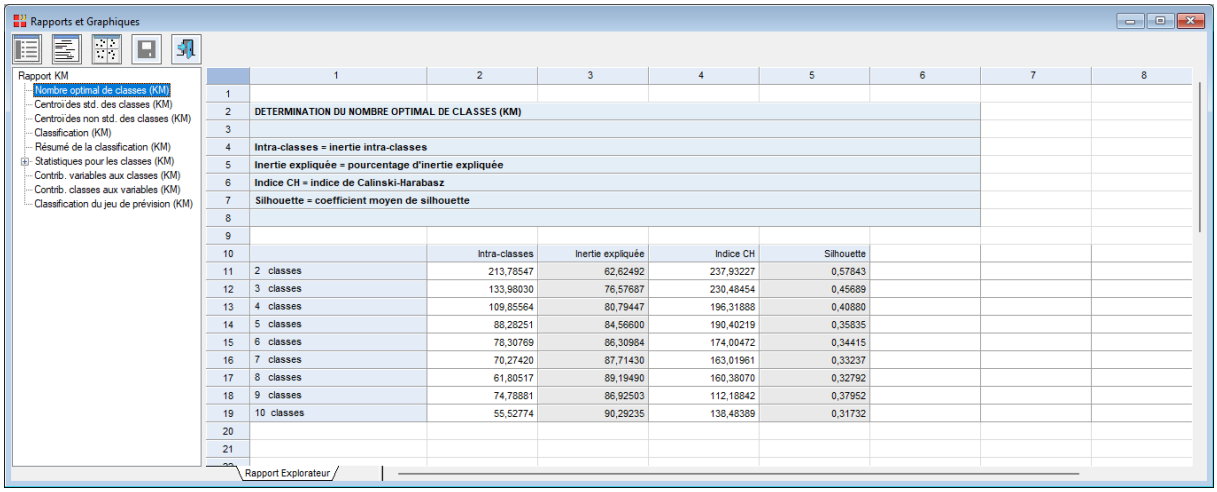
Cliquons sur le bouton ‘Sélection’ pour sélectionner les observations à utiliser comme jeu d’apprentissage.

Un message nous indique que 144 observations seront utilisées.

Cliquons sur Ok. Un message nous indique que les lignes non sélectionnées seront utilisées comme jeu de prévision.



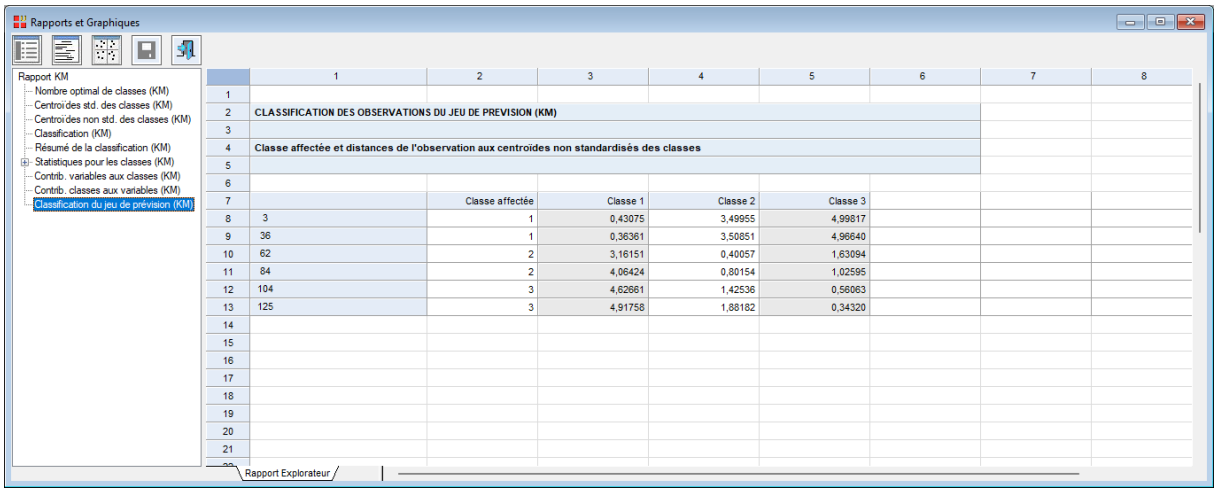
Visualisons quelques résultats obtenus.

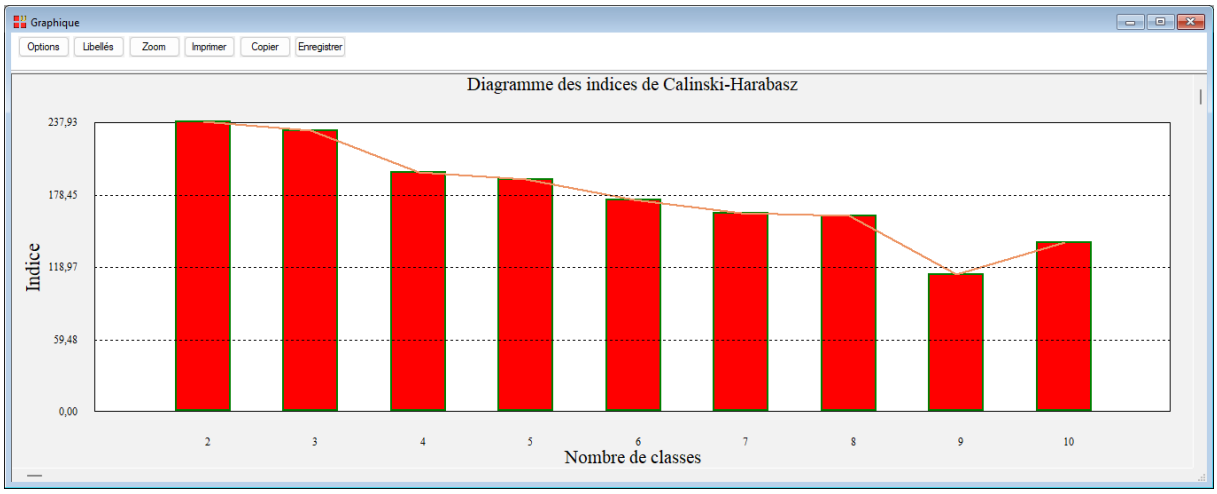


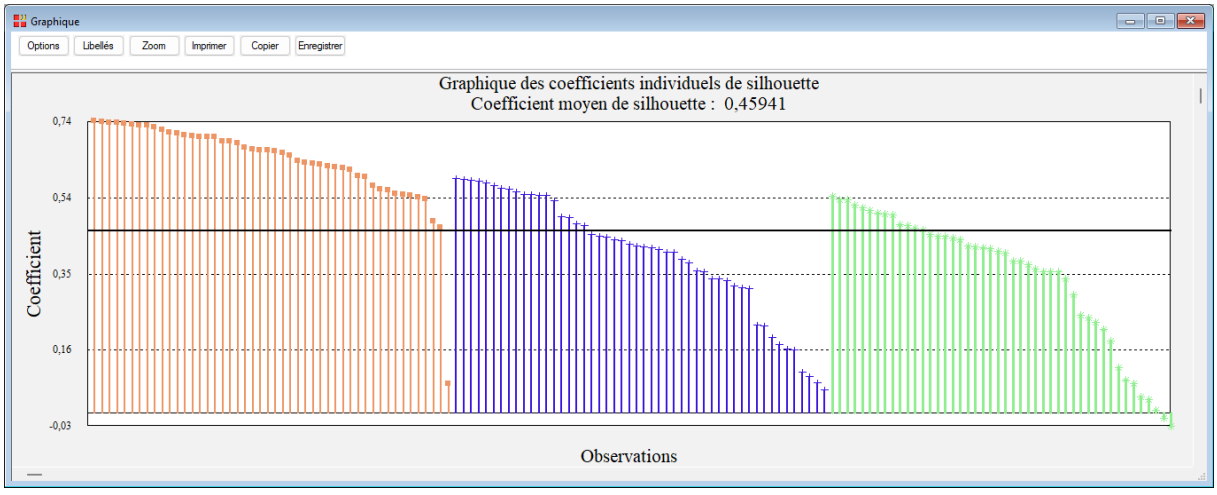
RESUME DE LA CLASSIFICATION (KM)

Silhouette : coefficient moyen de silhouette pour la classe

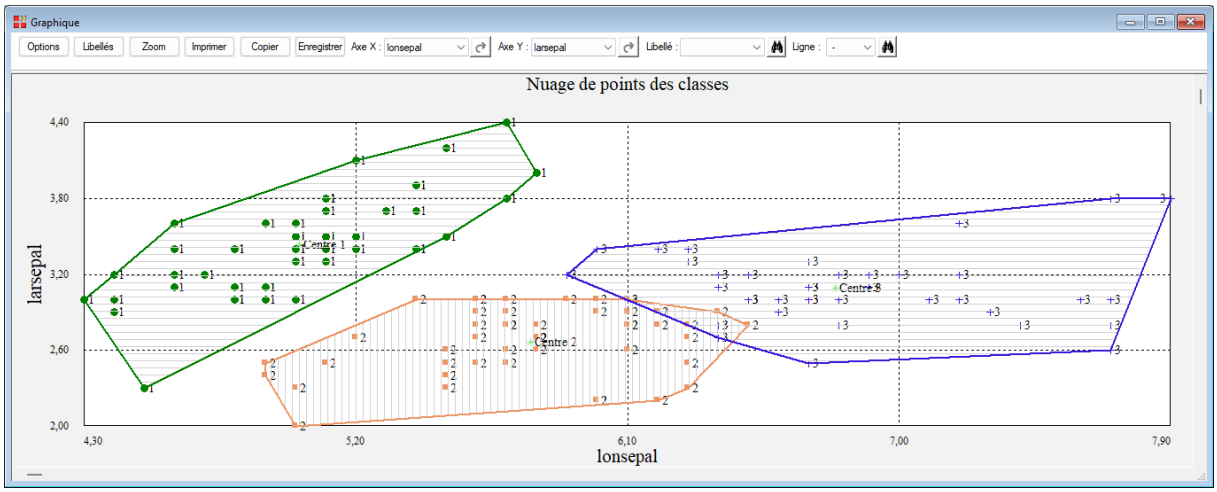
	Nombre d'observations	Pourcentage	Silhouette
Classe 1	48	33,33333	0,63389
Classe 2	50	34,72222	0,39738
Classe 3	46	31,94444	0,34476
Total	144	100,00000	0,45941







La ligne horizontale indique la position du coefficient moyen de silhouette. Si la plupart des observations d'une classe ont des coefficients inférieurs à ce coefficient moyen, cela indique probablement que la partition obtenue n'est pas bonne.



Exemple 3 : Fichier IRIS - Classification mixte

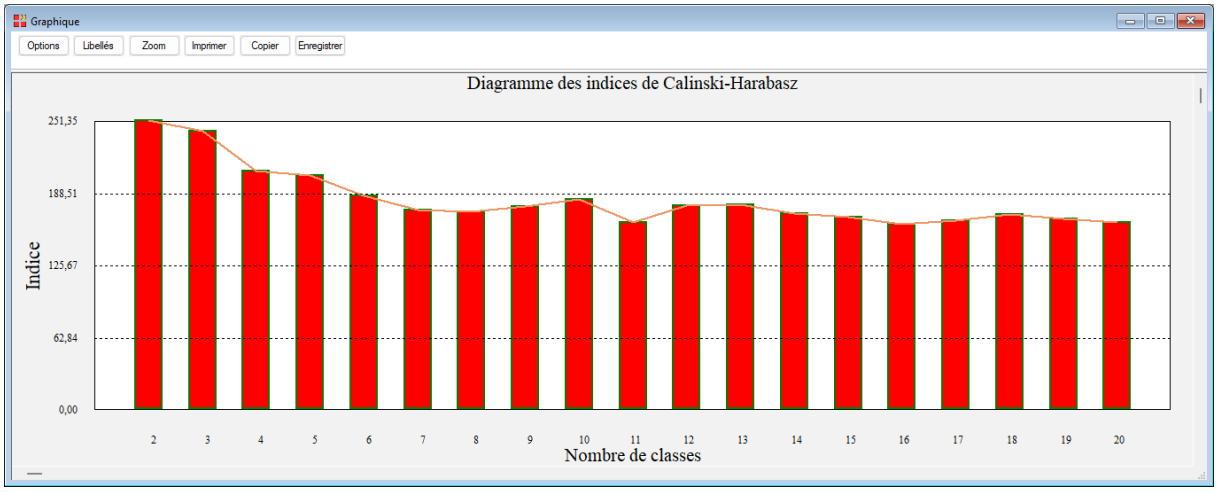
Si le nombre d'observations à classer est important, la classification mixte est une démarche intéressante. Elle se déroule en trois étapes :

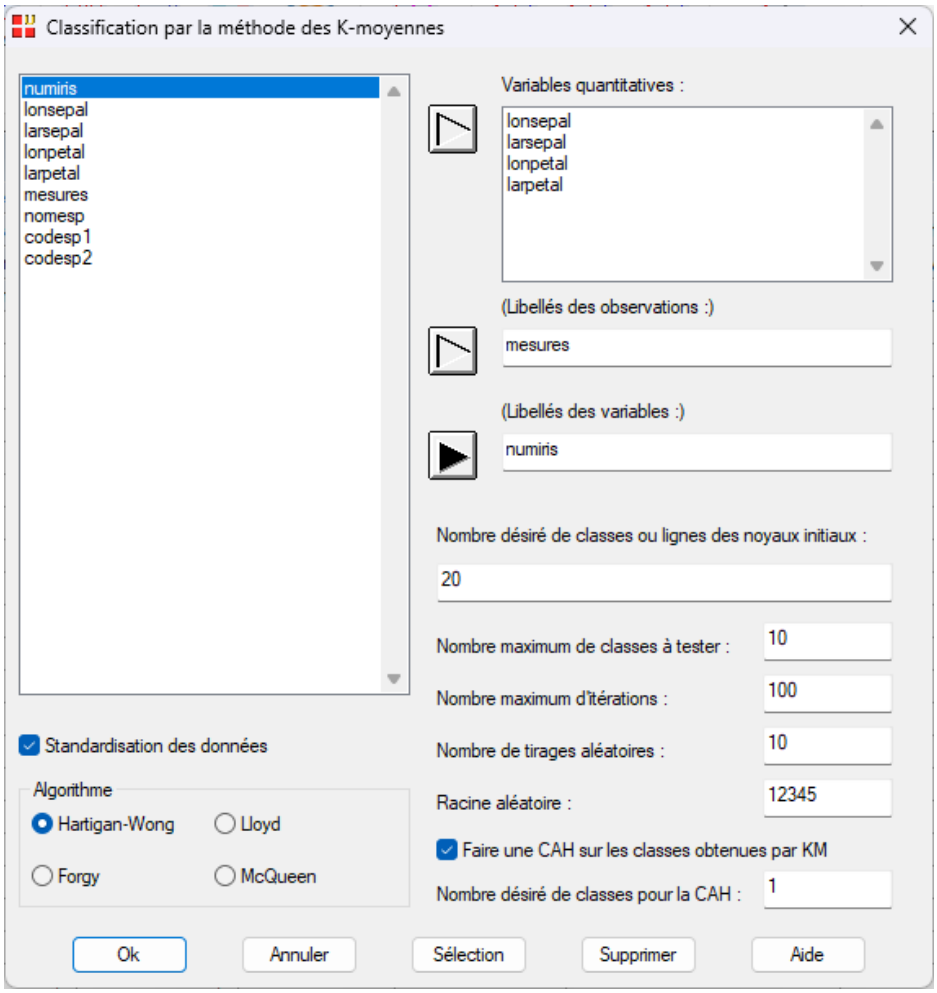
1. Partitionnement en *q* classes par la méthode K-moyennes
2. Classification ascendante hiérarchique (CAH) sur les *q* classes issues des K-moyennes avec pondération par les effectifs des classes
3. Partition finale obtenue par troncature de l'arbre CAH

A titre d'exemple illustratif, mettons cette démarche en œuvre en utilisant le fichier IRIS et en demandant un nombre de classes égal à 20 (le nombre maximum de classes à tester est alors automatiquement défini à 20).

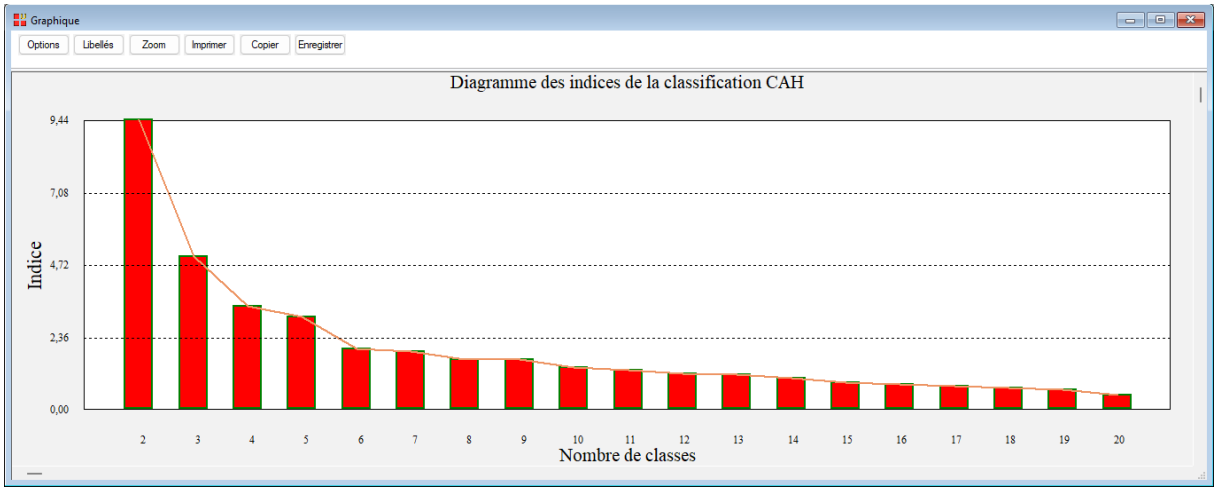
Cochons la case 'Faire une classification CAH sur les classes obtenues par KM'' et demandons 1 classe pour la CAH.

Visualisons le diagramme des indices de Calinski-Harabasz (CH).



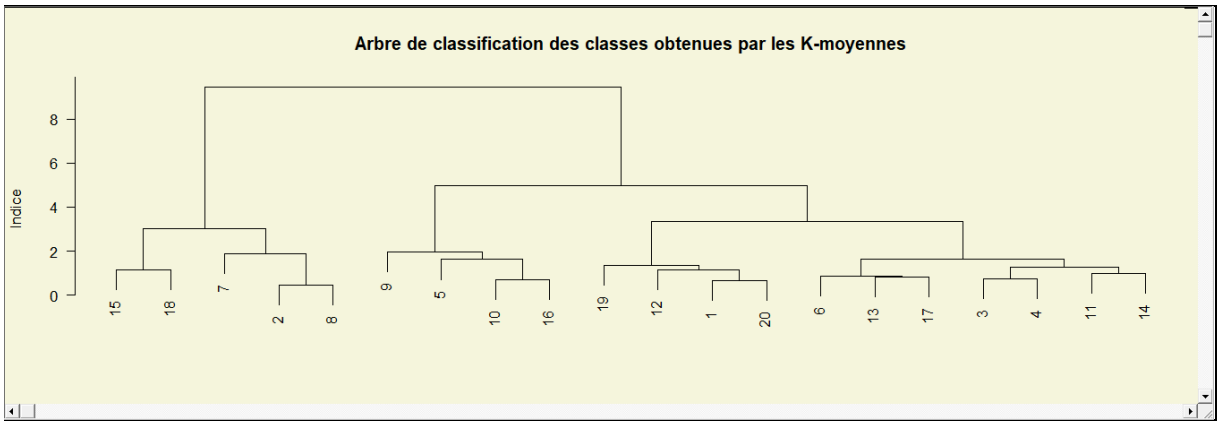


Affichons le diagramme des indices de la classification CAH.

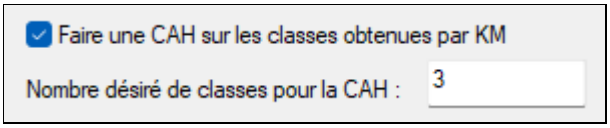


Ce graphique nous indique que 3 classes est un choix possible et confirme ainsi le résultat obtenu avec le diagramme des indices CH.

Visualisons l'arbre de la classification.

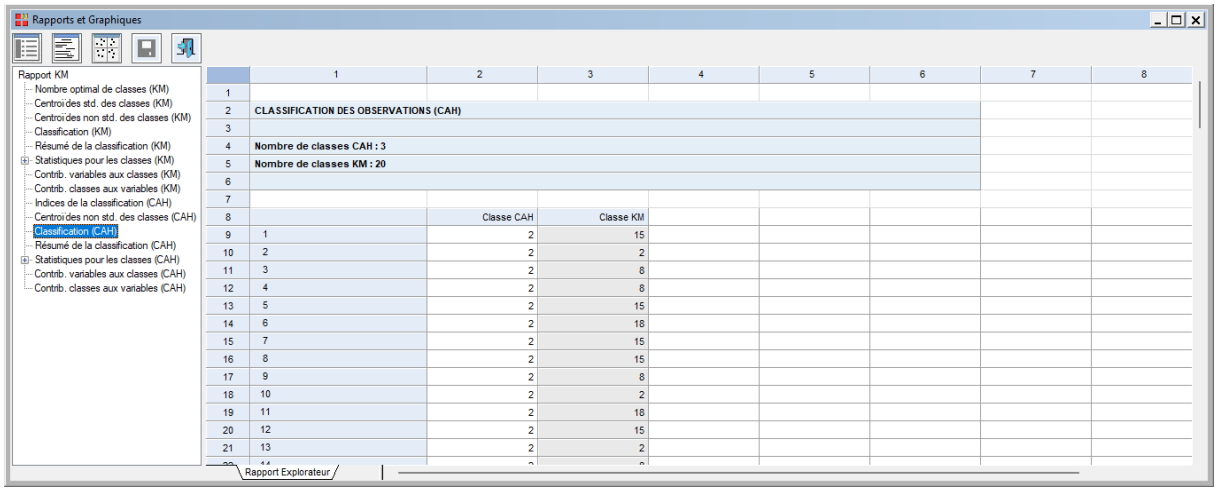


Revenons à la boîte de dialogue d'entrée des données en cliquant sur et indiquons un nombre de classes égal à 3 pour la classification CAH.



Exécutons à nouveau l'analyse.

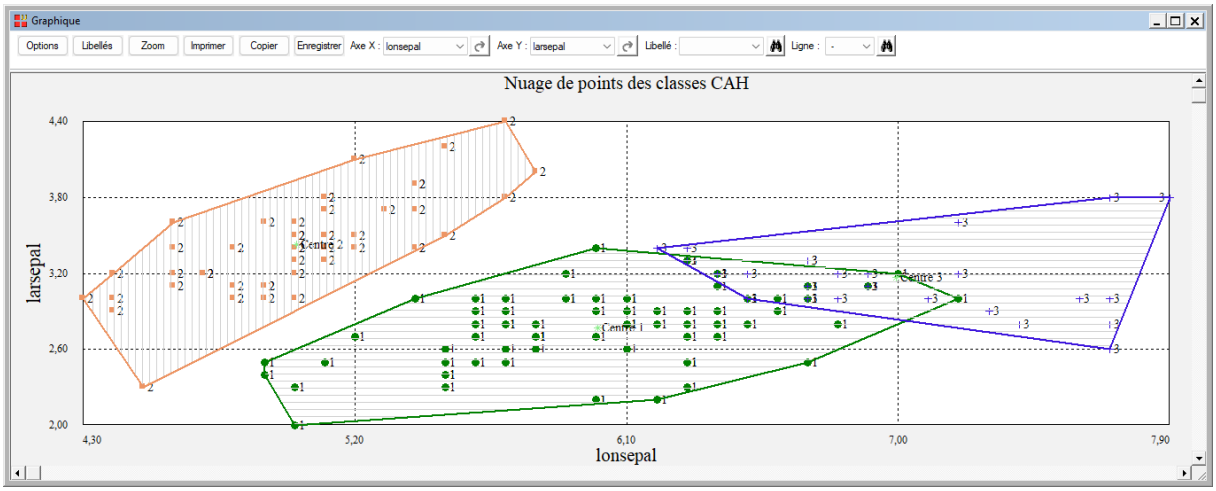
Le rapport nous indique comment chaque observation a été affectée à une classe CAH et la classe initiale KM.



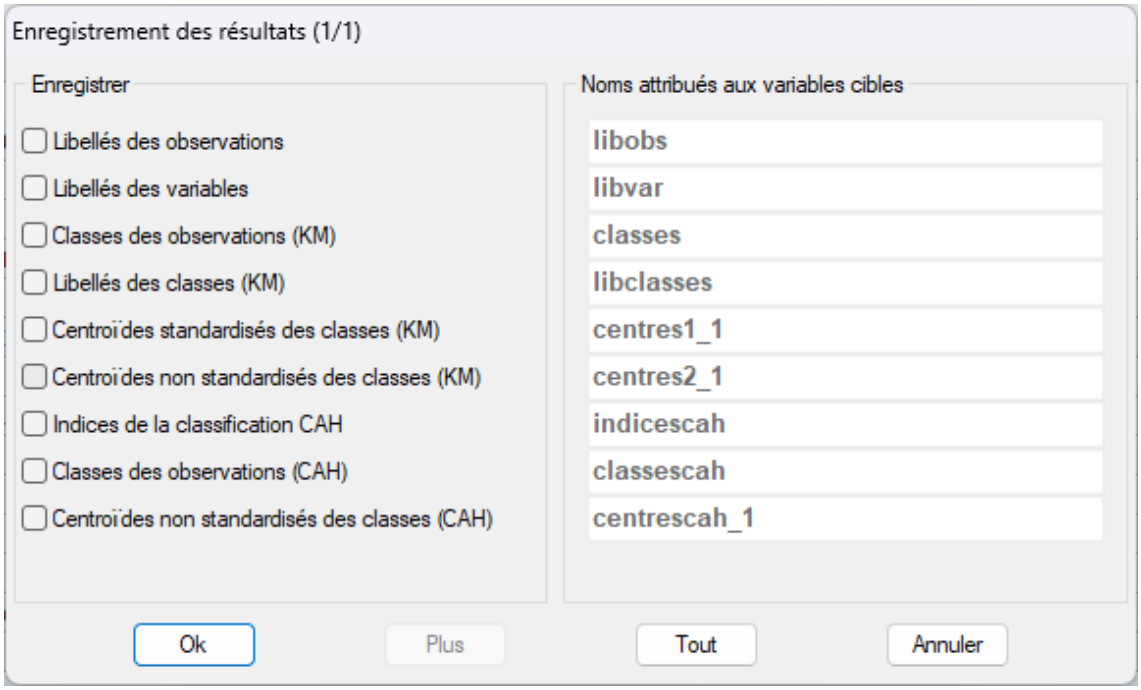
RESUME DE LA CLASSIFICATION (CAH)		
	Nombre d'observations	Pourcentage
Classe 1	74	49,33333
Classe 2	50	33,33333
Classe 3	26	17,33333

CENTROIDES NON STANDARDISES DES CLASSES (CAH)				
	lonsepal	larsepal	lonpetal	larpetal
Classe 1	6,00541	2,76892	4,56486	1,49324
Classe 2	5,00600	3,42800	1,46200	0,24600
Classe 3	6,99231	3,16538	5,87692	2,19615

Le graphique du nuage des points affiche les classes de la CAH.



Il est possible d'enregistrer ces résultats :



Les variables internes créées par la procédure

Voici la liste des variables internes créées par la procédure.

Variable	*Contenu*
libobs	Libellés des observations du jeu d'apprentissage
libvar	Libellés des variables
classes	Classes affectées aux observations du jeu d'apprentissage (KM)
libclasses	Libellés des classes (KM)
centres1	Centroïdes standardisés des classes (KM)
centres2	Centroïdes non standardisés des classes (KM)
libobsprev	Libellés des observations du jeu de prévision (KM)
classesprev	Classes affectées aux observations du jeu de prévision (KM)
indicescah	Indices de la classification CAH (classification mixte)
classescah	Classes des observations CAH (classification mixte)
centrescah	Centroïdes non standardisés des classes CAH (classification mixte)

Références

Documentation du package R 'stats' (2026)

https://rdrr.io/r/stats/stats-package.html

Documentation du package R 'cluster' (2026)

https://cran.r-project.org/web/packages/cluster/cluster.pdf