UNIWIN VERSION 9.7.0

METHODE SIMCA

Révision : 02/09/2023

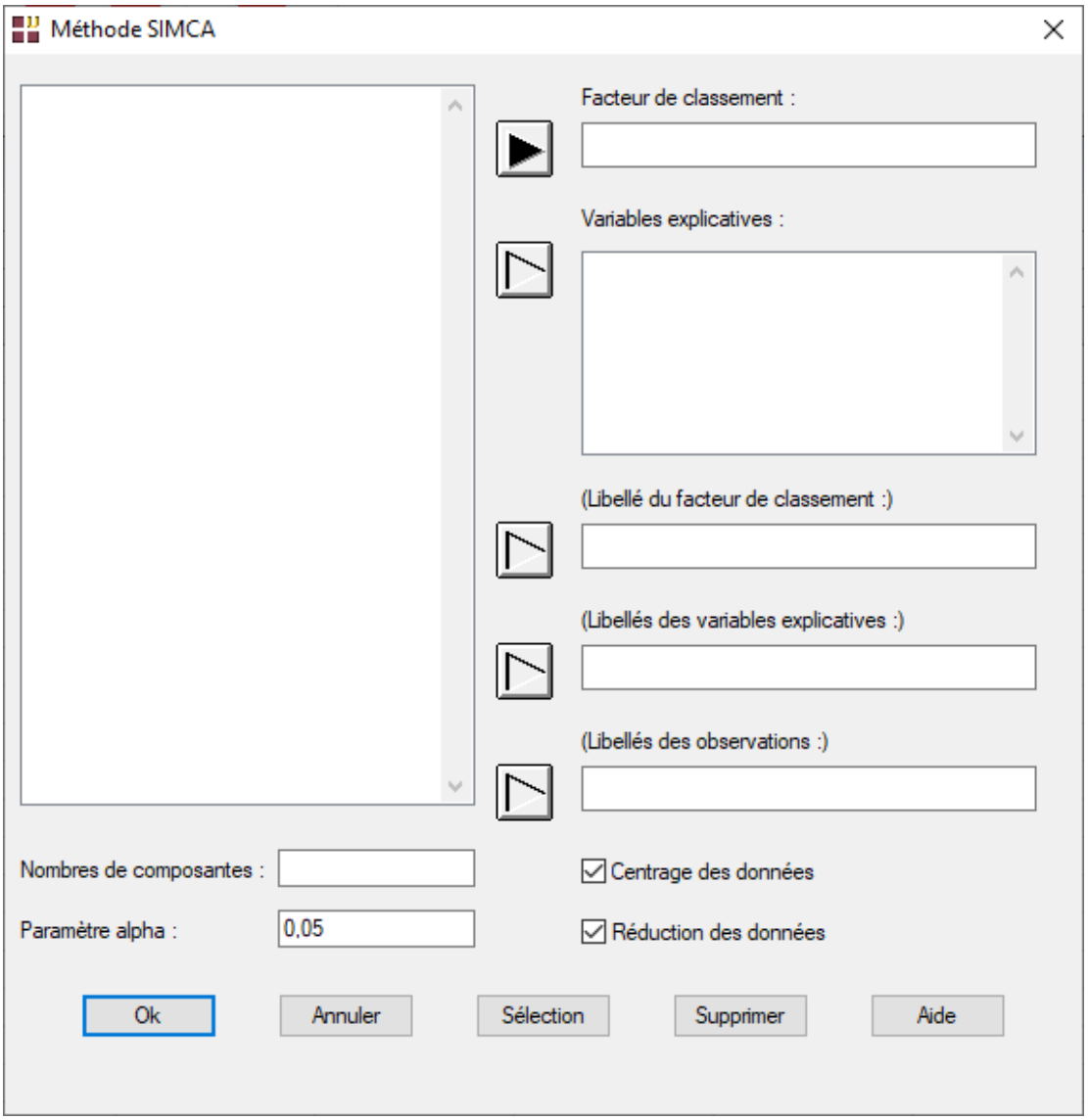
Définition

La méthode SIMCA (Soft Independent Modeling of Class Analogy) est une technique de classement proposée par Svante Wold dans les années 1970. Cette méthode supervisée est basée sur l'analyse en composantes principales (ACP). Pour chaque classe, une ACP est réalisée utilisant uniquement les observations de cette classe. Les différents modèles obtenus pour chacune des classes peuvent avoir des nombres de composantes différents. Ces modèles permettent de prévoir l'appartenance ou non d'une observation du jeu d'apprentissage ou de prévision à une classe prédéfinie mais également de déterminer si une observation appartient à plusieurs classes (recouvrement) ou à aucune classe (possible point aberrant ou nouvelle classe).

Cette procédure est basée sur le package R 'mdatools'.

Entrée des données

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



Cette boîte de dialogue permet de définir le facteur de classement qualitatif, les variables explicatives quantitatives, les libellés optionnels du facteur de classement, des variables explicatives et des observations.

Elle permet également de préciser si les données doivent être centrées et réduites, la valeur du paramètre alpha (niveau de signification pour les limites statistiques dans le graphique de Cooman) et les nombres de composantes pour chacun des modèles (obtenus par exemple par des analyses préalables ACP ou NIPALS des jeux d'apprentissage de chacun des modèles). Si ces nombres ne sont pas spécifiés, toutes les composantes sont extraites.

Données manquantes

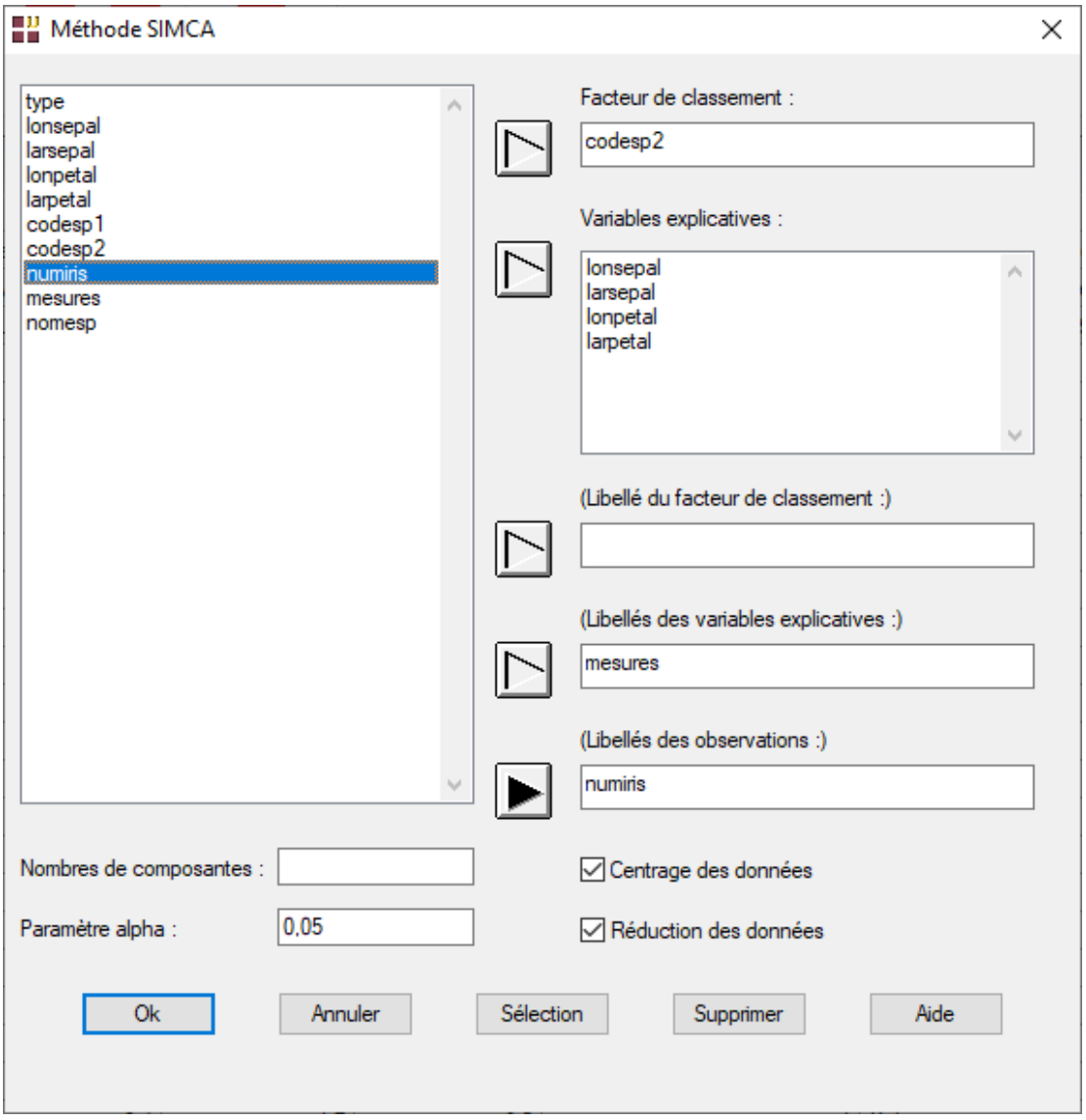
Les valeurs manquantes du facteur de classement définissent l'échantillon de prévision. Les données manquantes pour les variables explicatives sont autorisées.

Exemple 1 : Fichier IRIS3

Pour ce premier exemple, nous utiliserons le fichier IRIS3 pour illustrer cette procédure.

Ce fichier contient les données relatives à 150 iris de 3 espèces : Iris Setosa, Iris Versicolor et Iris Virginica. Les mesures effectuées sont, en millimètres, la longueur du sépale (lonsepal), la longueur du pétale (lonpetal), la largeur du sépale (larsepal) et la largeur du pétale (larpetal).

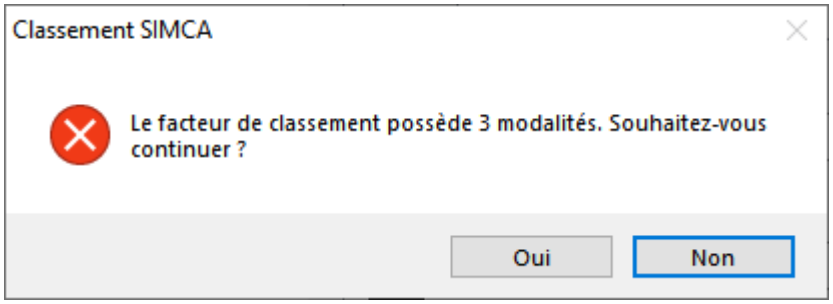
Renseignons la boîte de dialogue comme montré ci-dessous.



Sélectionnons la variable *codesp2* comme facteur de classement et les variables *lonsepal*, *larsepal*, *lonpetal* et *lonsepal* comme variables explicatives.

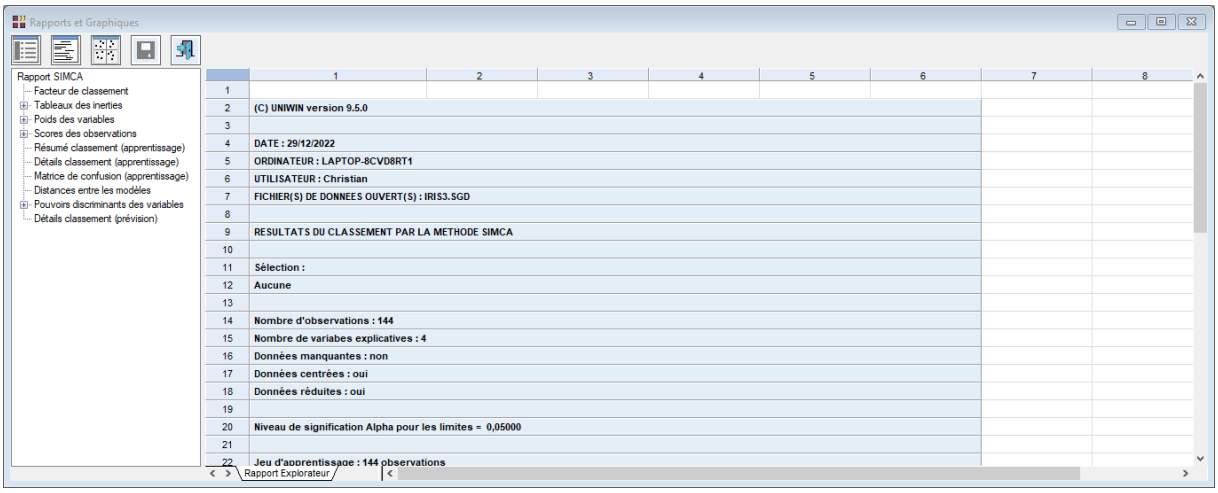
Conservons les autres paramètres aux valeurs par défaut.

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



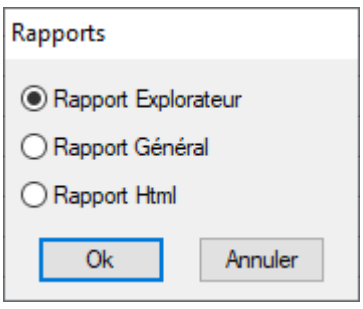
Après avoir visualisé le message informatif nous indiquant le nombre de modalités du facteur de classement, nous continuons l'analyse.

Après quelques instants, la fenêtre « Rapports et Graphiques » s'affiche :

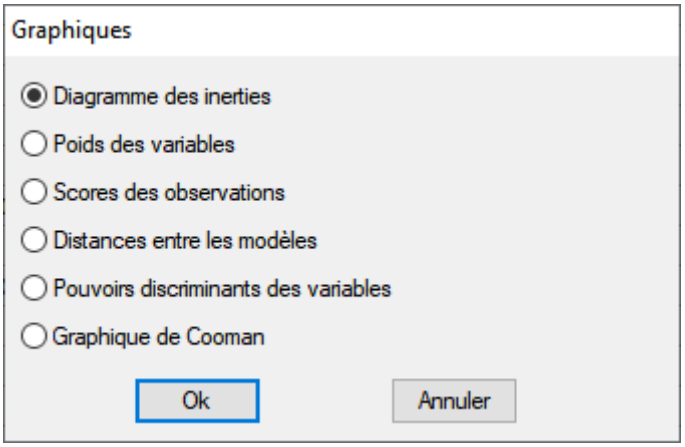


La barre d'outils 'Rapports et Graphiques' permet par l'icône 'Données'  de rappeler la boîte de dialogue d'entrée des données.

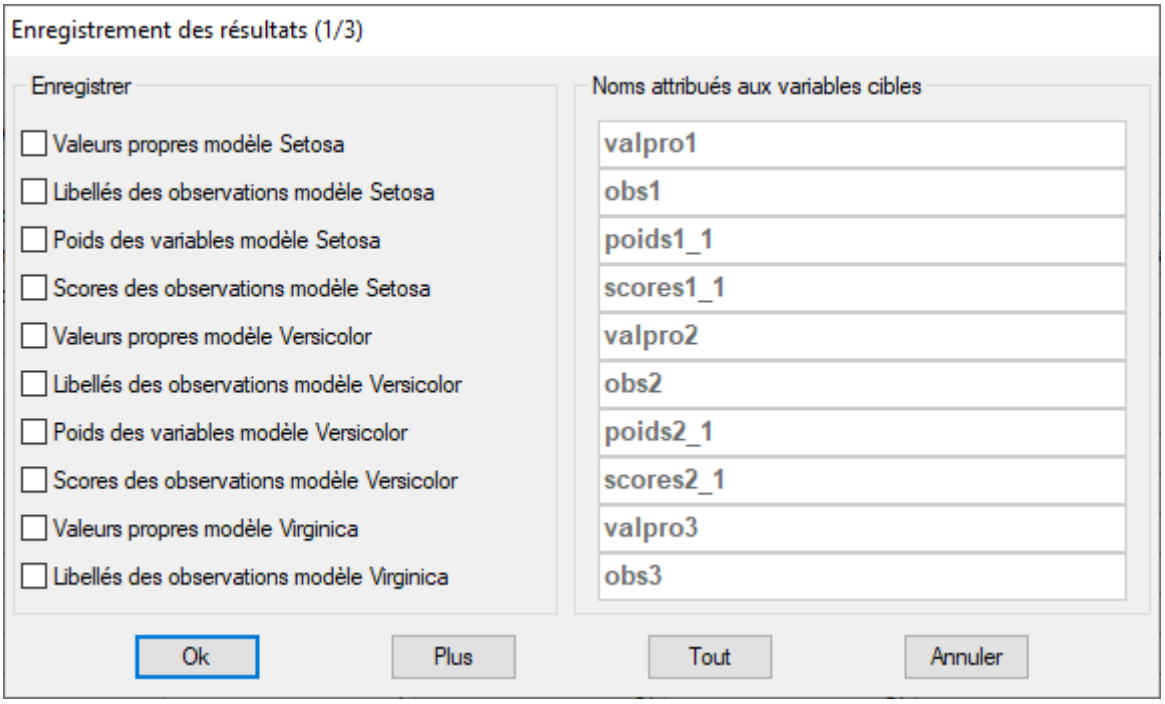
L'icône 'Rapports'  affiche la boîte de dialogue des options pour les rapports :



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.

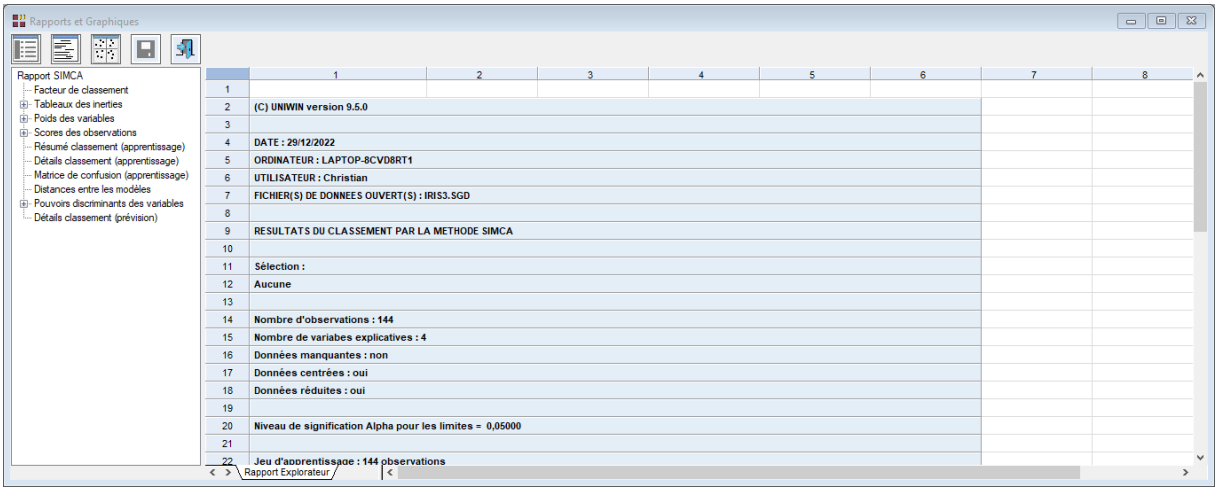


L'icône 'Quitter'  permet de quitter l'analyse.

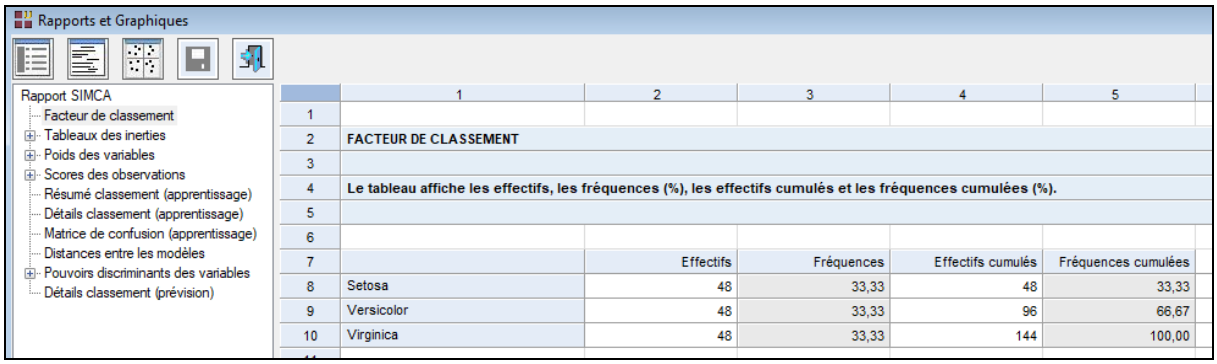
L'option Rapports

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

Le premier tableau affiche les nombres d'observations pour les jeux d'apprentissage et de prévision et rappelle les paramètres de l'étude.

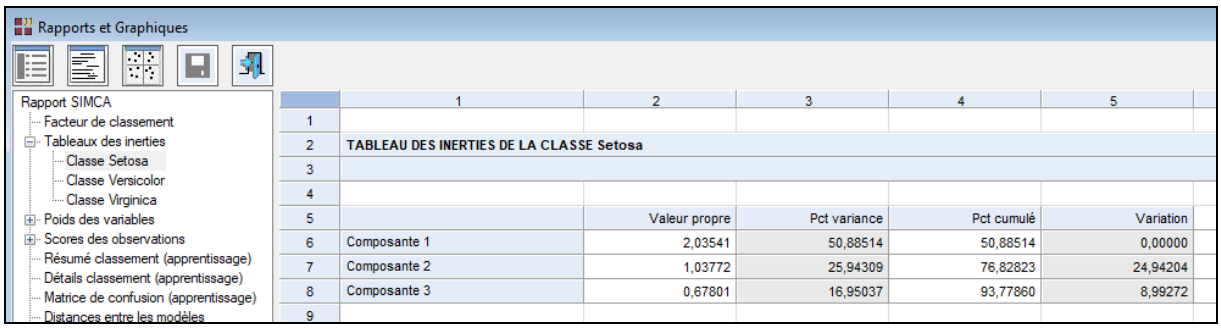


Le deuxième tableau affiche un tri à plat du facteur de classement.



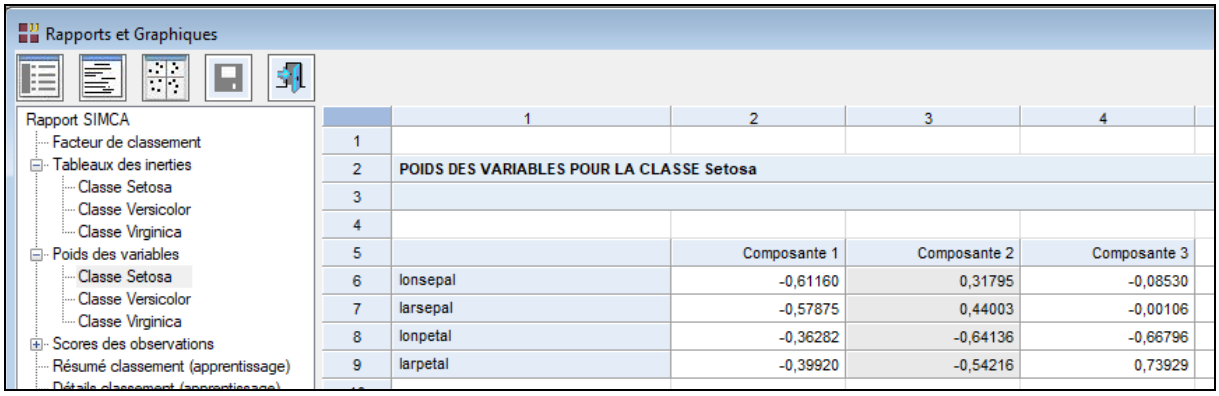
	Effectifs	Fréquences	Effectifs cumulés	Fréquences cumulées
Setosa	48	33,33	48	33,33
Versicolor	48	33,33	96	66,67
Virginica	48	33,33	144	100,00

Le troisième tableau affiche les inerties pour chacun des modèles construits.



	Valeur propre	Pct variance	Pct cumulé	Variation
Composante 1	2,03541	50,88514	50,88514	0,00000
Composante 2	1,03772	25,94309	76,82823	24,94204
Composante 3	0,67801	16,95037	93,77860	8,99272

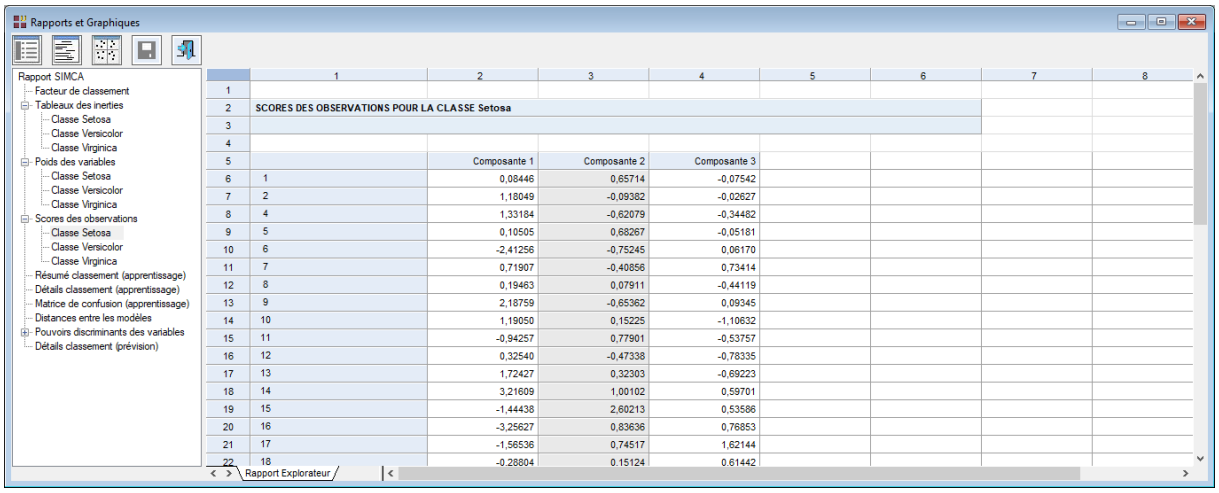
Le quatrième tableau affiche les poids des variables pour chacun des modèles construits.



POIDS DES VARIABLES POUR LA CLASSE Setosa

	Composante 1	Composante 2	Composante 3
lonsepal	-0,61160	0,31795	-0,08530
larsepal	-0,57875	0,44003	-0,00106
lonpetal	-0,36282	-0,64136	-0,66796
larpetal	-0,39920	-0,54216	0,73929

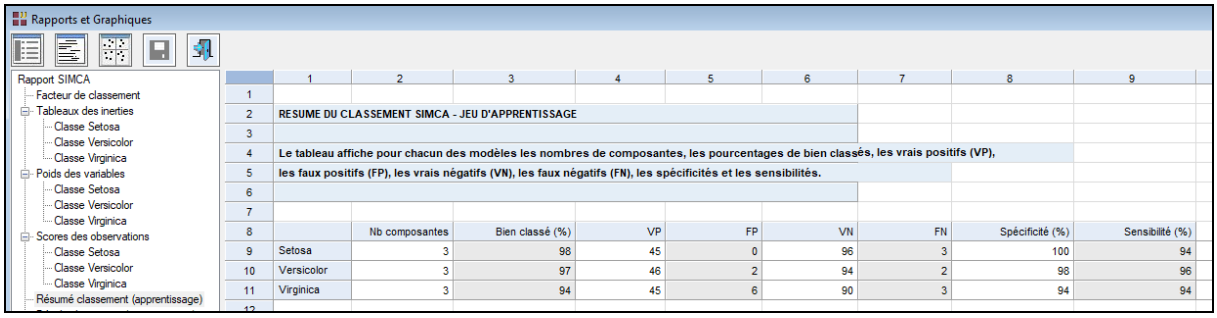
Le cinquième tableau affiche les scores des observations pour chacun des modèles construits.



SCORES DES OBSERVATIONS POUR LA CLASSE Setosa

	Composante 1	Composante 2	Composante 3
1	0,08446	0,65714	-0,07542
2	1,18049	-0,09382	-0,02627
3	1,33184	-0,62079	-0,34482
4	0,10505	0,68267	-0,05181
5	-2,41256	-0,75245	0,06170
6	0,71907	-0,40856	0,73414
7	0,19463	0,07911	-0,44119
8	2,18759	-0,65362	0,09345
9	1,19050	0,15225	-1,10632
10	-0,94257	0,77901	-0,53757
11	0,32540	-0,47338	-0,78335
12	1,72427	0,32303	-0,69223
13	3,21609	1,00102	0,59701
14	-1,44438	2,60213	0,53586
15	-3,25627	0,83636	0,76853
16	-1,56536	0,74517	1,62144
17	-0,28804	0,15124	0,61442

Le sixième tableau affiche pour chacun des modèles élaborés en utilisant le jeu d'apprentissage les nombres de composantes extraites, les pourcentages de bien classés, les vrais positifs (VP), les faux positifs (FP), les vrais négatifs (VN), les faux négatifs (FN), les spécificités et sensibilités.



RESUME DU CLASSEMENT SIMCA - JEU D'APPRENTISSAGE

Le tableau affiche pour chacun des modèles les nombres de composantes, les pourcentages de bien classés, les vrais positifs (VP), les faux positifs (FP), les vrais négatifs (VN), les faux négatifs (FN), les spécificités et les sensibilités.

	Nb composantes	Bien classé (%)	VP	FP	VN	FN	Spécificité (%)	Sensibilité (%)
Setosa	3	98	45	0	96	3	100	94
Versicolor	3	97	46	2	94	2	98	96
Virginica	3	94	45	6	90	3	94	94

La spécificité (proportion des vrais négatifs bien détectés par le modèle) est définie par :

$$\frac{100 * VN}{(VN + FP)}$$

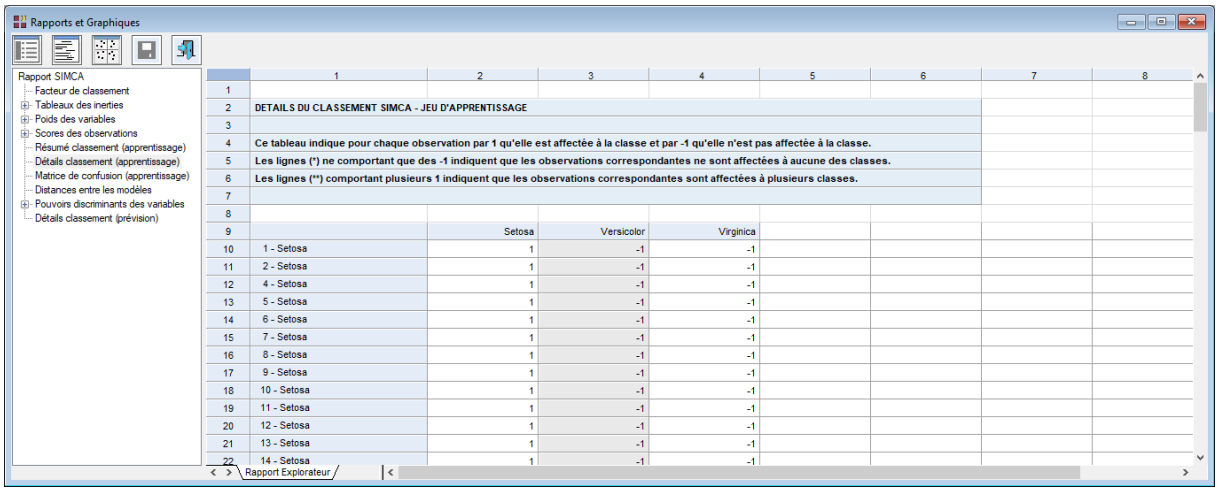
et la sensibilité (proportions des vrais positifs bien détectés par le modèle) par :

$$\frac{100 * VP}{(VP + FN)}$$

Le pourcentage de bien classés est défini par :

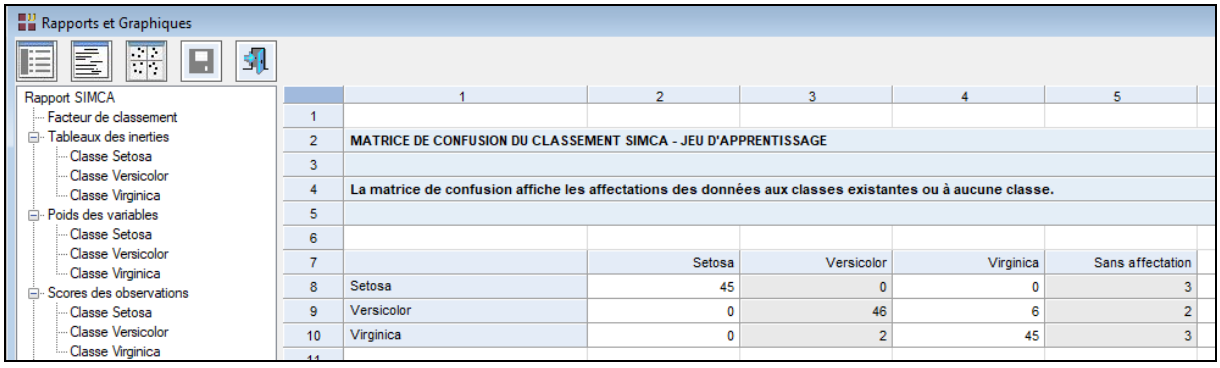
$$\frac{100 * (VP + VN)}{(VP + VN + FP + FN)}$$

Le tableau suivant affiche les détails du classement pour le jeu d'apprentissage.

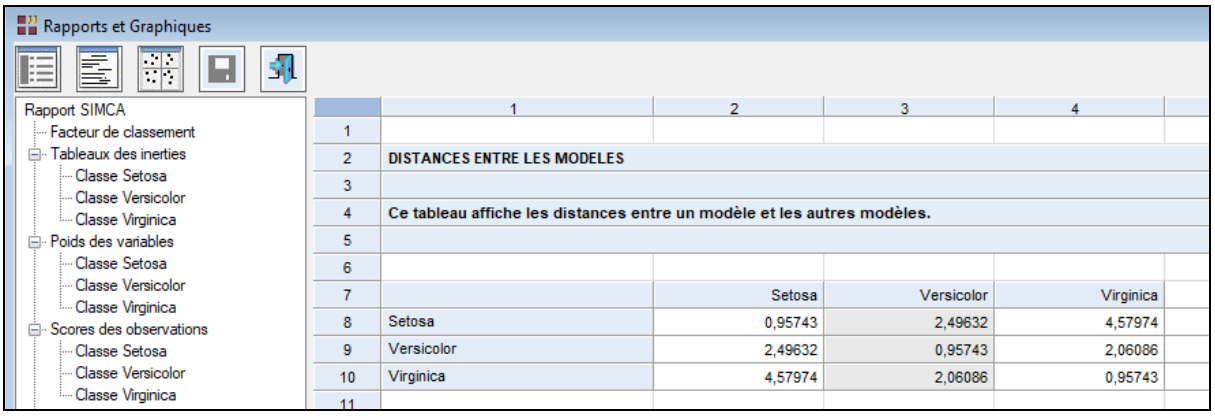


Chaque observation est affectée (1) ou non affectée (-1) à un modèle. Une observation peut être affectée à plusieurs modèles, cela indique des recouvrements des classes observées. Une observation peut être affectée à aucun des modèles, cela indique un possible point aberrant ou une observation définissant une éventuelle nouvelle classe.

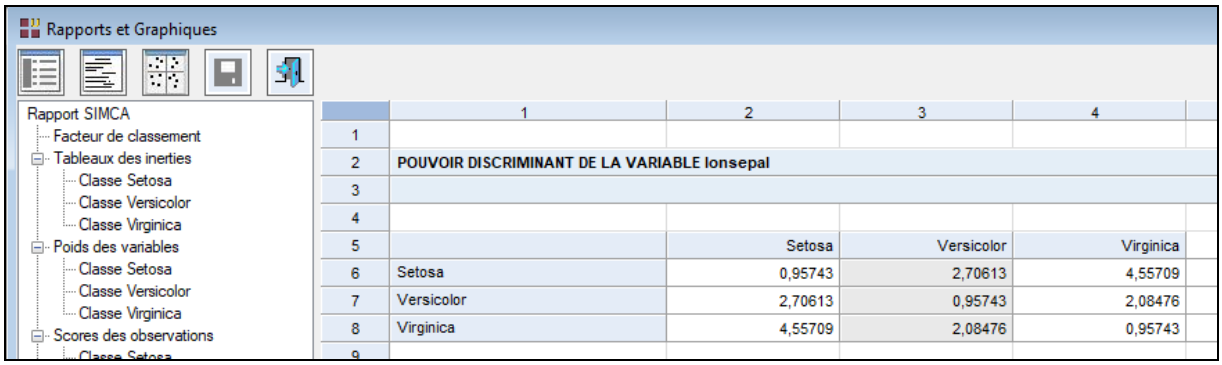
Le tableau suivant affiche la matrice de confusion :



Les tableaux suivants affiche les distances entre les modèles et les pouvoirs discriminants des variables pour chacun des modèles :



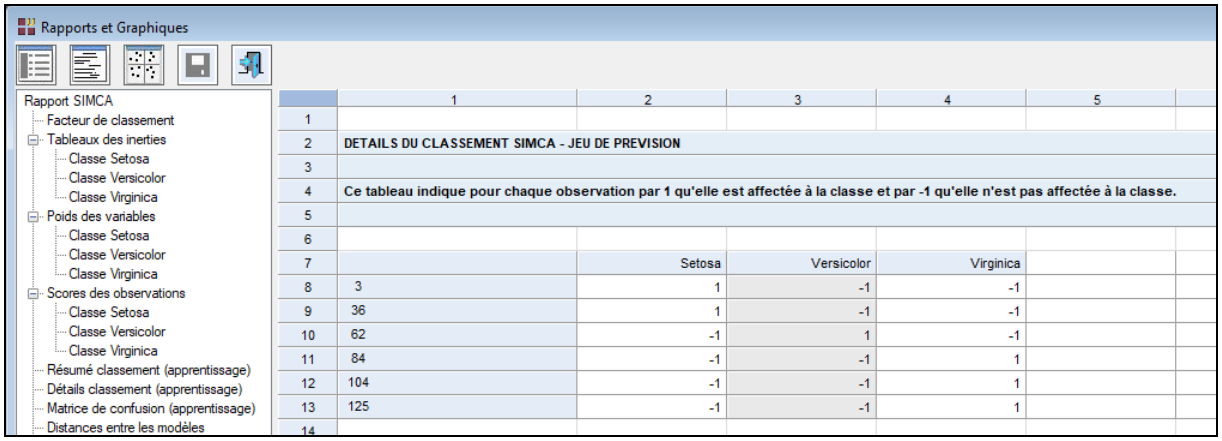
	1	2	3	4
1				
2	**DISTANCES ENTRE LES MODELES**			
3				
4	**Ce tableau affiche les distances entre un modèle et les autres modèles.**			
5				
6				
7		Setosa	Versicolor	Virginica
8	Setosa	0,95743	2,49632	4,57974
9	Versicolor	2,49632	0,95743	2,06086
10	Virginica	4,57974	2,06086	0,95743
11				



	1	2	3	4
1				
2	**POUVOIR DISCRIMINANT DE LA VARIABLE lonsepal**			
3				
4				
5		Setosa	Versicolor	Virginica
6	Setosa	0,95743	2,70613	4,55709
7	Versicolor	2,70613	0,95743	2,08476
8	Virginica	4,55709	2,08476	0,95743
9				

Voir l'option 'Graphiques' pour les détails sur les calculs effectués.

Le dernier tableau affiche les détails du classement pour les observations du jeu de prévision.

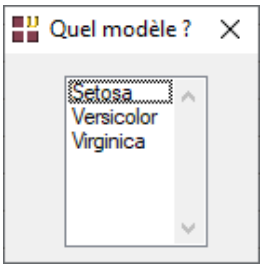


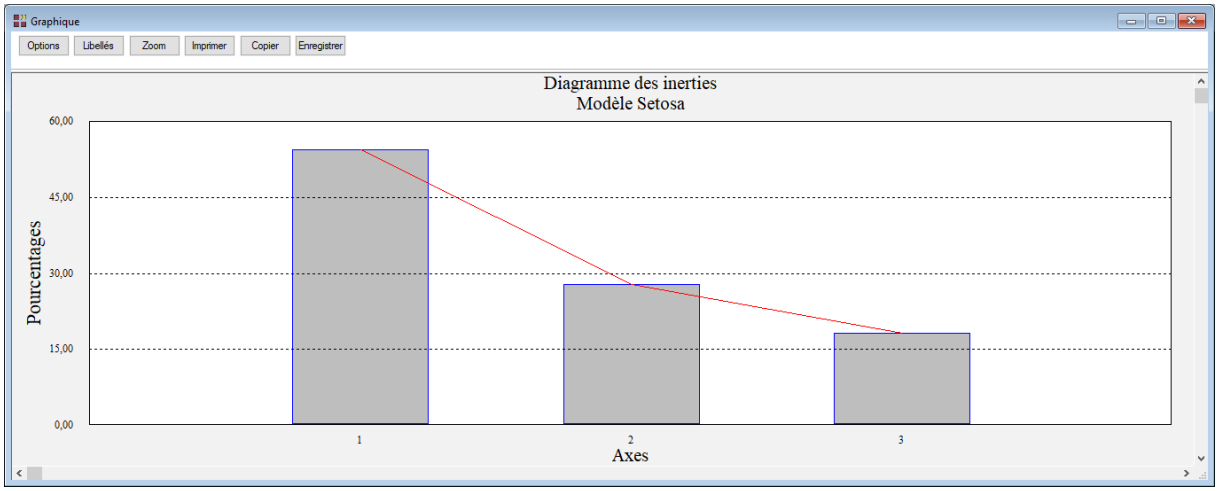
	1	2	3	4	5
1					
2	**DETAILS DU CLASSEMENT SIMCA - JEU DE PREVISION**				
3					
4	**Ce tableau indique pour chaque observation par 1 qu'elle est affectée à la classe et par -1 qu'elle n'est pas affectée à la classe.**				
5					
6					
7		Setosa	Versicolor	Virginica	
8	3	1	-1	-1	
9	36	1	-1	-1	
10	62	-1	1	-1	
11	84	-1	-1	1	
12	104	-1	-1	1	
13	125	-1	-1	1	
14					

L'option Graphiques

- Diagramme des inerties

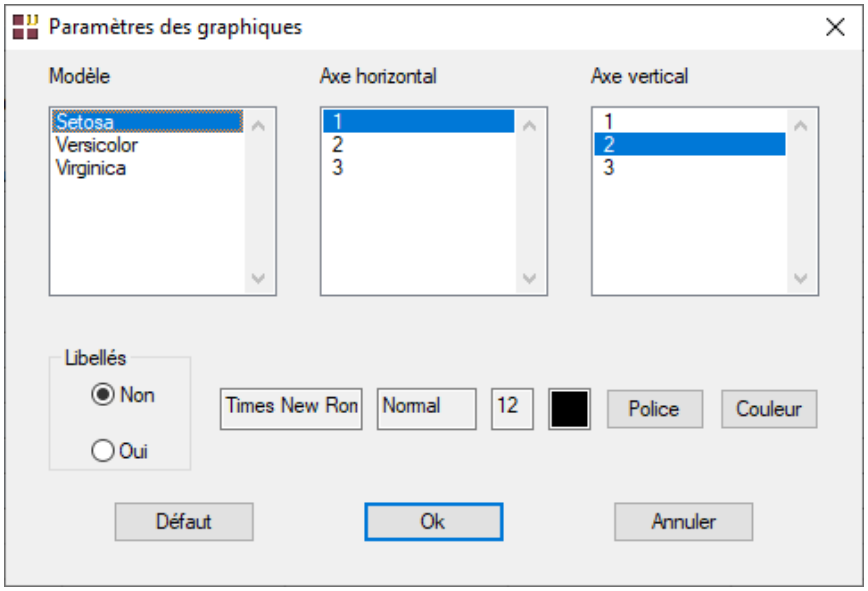
Ce graphique affiche le diagramme des inerties pour chacun des modèles. Il permet de choisir le modèle.

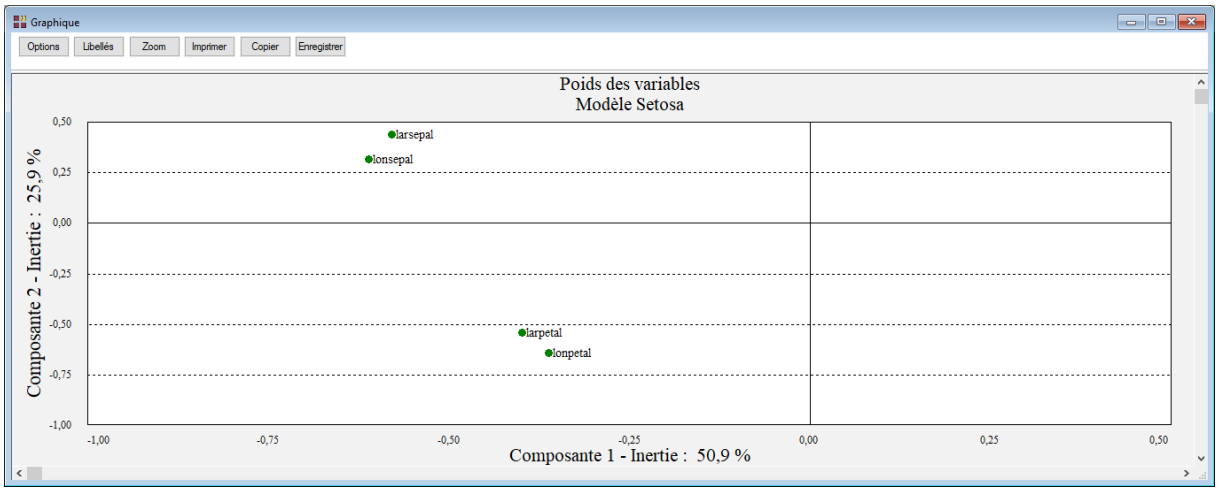




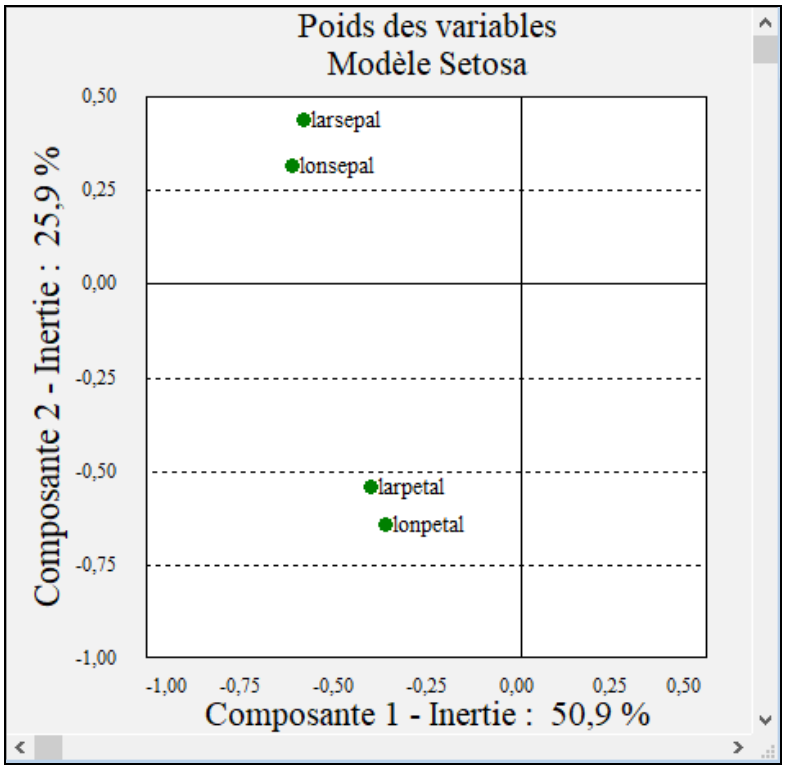
- Poids des variables

Ce graphique affiche les poids des variables pour chacun des modèles. Il permet de choisir le plan factoriel désiré et d'afficher ou non les libellés des variables.





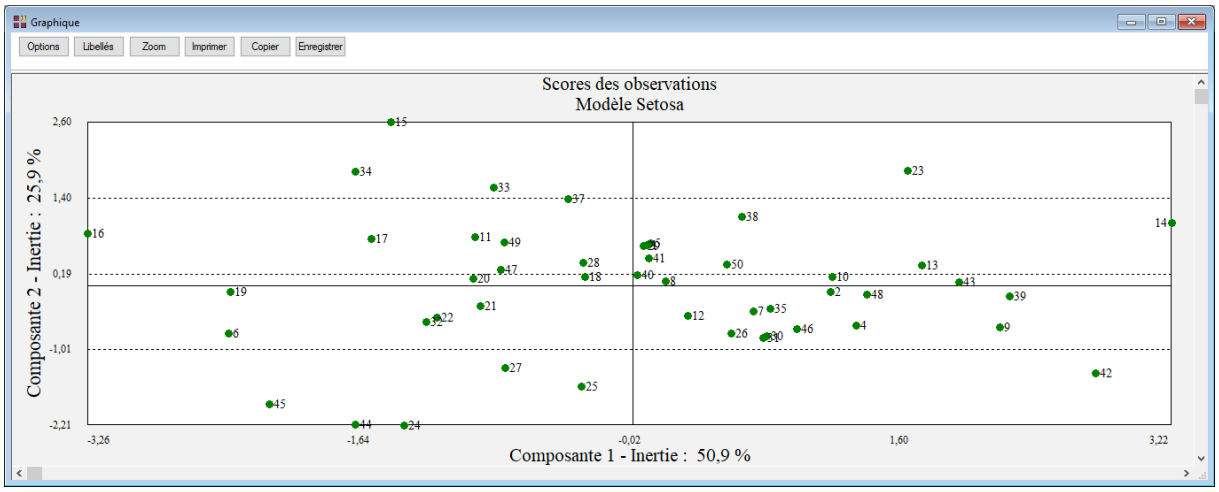
Par les options graphiques (échelles des axes), il est possible d'afficher un graphique orthonormé.



- Scores des observations

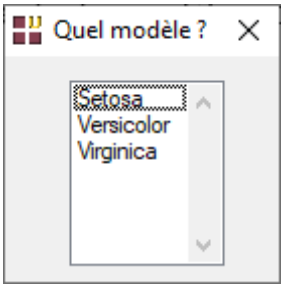
Ce graphique affiche les scores des observations pour chacun des modèles.

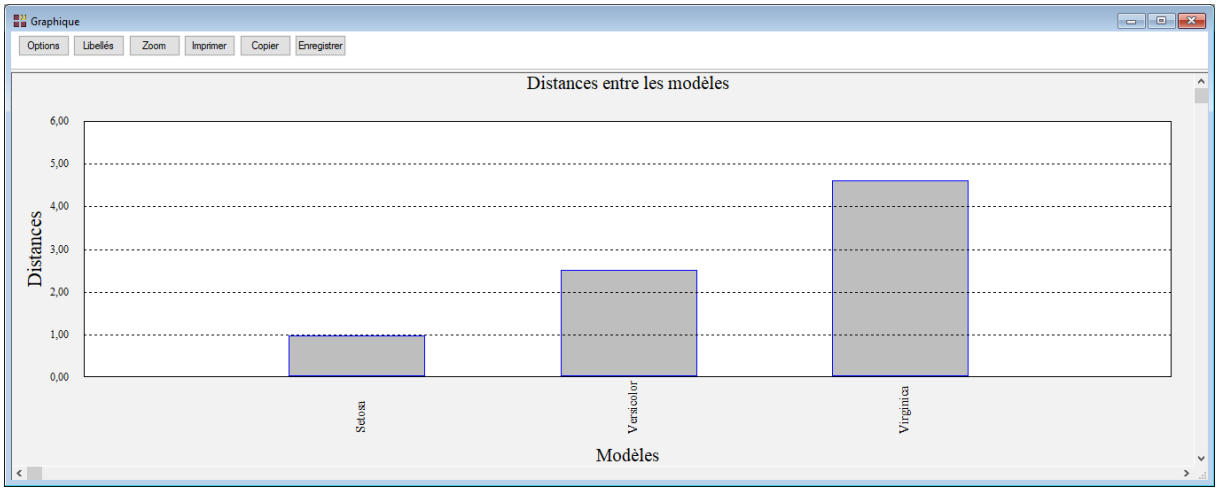
Il permet comme pour le poids des variables de choisir le plan factoriel désiré et d'afficher ou non les libellés des observations.



- Distances entre les modèles

Ce graphique affiche les similarités entre un modèle sélectionné et les autres modèles en utilisant les variances résiduelles.





Plus précisément, voici l'algorithme mis en œuvre.

Soient m1 et m2 deux modèles ayant les nombres respectifs de composantes A1 et A2. Ces modèles ont été élaborés en utilisant les jeux d'apprentissage X1 et X2 ayant respectivement n1 et n2 observations.

Faisons les calculs suivants :

1. Etablissons le modèle m1 pour X2 et calculons les résidus E12
2. Calculons la variance des résidus s12 = somme(E12^2) / n1
3. Etablissons le modèle m2 pour X1 et calculons les résidus E21
4. Calculons la variance des résidus s21 = somme(E21^2) / n2
5. Calculons la variance des résidus pour m1 avec s1 = somme(E1^2) / (n1 - A1 - 1)
6. Calculons la variance des résidus pour m2 avec s2 = somme(E2^2) / (n2 - A2 - 1)

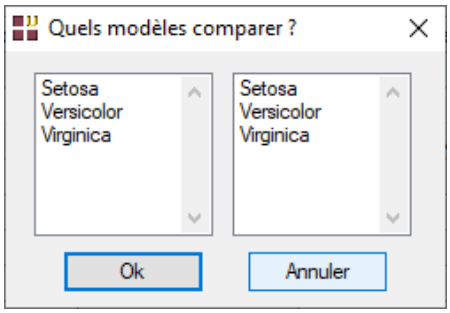
La distance au modèle est calculée par : d = racine((s12 + s21) / (s1 + s2))

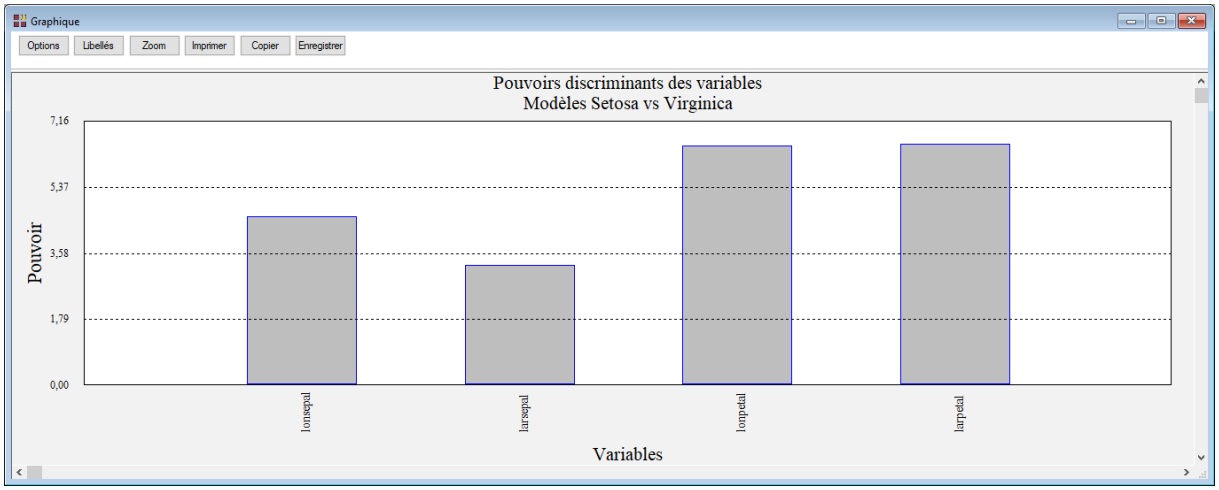
Si les deux modèles et les deux jeux d'apprentissage sont identiques, alors la distance est égale à racine((n - A - 1) / n).

En général, si la distance entre les modèles est inférieure à 1, cela indique que les classes se recouvrent. Si la distance est supérieure à 3, les classes sont bien séparées.

- Pouvoirs discriminants des variables

Ce graphique affiche pour deux modèles sélectionnés les pouvoirs discriminants des variables c'est-à-dire la capacité de ces variables à séparer les classes.

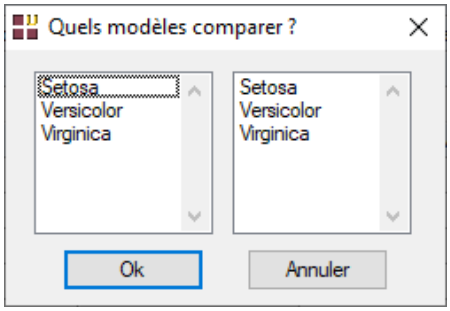


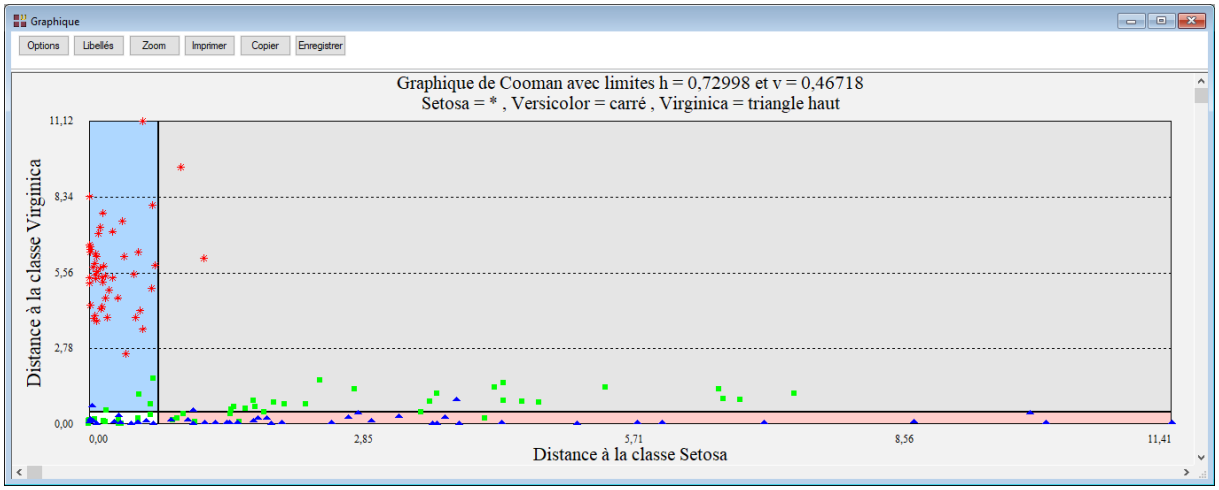


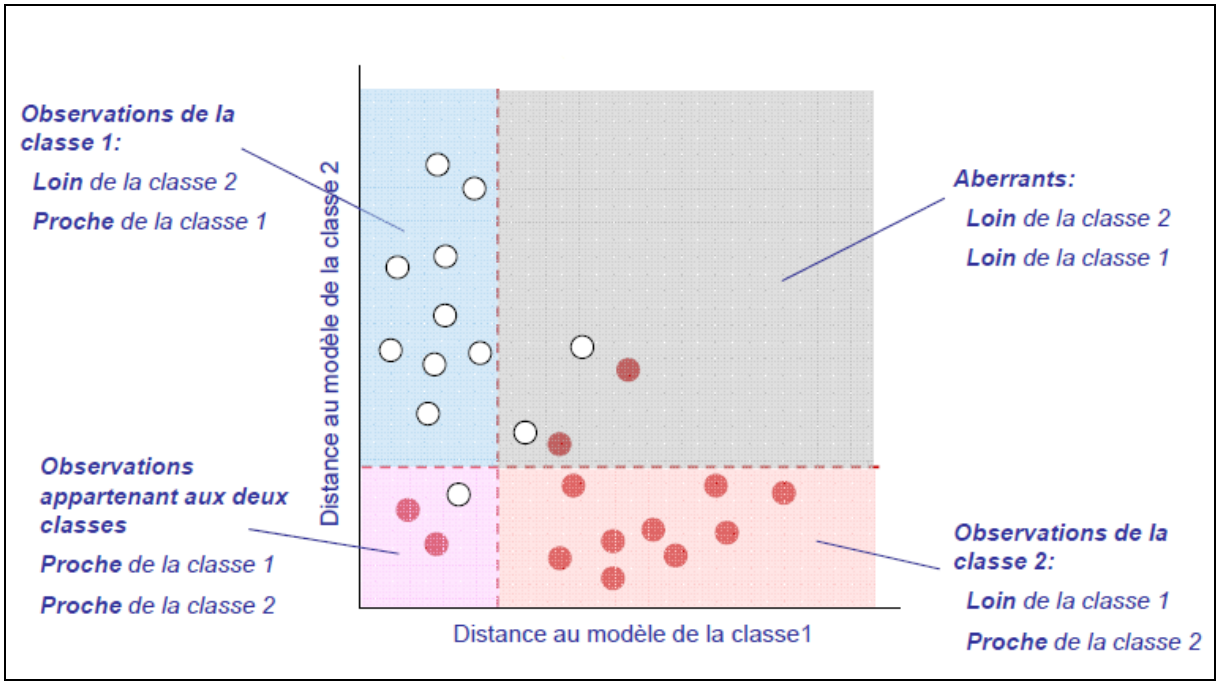
Le pouvoir discriminant est calculé comme la distance au modèle en utilisant la variance des résidus. Toutefois, dans ce cas, au lieu de sommer la variance sur toutes les variables, le calcul est fait de façon séparé pour chaque variable. Un pouvoir discriminant égal ou supérieur à 3 est considéré comme élevé.

- Graphique de Cooman

Cette option affiche pour les deux modèles sélectionnés le graphique de Cooman.







Origine : Julien Boccard – Support formation « omics »

Les limites horizontale et verticale sont calculées en utilisant la loi du Khi-carré avec le niveau de signification alpha.

Exemple 2 : Fichier WINES2

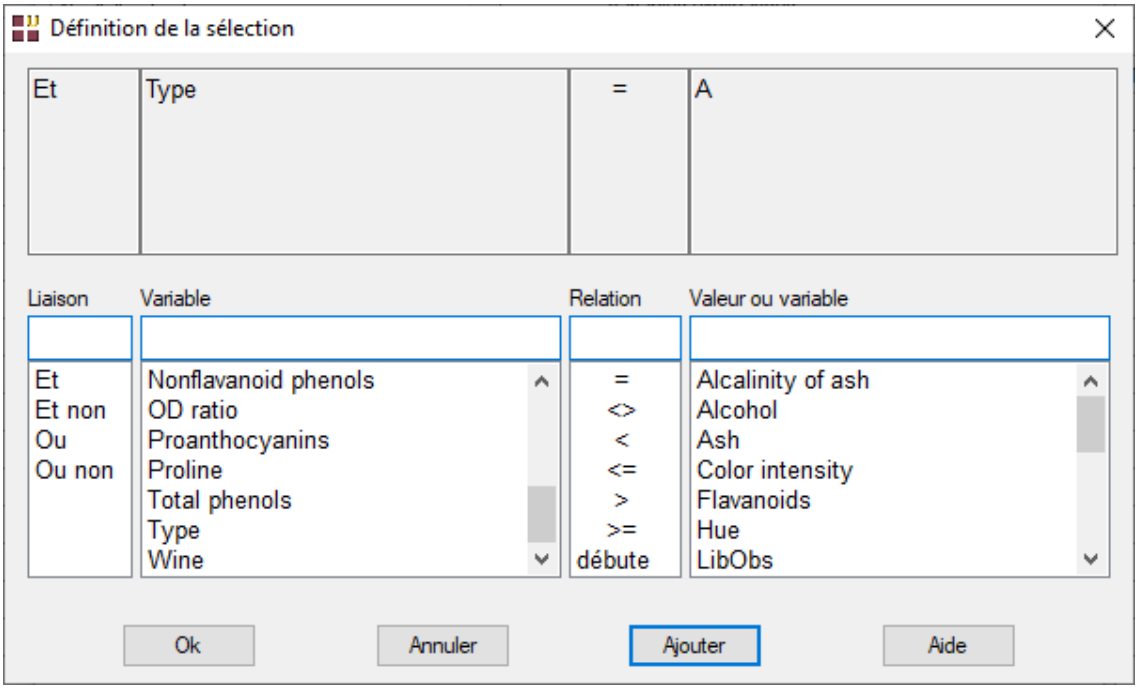
Les données sont le résultat de l'analyse chimique de 180 échantillons de vins issus de trois cultivars différents et provenant d'une même région en Italie (source : https://archive.ics.uci.edu/ml/datasets/wine). Elles sont constituées des treize caractéristiques chimiques et spectroscopiques suivantes :

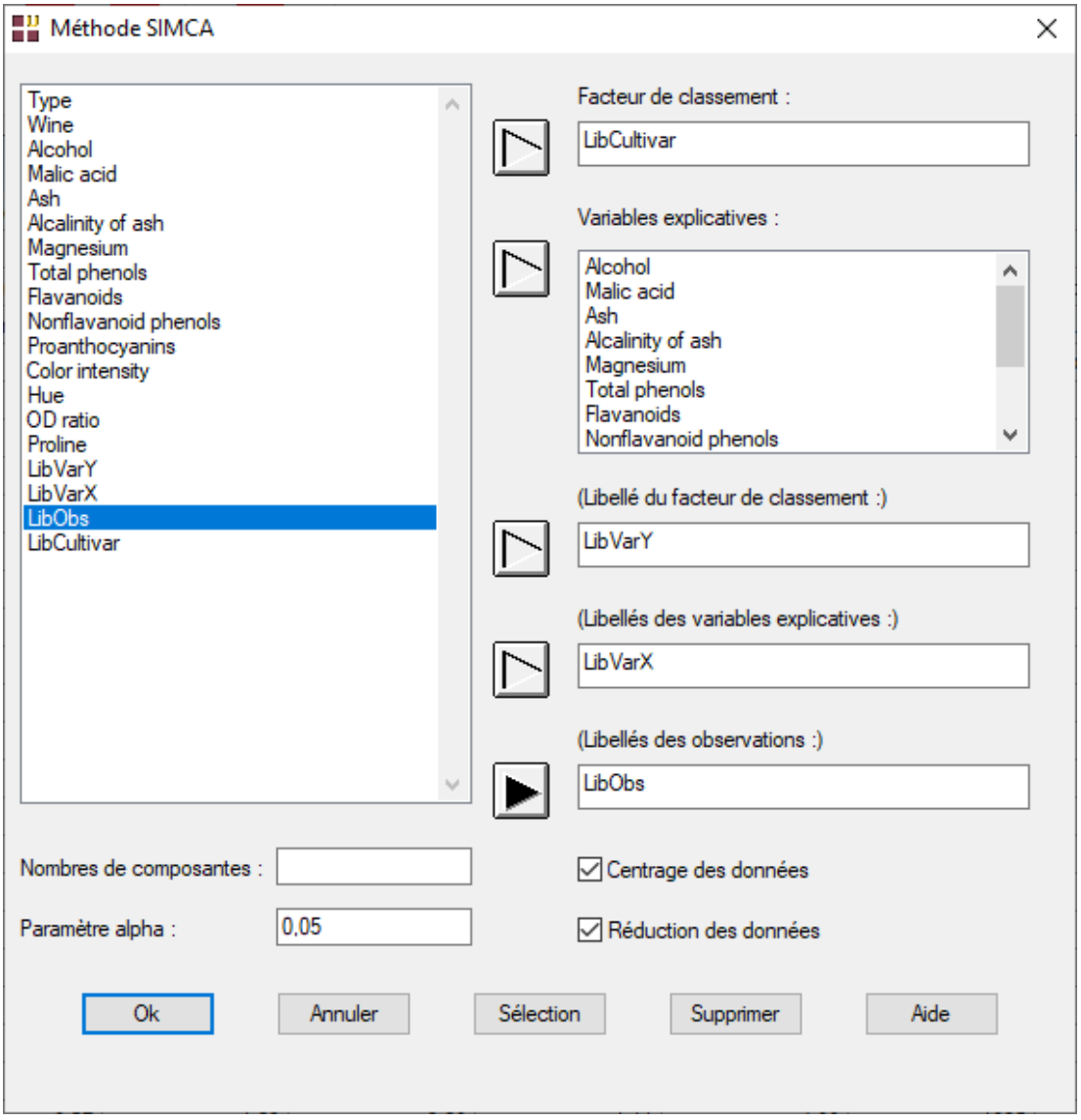
wine	N° du cultivar
alcohol	Alcool
malic acid	Acide malique
ash	Cendres
ash alkalinity	Alcalinité des cendres
magnesium	Magnésium
tot. phenols	Total des phénols
flavonoids	Flavonoïdes
non-flav. phenols	Phénols non-flavonoïdes
proanth	Proanthocyanidines
col. int.	Intensité de couleur
col. hue	Teinte de couleur
OD ratio	Rapport OD
proline	Proline

Cliquons sur l'icône SIMCA dans le ruban Expliquer et renseignons la boîte de dialogue comme montré ci-après (les variables explicatives de 'Alcohol' à 'Proline' sont sélectionnées).

Cliquons sur le bouton 'Sélection' pour définir les données du jeu d'apprentissage et utilisons la colonne 'Type' du fichier des données pour cela.

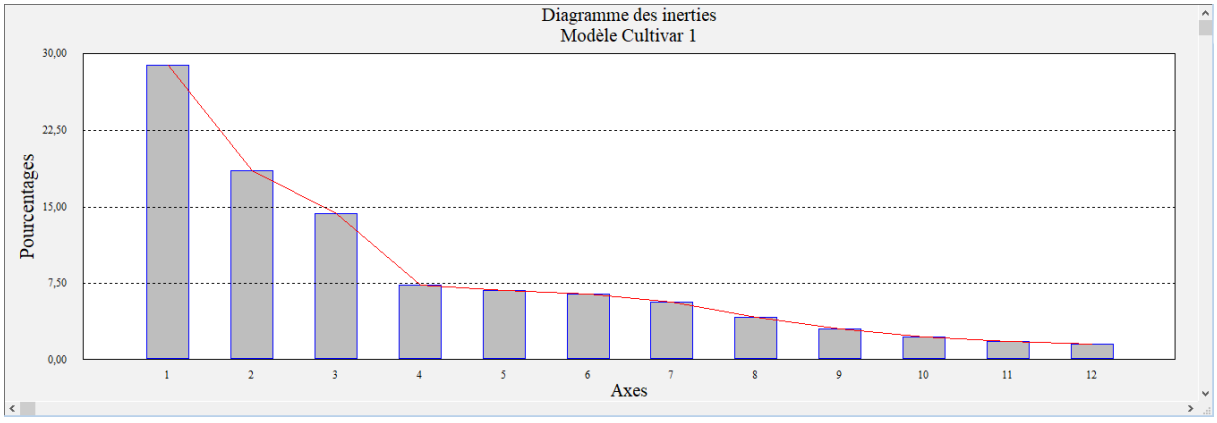
151 observations sont sélectionnées pour le jeu d'apprentissage.

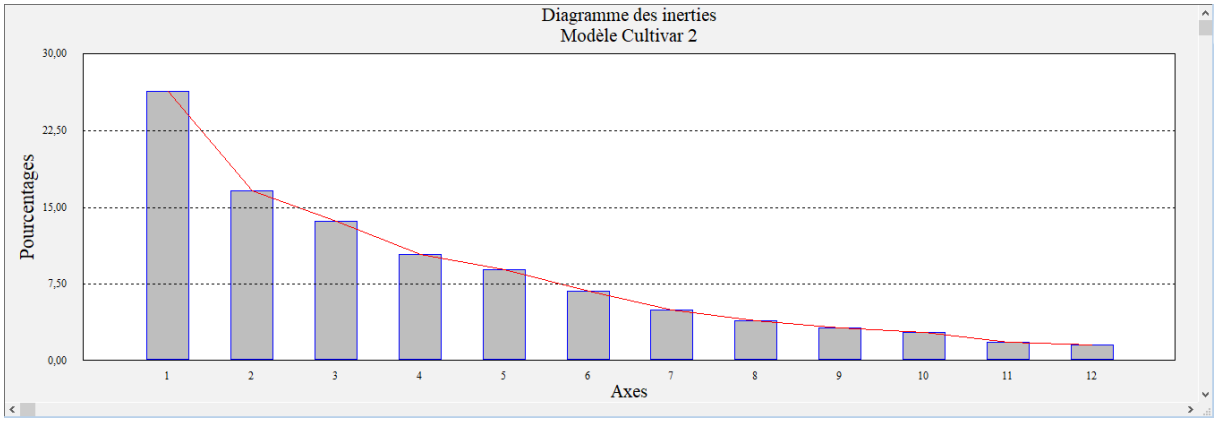


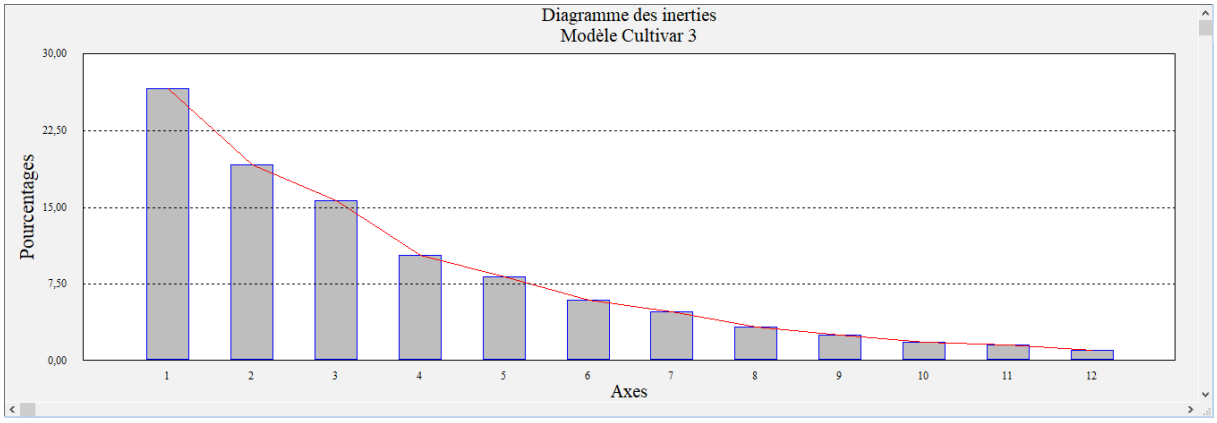


Cliquons sur Ok. Après quelques instants la fenêtre 'Rapports et Graphiques' s'affiche.

Regardons les diagrammes des inerties pour chacun des trois modèles.







Ces résultats nous indiquent que les nombres de composantes à retenir pour chacun des modèles sont probablement respectivement de 3, 5 et 5.

Exécutons donc à nouveau l'analyse en cliquant sur l'icône  de la barre d'outils et entrons ces nombres dans le champ 'Nombres de composantes' puis redéfinissons notre sélection des données du jeu d'apprentissage 'Type = A'. Cliquons sur Ok.

Voici quelques-uns des résultats obtenus.

RESUME DU CLASSEMENT SIMCA - JEU D'APPRENTISSAGE

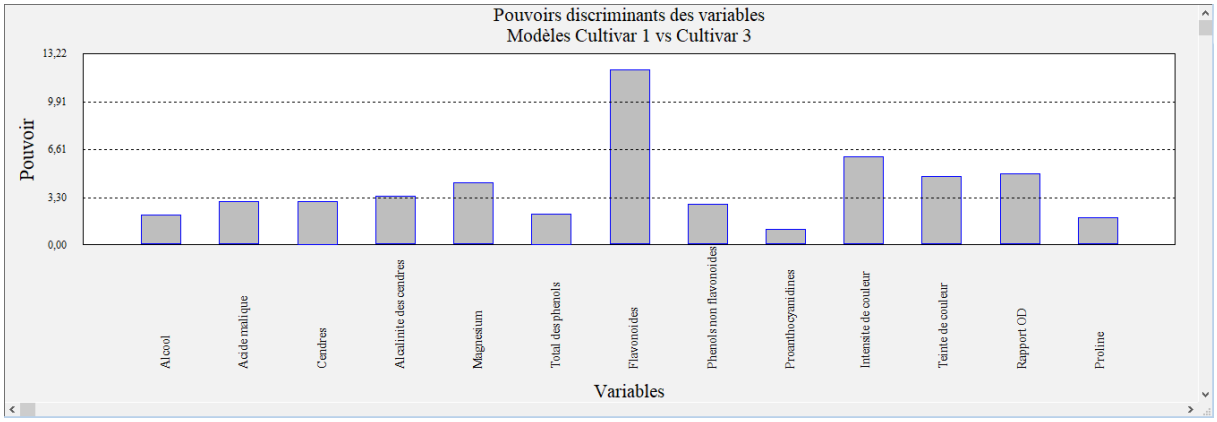
Le tableau affiche pour chacun des modèles les nombres de composantes, les pourcentages de bien classés, les vrais positifs (VP), les faux positifs (FP), les vrais négatifs (VN), les faux négatifs (FN), les spécificités et les sensibilités.

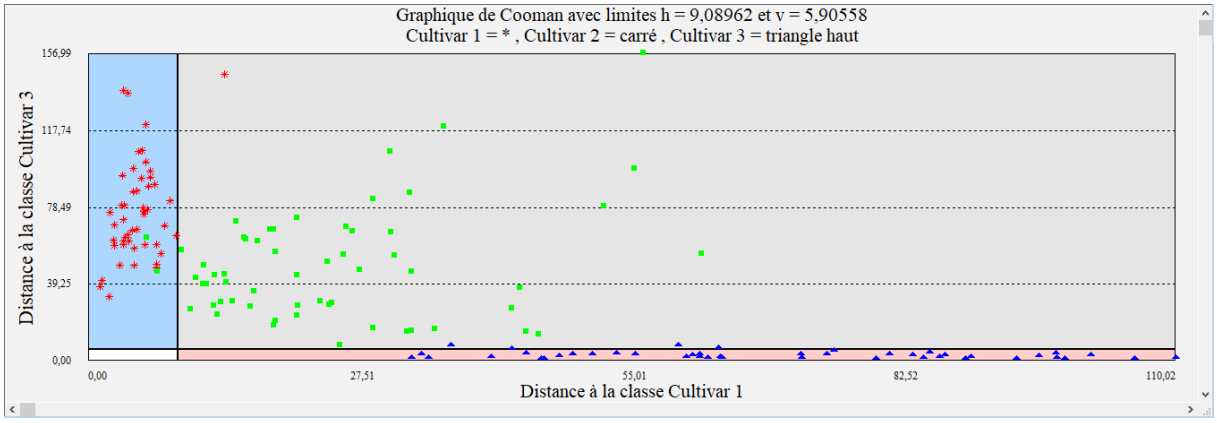
	Nb composantes	Bien classé (%)	VP	FP	VN	FN	Spécificité (%)	Sensibilité (%)
Cultivar 1	3	99	47	1	102	1	99	98
Cultivar 2	5	88	54	13	79	5	86	92
Cultivar 3	5	100	44	0	107	0	100	100

MATRICE DE CONFUSION DU CLASSEMENT SIMCA - JEU D'APPRENTISSAGE

La matrice de confusion affiche les affectations des données aux classes existantes ou à aucune classe.

	Cultivar 1	Cultivar 2	Cultivar 3	Sans affectation
Cultivar 1	47	6	0	1
Cultivar 2	1	54	0	5
Cultivar 3	0	7	44	0





Les variables créées par la procédure

Voici la liste des variables créées par la procédure.

Variable	*Contenu*
valpro	Valeurs propres pour le modèle
obs	Libellés des observations de la classe
poids	Poids des variables pour le modèle
scores	Scores des observations pour le modèle
obsapp	Libellés des observations (jeu d'apprentissage)
cpa	Classes prévues (jeu d'apprentissage)
coa	Classes observées apprentissage
resume	Résumé statistique
confusion	Matrice de confusion
distmod	Distances entre les modèles
discrivar	Pouvoirs discriminants de chaque variable
obsprev	Libellés des observations (jeu de prévision)
cpp	Classes prévues (jeu de prévision)

Références

Documentation du package R ‘mdatools’ (2021)

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

Exemple 2

https://archive.ics.uci.edu/ml/datasets/wine