UNIWIN VERSION 10.2.0

ANALYSE DISCRIMINANTE QUALITATIVE

Révision : 25/03/2025

Définition

L'Analyse Discriminante Qualitative (ADQ) est une généralisation de l'Analyse Factorielle Discriminante (AFD) dans le cas où les variables explicatives sont qualitatives et non plus quantitatives.

La première étape de l'analyse consiste à mettre en œuvre une Analyse des Correspondances Multiples (ACM) des variables qualitatives.

La deuxième étape remplace les variables qualitatives d'origine par les coordonnées sur les axes factoriels issus de l'ACM et effectue sur ces données une Analyse Factorielle Discriminante (AFD).

Les fonctions discriminantes sont ensuite exprimées en fonction des indicatrices des modalités des variables qualitatives d'origine.

En fonction des données et des paramètres définis par l'utilisateur, l'analyse ADQ réalise automatiquement les études de la population d'apprentissage et des éventuelles populations de validation et de prévision.

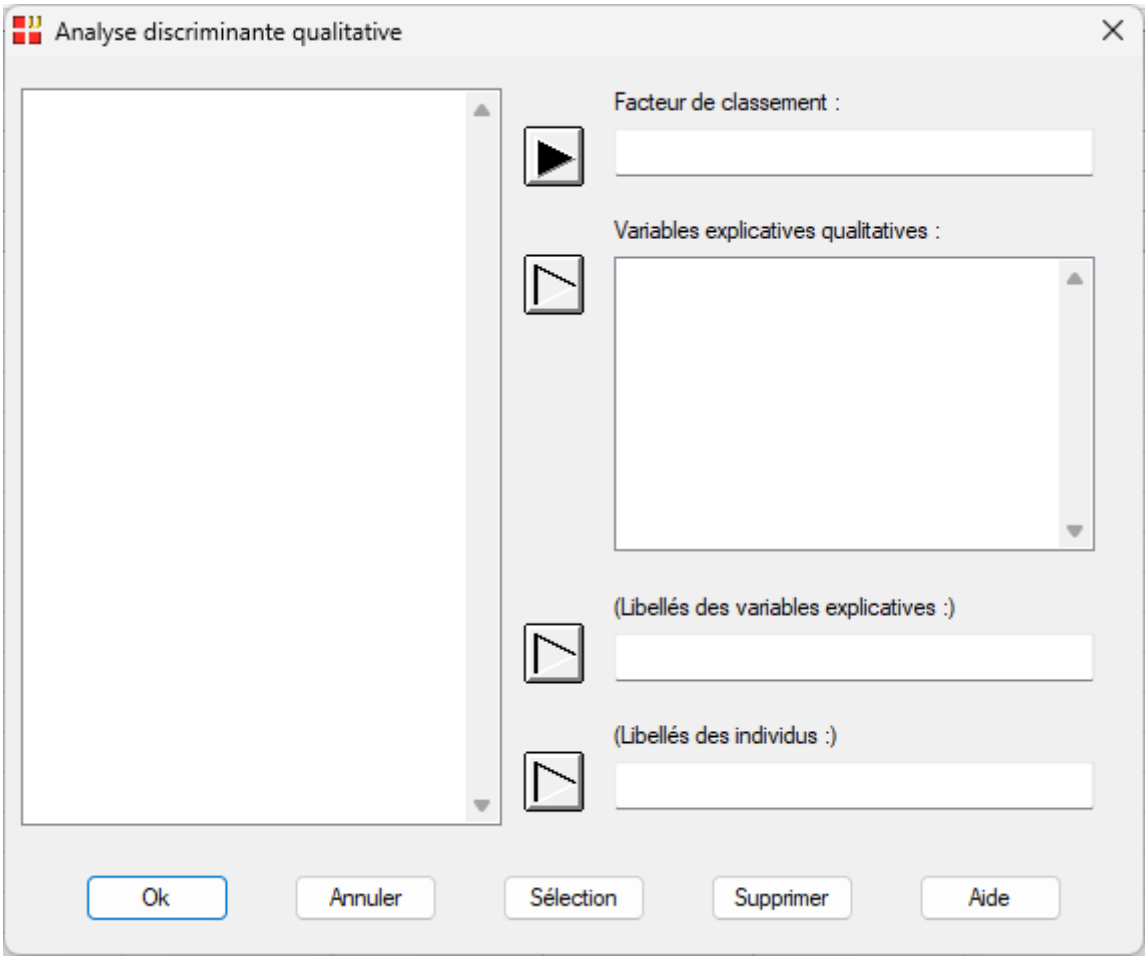
De façon plus précise, la méthode peut se décomposer en trois étapes. Supposons une population de n individus. Découpons cette population en trois sous-populations de tailles n_1, n_2 et n_3 avec $n_1 + n_2 + n_3 = n$. Les trois étapes sont :

- une étude initiale sur la population d'apprentissage de taille n_1
- une étude de validation sur la population de validation de taille n_2
- une étude prospective sur une population de prévision de taille n_3

Des tableaux résumés et détaillés des classements sont calculés. Le tracé de plans factoriels et un rapport général de synthèse sont proposés.

Entrée des données

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



Cette boîte de dialogue permet de définir le facteur de classement qualitatif, les variables explicatives qualitatives, la variable contenant les libellés de ces variables explicatives et la variable contenant les libellés des individus.

Données manquantes

Les données manquantes ne sont pas autorisées.

Exemple : Fichier CHIENS2

Nous utiliserons le fichier CHIENS2 pour illustrer cette procédure. Ce fichier contient 8 variables descriptives de 27 races de chiens :

Libellé long	*Libellé court*
Beauceron	BEA
Basset	BAS
Berger allemand	BER
Boxer	BOX
Bull-dog	BUD
Bull-mastiff	BUM
Caniche	CAN
Chihuahua	CHI
Cocker	COC
Colley	COL
Dalmatien	DAL
Doberman	DOB
Dogue allemand	DOG
Epagneul breton	EPB
Epagneul français	EPF
Fox-Hound	FOH
Fox-terrier	FOT
Grand bleu de Gascogne	GRB
Labrador	LAB
Lévrier	LEV
Mastiff	MAS
Pékinois	PEK
Pointer	POI
Saint-Bernard	STB
Setter	SET
Teckel	TEC
Terre-Neuve	TER

Les libellés courts sont dans la variable *librace*, les libellés longs dans la variable *race.*

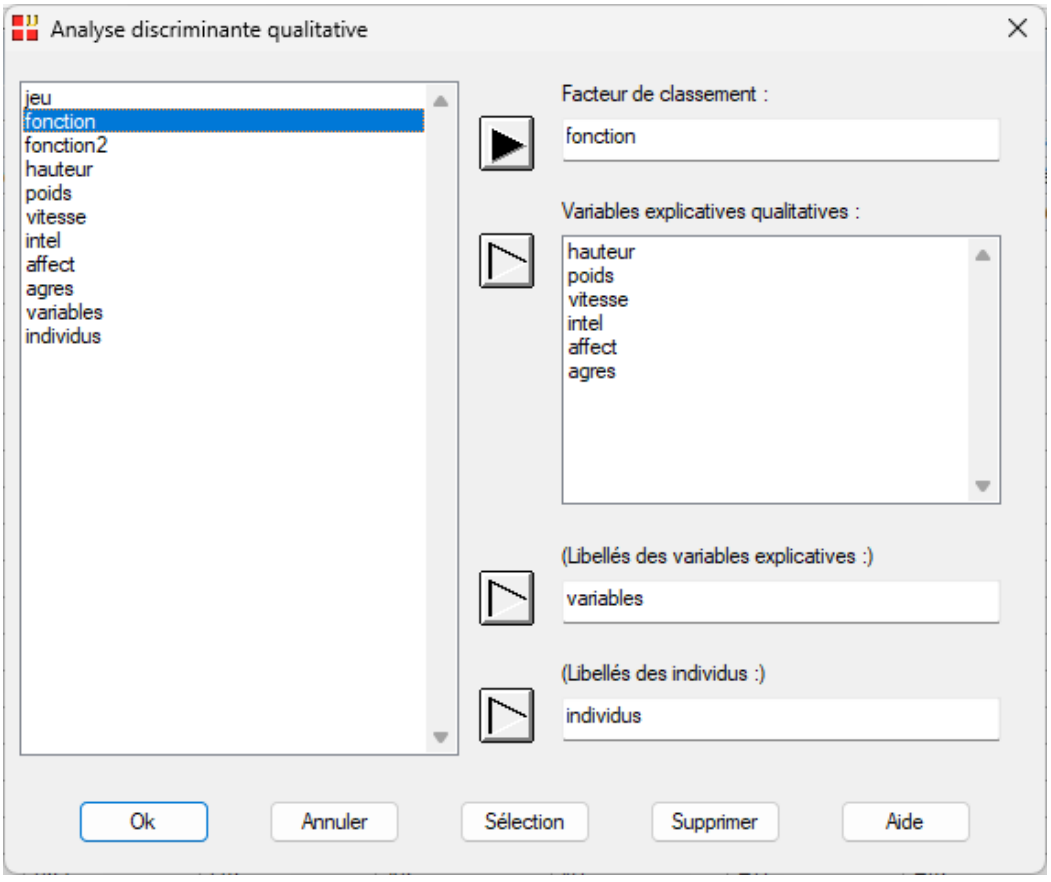
Les 8 variables descriptives sont les suivantes :

Mesures	*Variables*	*Modalités*	*Libellés des modalités*
Hauteur	hauteur	Ha1, Ha2, Ha3	libhauteur
Poids	poids	Po1, Po2, Po3	libpoids
Vitesse	vitesse	Vi1, Vi2, Vi3	libvitesse
Intelligence	intel	In1, In2, In3	libintel
Affectivité	affect	Af1, Af2	libaffect
Agressivité	agres	Ag1, Ag2	libagres
Fonction	fonction	Chasse Compagnie Utile	
Fonction2	fonction2	Autre Compagnie	

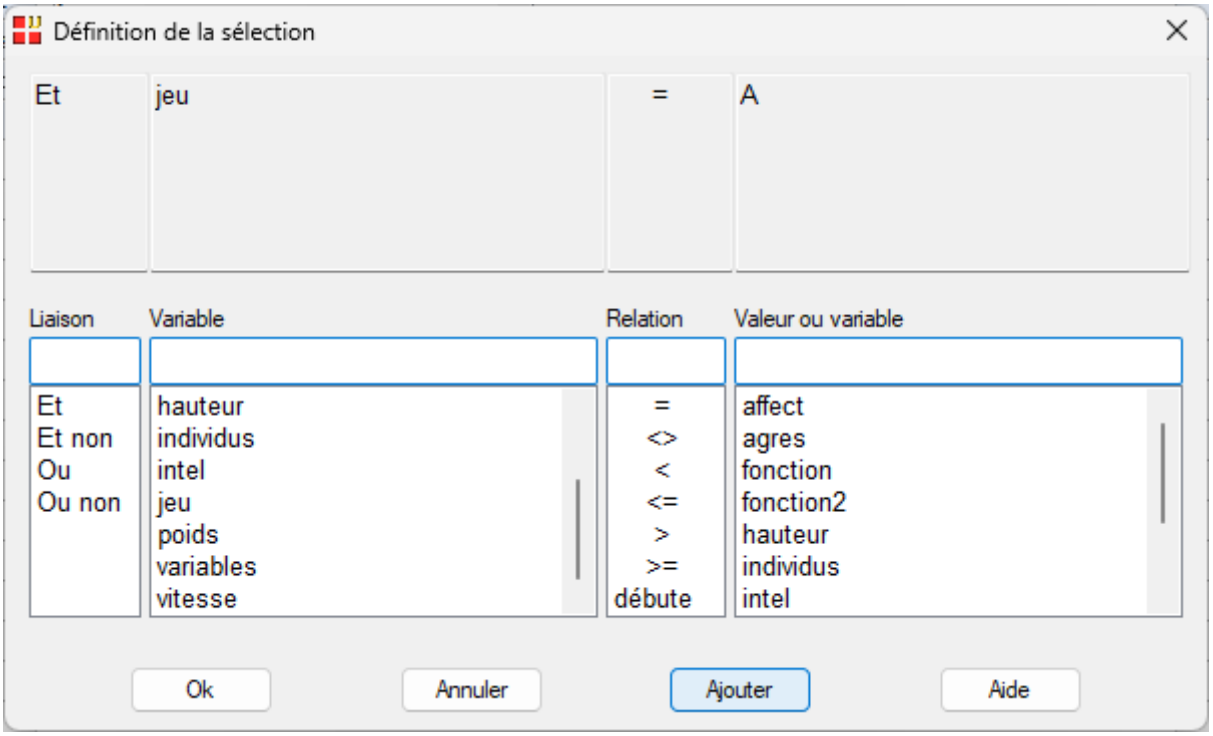
La variable ‘jeu‘ permet de distinguer les populations d’apprentissage (A), de validation (V) et de prévision (P).

Les 6 premières variables descriptives seront utilisées comme variables explicatives, la 7ième variable (*fonction*) comme variable de classement définissant nos groupes de chiens.

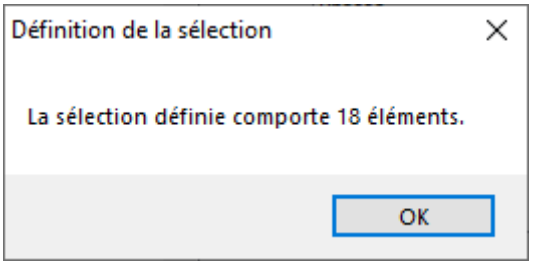
Cliquons sur l’icône ADQ dans le ruban Expliquer. La boîte de dialogue montrée ci-après s’affiche.



et sélectionnons la population d'apprentissage :



Il y a 18 chiens dans la population d'apprentissage :

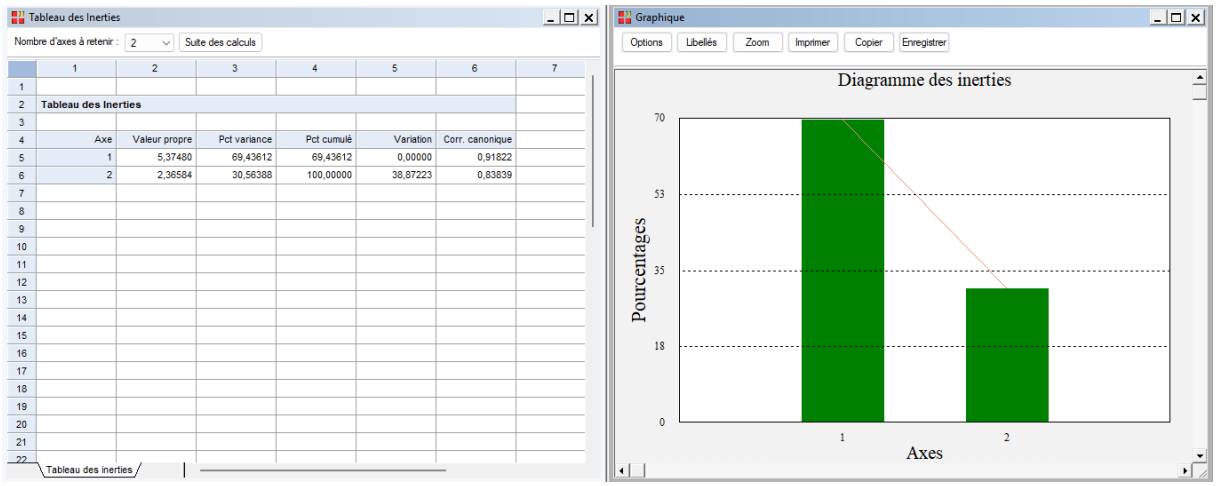


Les chiens non sélectionnés pour lesquels les valeurs du facteur de classement sont connues constituent la population de validation.

Les chiens non sélectionnés pour lesquels les valeurs du facteur de classement ne sont pas connues constituent la population de prévision.

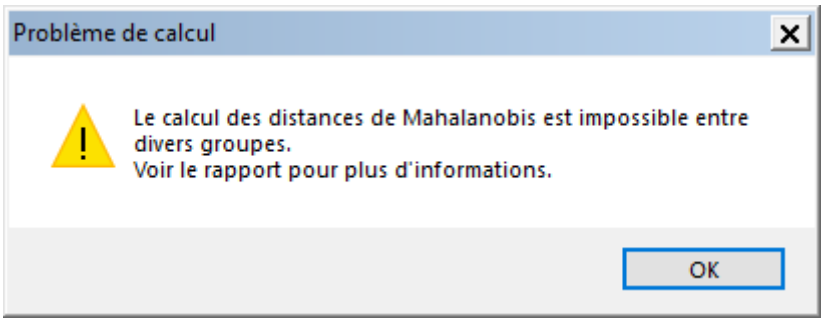
Après avoir renseigné cette boîte de dialogue, UNIWIN débute le calcul de l'Analyse Discriminante Qualitative.

Après quelques instants, un tableau précisant l'inertie expliquée par les différents vecteurs propres issus de l'analyse apparaît ainsi qu'un diagramme des pourcentages d'inertie expliquée par chacun des axes.

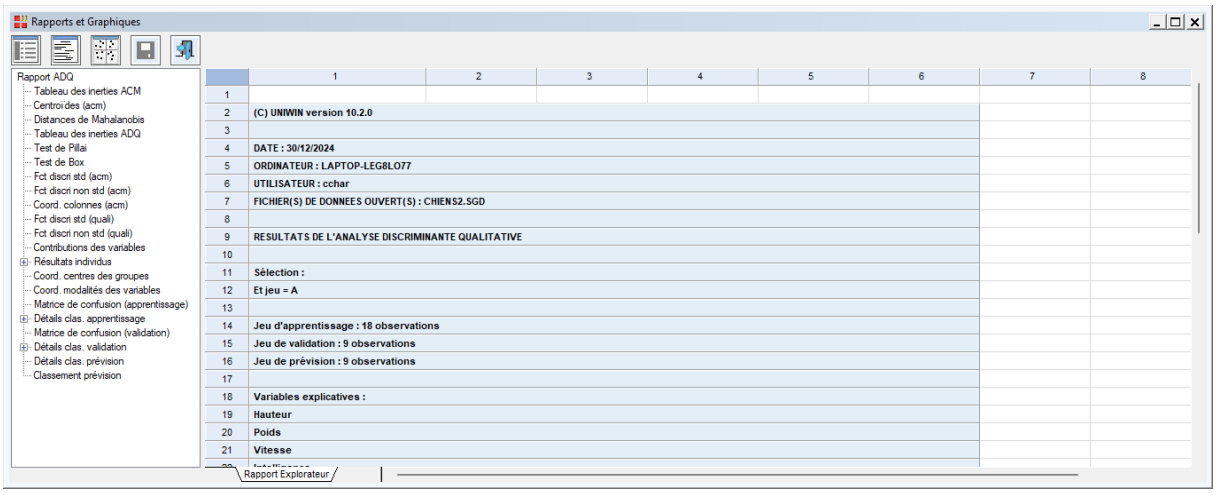


L'option 'Nombre d'axes à retenir' permet de préciser le nombre de composantes principales à extraire. Cliquons sur le bouton 'Suite des calculs'.

Le logiciel nous précise que les distances de Mahalanobis ne sont pas calculables pour tous les groupes. Plus de détails seront donnés dans le rapport.

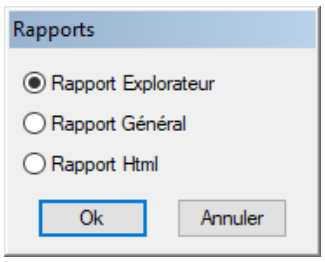


Après quelques instants, l'écran montré ci-dessous s'affiche :

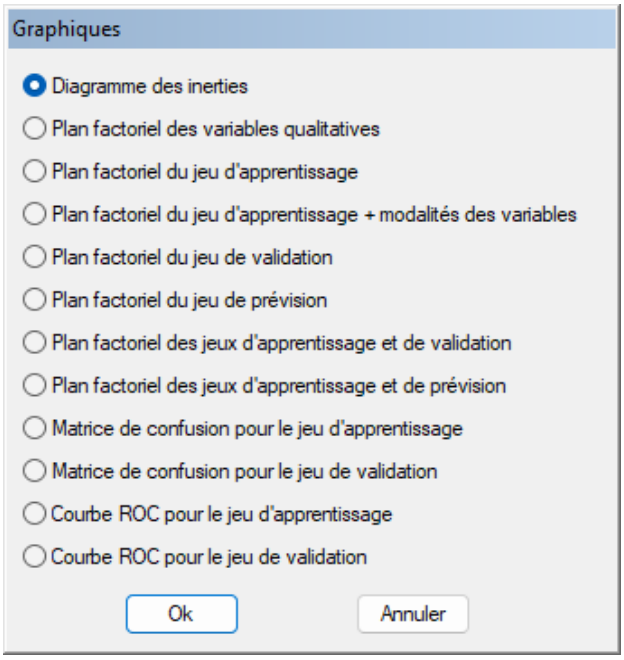


La barre d'outils 'Rapports et Graphiques' permet par l'icône 'Données'  de rappeler la boîte de dialogue d'entrée des données.

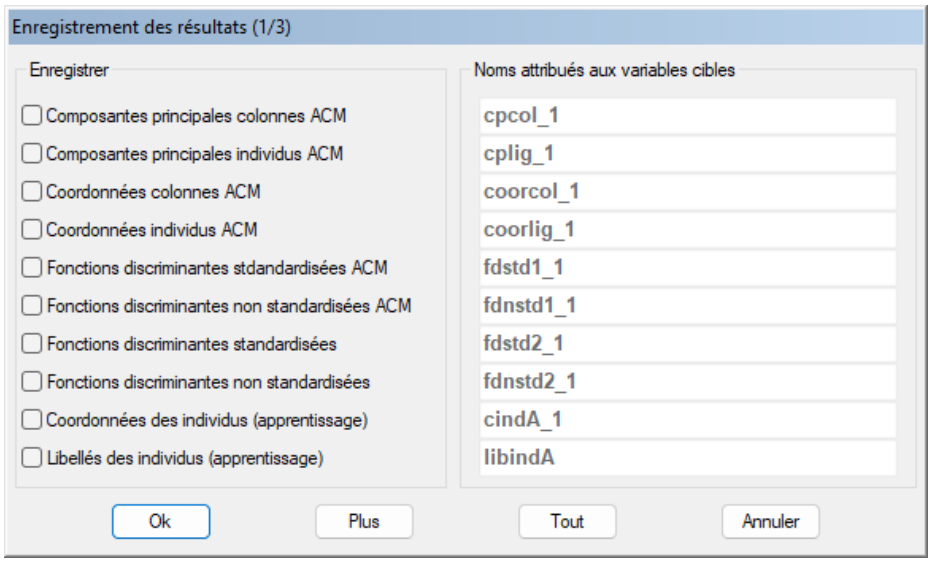
L'icône 'Rapports'  affiche la boîte de dialogue des options pour les rapports :



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.



Note : le bouton ‘Plus’ permet d’afficher la suite de la liste des variables.

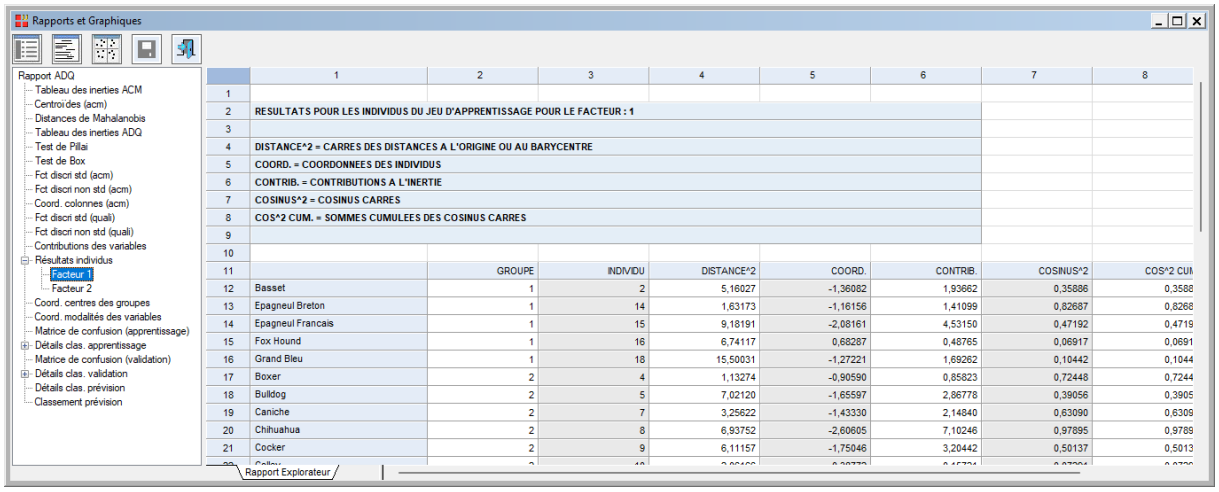
L’icône ‘Quitter’  permet de quitter l’analyse.

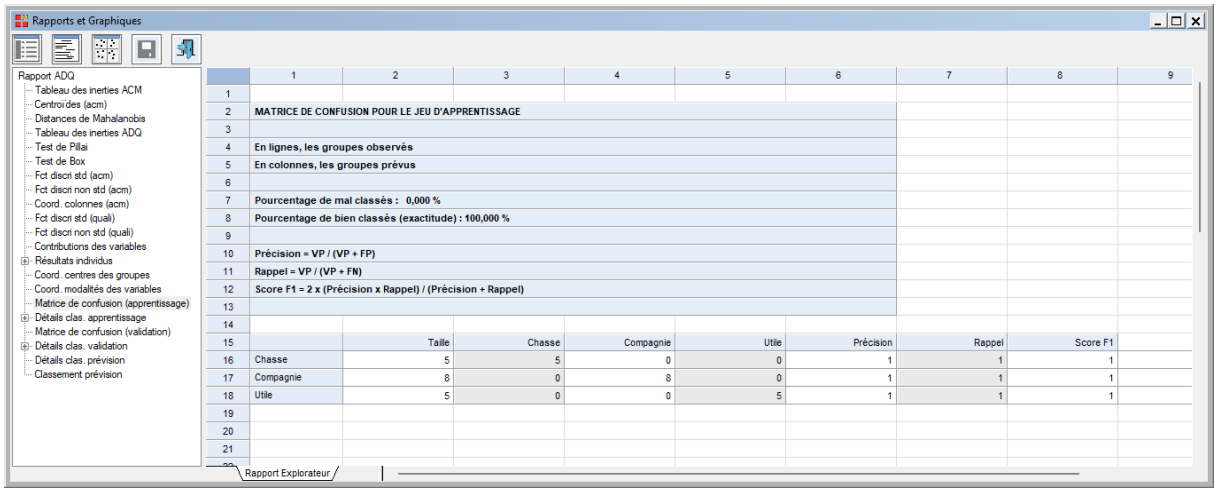
L’option Rapports

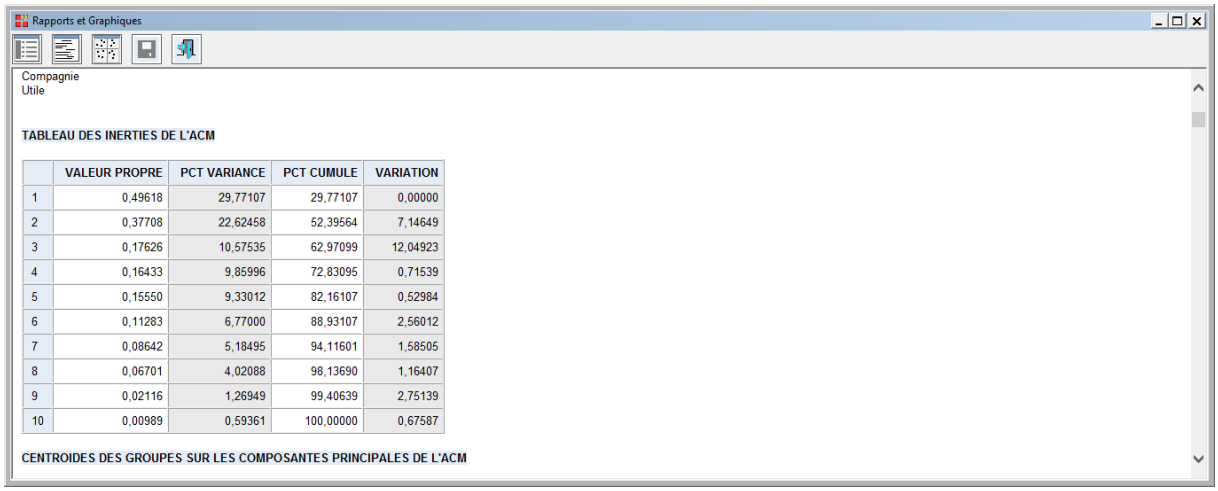
Cette option permet d’obtenir le rapport à l’écran sous la forme d’un explorateur, d’un tableur ou au format HTML.

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 trois exemples du rapport pour notre ADQ : Explorateur, Général, HTML.







	VALEUR PROPRE	PCT VARIANCE	PCT CUMULE	VARIATION
1	0,49618	29,77107	29,77107	0,00000
2	0,37708	22,62458	52,39564	7,14649
3	0,17626	10,57535	62,97099	12,04923
4	0,16433	9,85996	72,83095	0,71539
5	0,15550	9,33012	82,16107	0,52984
6	0,11283	6,77000	88,93107	2,56012
7	0,08642	5,18495	94,11601	1,58505
8	0,06701	4,02088	98,13690	1,16407
9	0,02116	1,26949	99,40639	2,75139
10	0,00989	0,59361	100,00000	0,67587

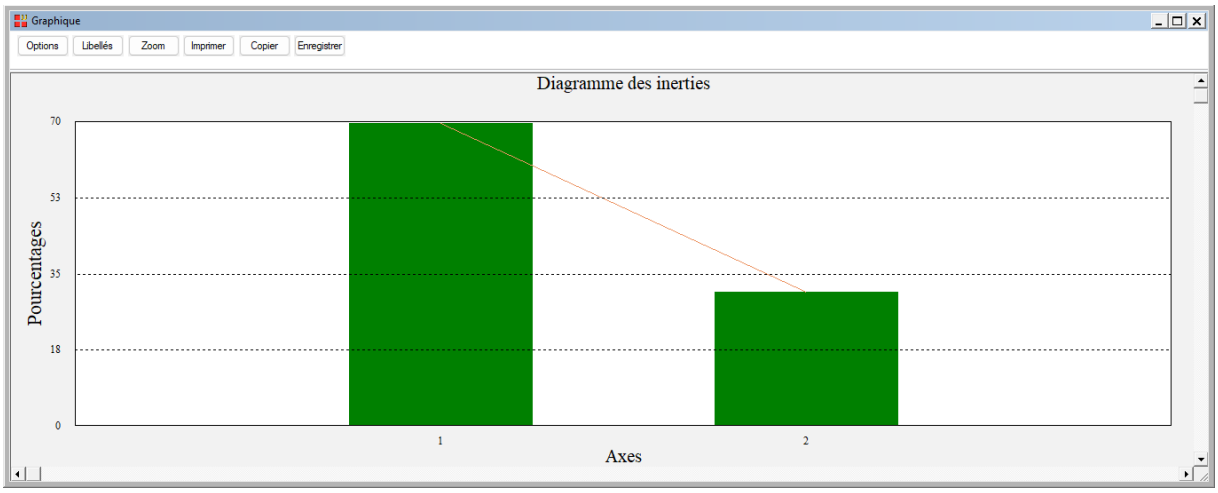
Ces rapports nous fournissent les renseignements suivants :

- Tableau des inerties de l'analyse des correspondances multiples (ACM)
- Centroïdes des groupes sur les composantes principales de l'ACM
- Distances de Mahalanobis entre les groupes, Fishers, niveaux de signification
- Tableau des inerties de l'analyse discriminante qualitative (ADQ)
- Test de Pillai
- Test de Box (égalité des matrices de covariances)
- Fonctions discriminantes standardisées de l'ACM
- Fonctions discriminantes non standardisées de l'ACM
- Coordonnées des variables dans l'ACM
- Fonctions discriminantes standardisées exprimées en fonction des variables qualitatives d'origine
- Fonctions discriminantes non standardisées exprimées en fonction des variables qualitatives d'origine
- Contributions des variables à l'inertie (carré du rapport de corrélation)
- Résultats pour les individus sur les différents facteurs (carré de la distance à l'origine, coordonnée, contribution, cosinus carré, cosinus carré cumulé)
- Coordonnées des centres des groupes
- Coordonnées des modalités des variables qualitatives
- Matrice de confusion (apprentissage)
- Détails du classement (apprentissage)
- Seuil, sensibilités et spécificités pour la population d'apprentissage
- Matrice de confusion (validation)
- Détails du classement (validation)
- Seuil, sensibilités et spécificités pour la population de validation
- Classement pour la population de prévision

L'option Graphiques

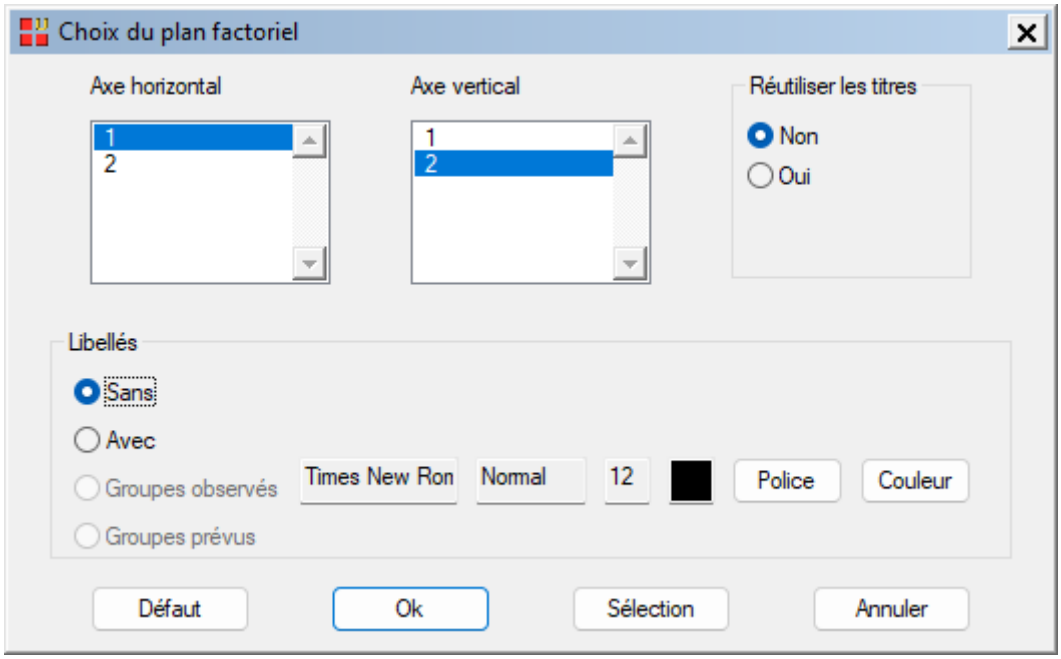
- Diagramme des inerties

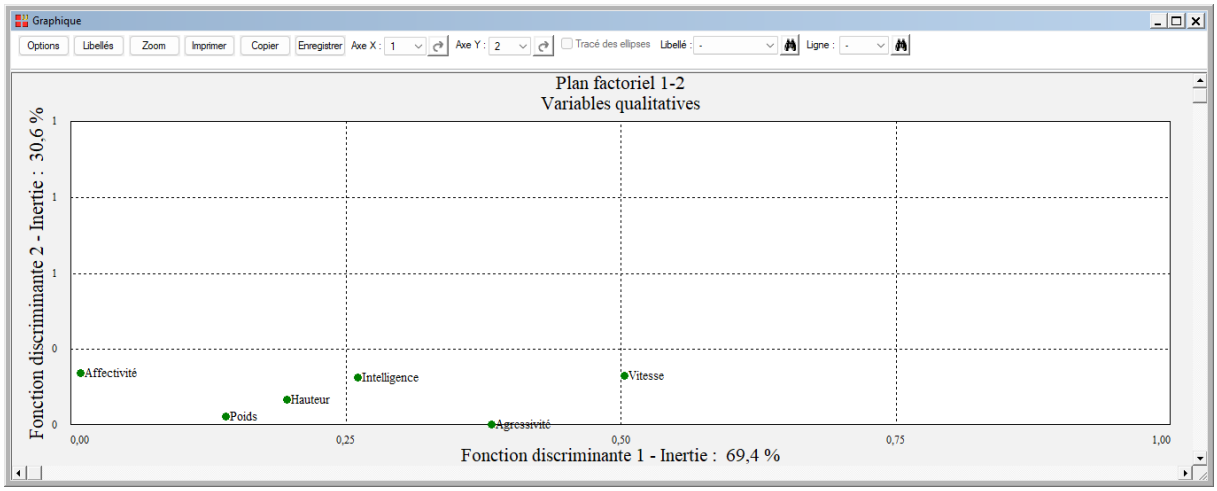
Ce graphique affiche les pourcentages d'inertie pour chacun des axes factoriels.



- Plan factoriel des variables qualitatives

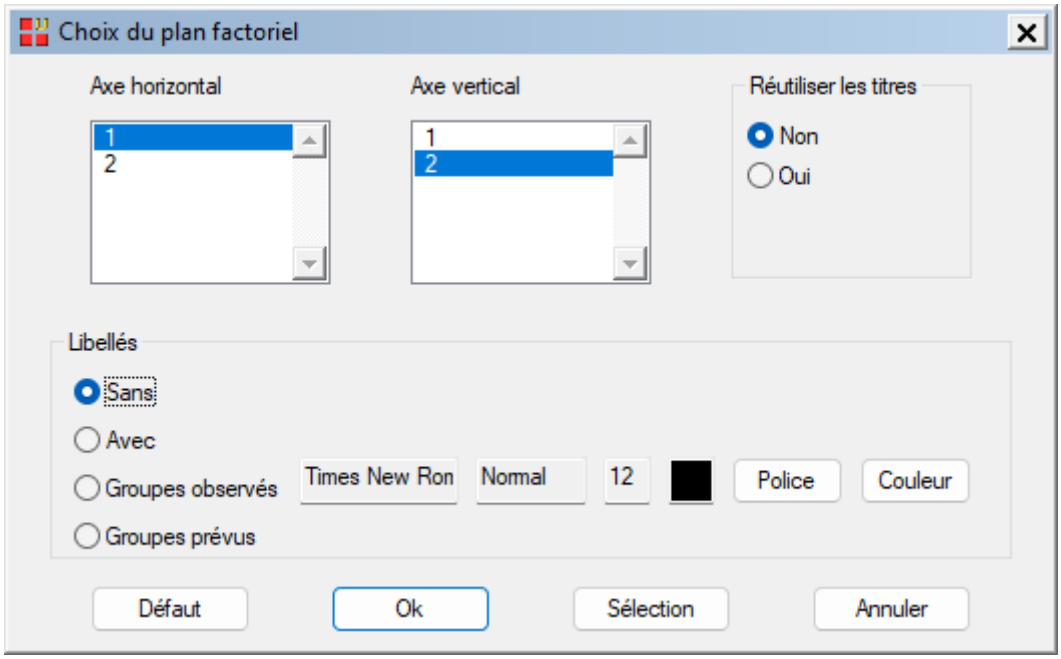
Cette option nous permet de représenter les variables qualitatives par les contributions de celles-ci à l'inertie (carré du rapport de corrélation). Une boîte de dialogue permettant de choisir le plan factoriel s'affiche. Elle permet également de préciser si l'on désire afficher les libellés des variables, de choisir la couleur et la police et d'indiquer si les titres du graphique (titre 1, titre 2), doivent être conservés pour être réutilisés ultérieurement dans d'autres graphiques créés lors de cette même session de travail.





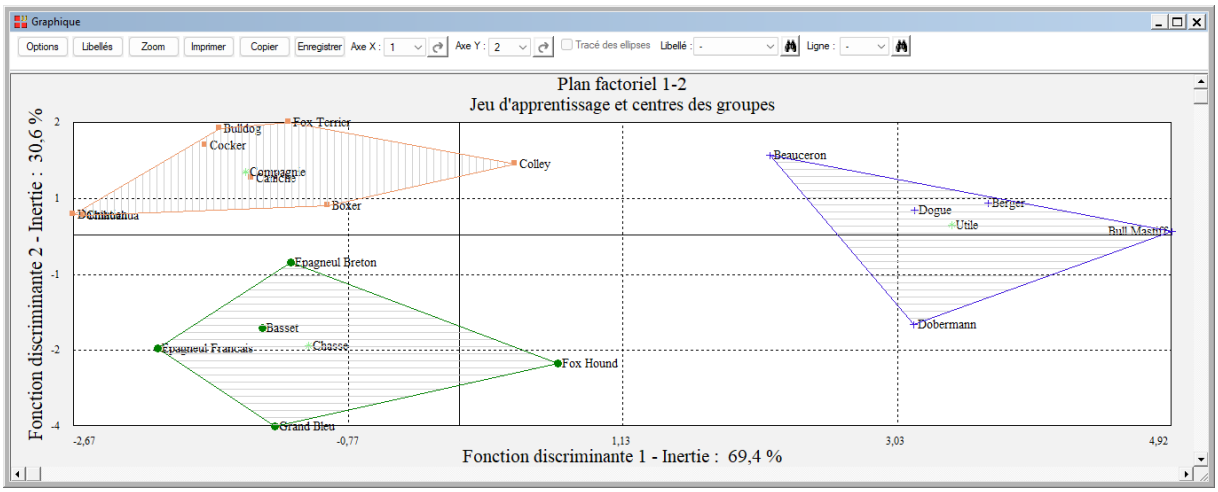
- Plans factoriels des jeux d'apprentissage, de validation et de prévision

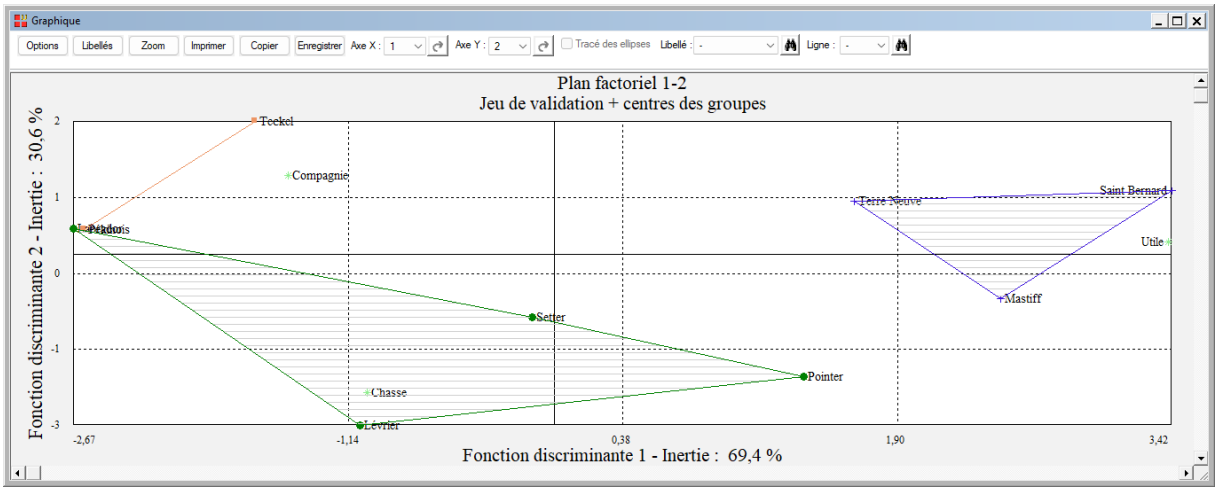
Cette option permet d'afficher des plans factoriels des individus des jeux d'apprentissage, de validation et de prévision ainsi que les centres des groupes. Une boîte de dialogue permettant de choisir le plan factoriel s'affiche.



Elle permet de préciser si l'on désire afficher ou non les libellés des individus, de préciser si ces libellés sont les codes des groupes observés ou les codes des groupes prévus, de choisir la couleur et la police pour ces libellés.

Il est également possible d'indiquer si les titres du graphique (titre 1, titre 2), doivent être conservés pour être réutilisés ultérieurement dans d'autres graphiques créés lors de cette même session de travail.

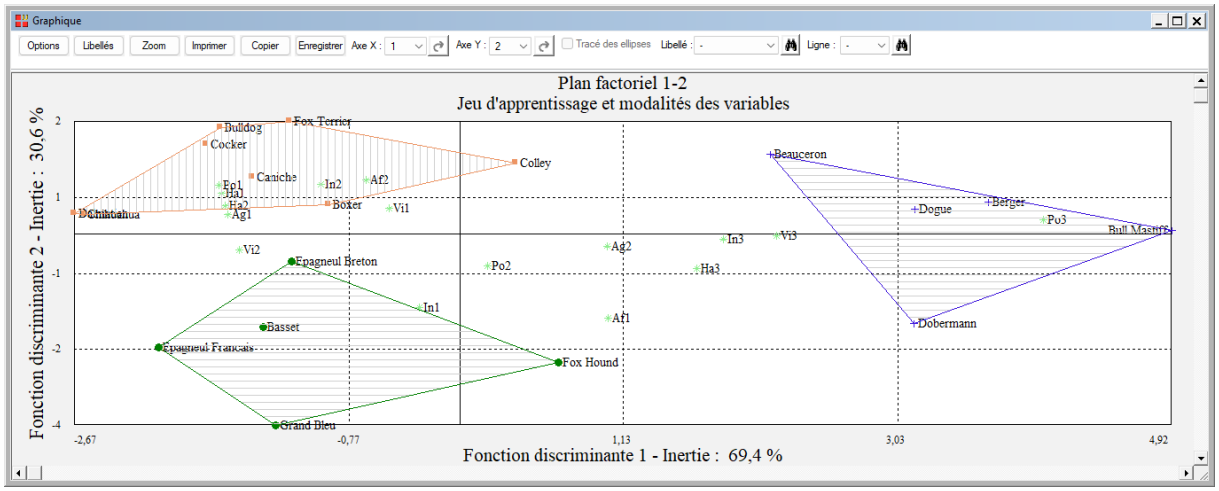




- Plan factoriel du jeu d'apprentissage et modalités des variables

Cette option permet d'afficher des plans factoriels des individus d'apprentissage et les modalités des variables qualitatives.

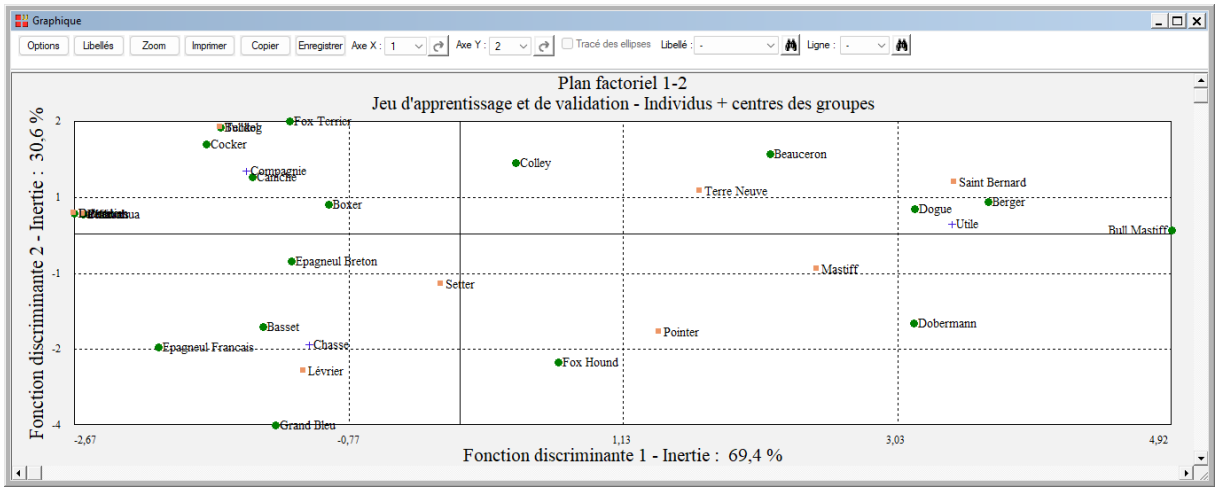
Comme pour l'option précédente, une boîte de dialogue permettant de choisir le plan factoriel et les libellés à afficher s'affiche.



- Plans factoriels des jeux d'apprentissage et de validation ou de prévision

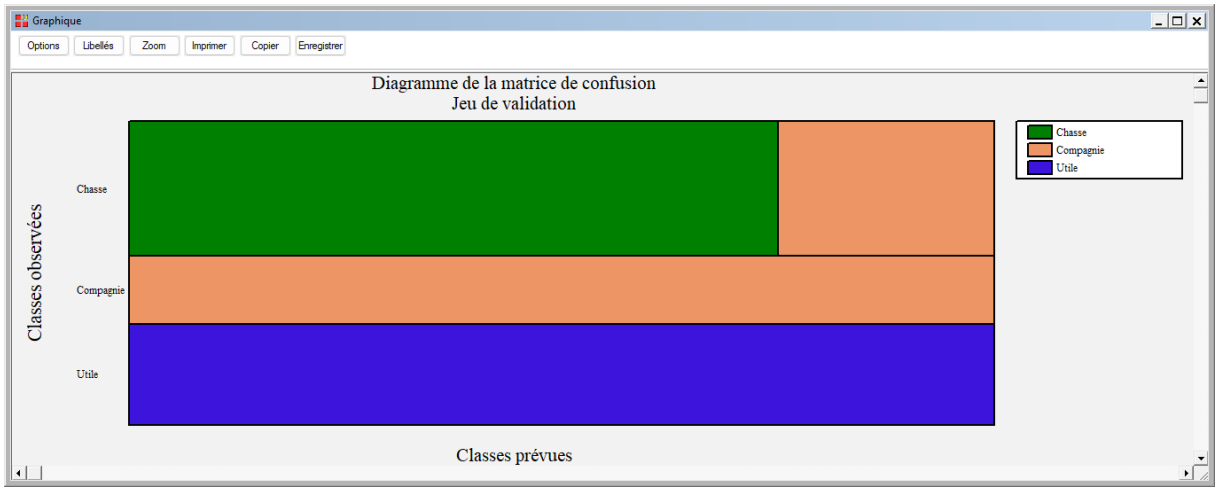
Cette option permet d'afficher des plans factoriels dans lesquels les individus d'apprentissage et de validation ou de prévision sont visualisés simultanément.

Comme pour les options précédentes, une boîte de dialogue permettant de choisir le plan factoriel et les libellés à afficher s'affiche.



- Graphiques des matrices de confusion

Ces graphiques affichent les matrices de confusion des jeux d'apprentissage et de validation sous la forme de diagrammes en mosaïque.



- Courbes ROC

Ces deux options ne sont pas actives dans cet exemple car le facteur de classement possède plus de 2 classes.

Exemple : Fichier ASSURANCE

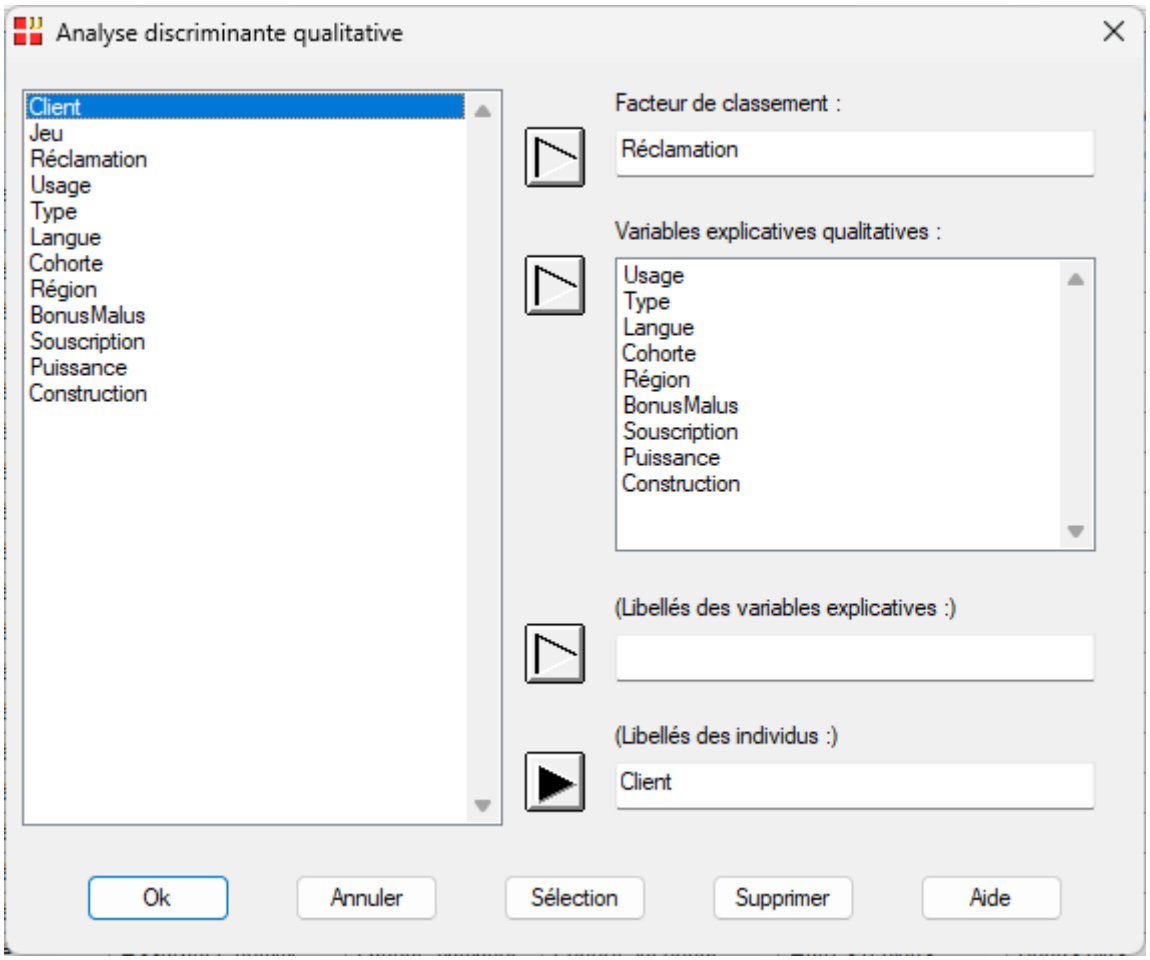
Le fichier ASSURANCE contient des informations collectées en 1992 concernant 1106 contrats d'assurance d'automobilistes belges.

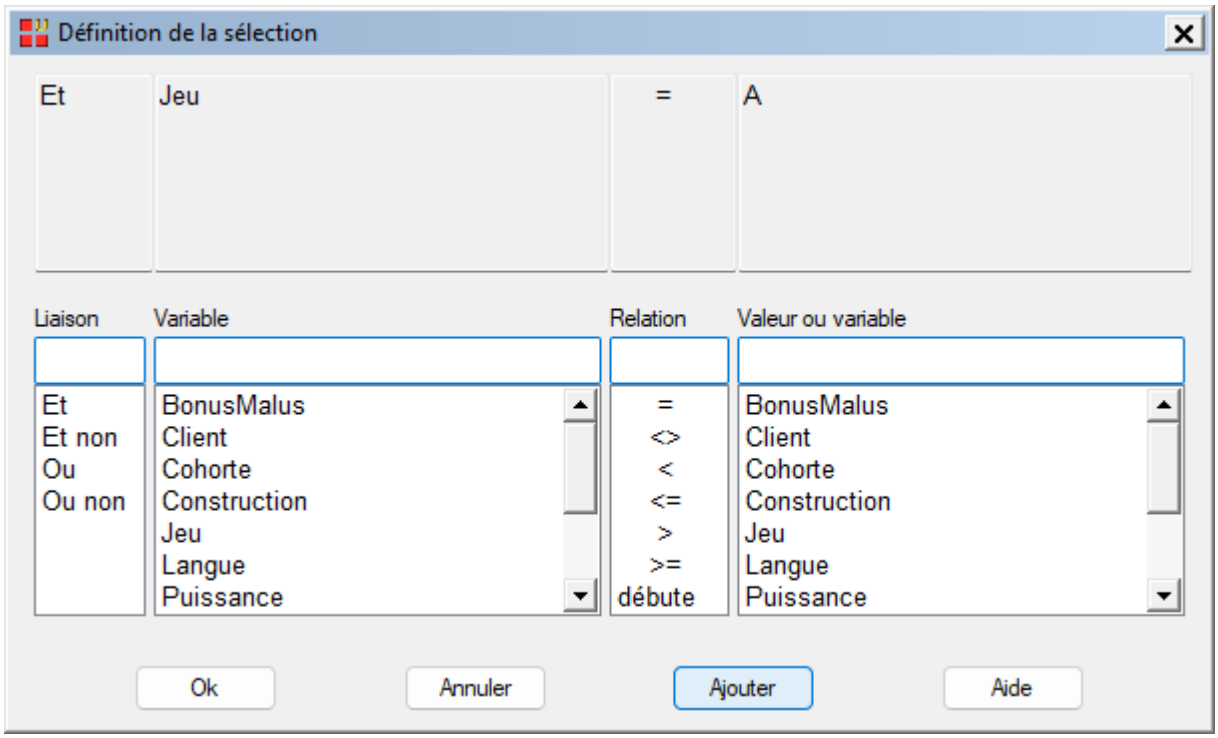
• Réclamation	Valide, Non valide
• Usage du véhicule	Usage privé, Usage professionnel
• Type d'assurance	Usage entreprise, Usage femme Usage homme
• Langue parlée	Langue flamande, Langue française
• Cohorte de naissance	Cohorte 1890 à 1949, Cohorte 1950 à 1973 Cohorte inconnue
• Région géographique	Autres régions, Bruxelles
• Niveau de bonus-malus	Bonus moins, Bonus plus
• Année de souscription	Souscription 86 et au-delà Souscription avant 86
• Puissance du véhicule	Puissance inférieure à 40 Puissance supérieure ou égale à 40
• Année de construction du véhicule	Construction 33-89, Construction 90-91

Référence : Saporta G., Niang N. (2006) Correspondence Analysis and Classification. In *Multiple Correspondence Analysis and Related Methods*, M. Greenacre and J. Blasius, Eds., pp 371-392. Chapman & Hall/CRC, Boca Raton, Florida, USA.

La variable 'Jeu' permet de distinguer les populations d'apprentissage (A) et de validation (V).

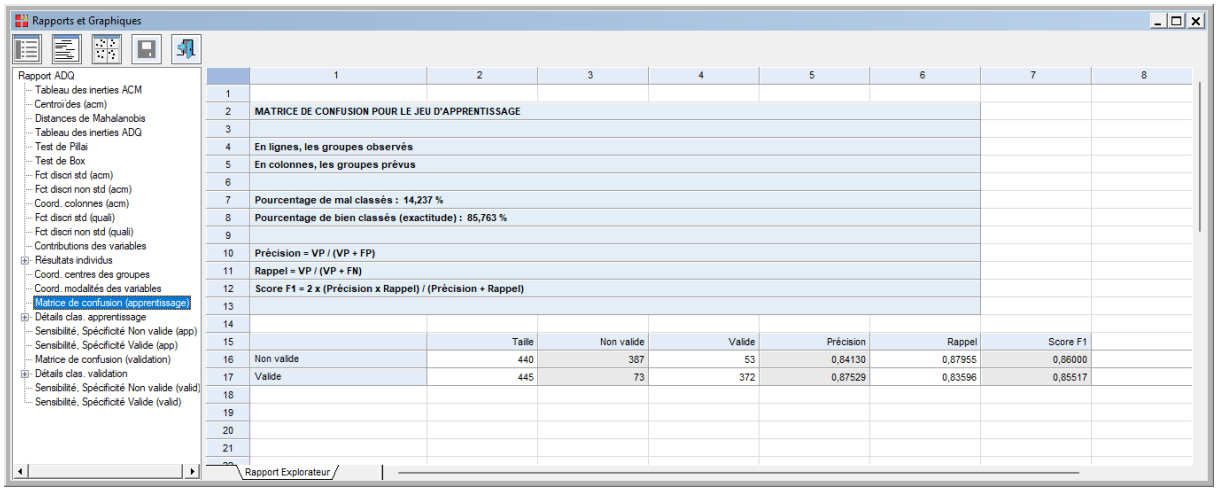
Cliquons sur l'icône ADQ dans le ruban Expliquer et renseignons la boîte de dialogue comme montré ci-après puis sélectionnons la population d'apprentissage (Jeu = A).

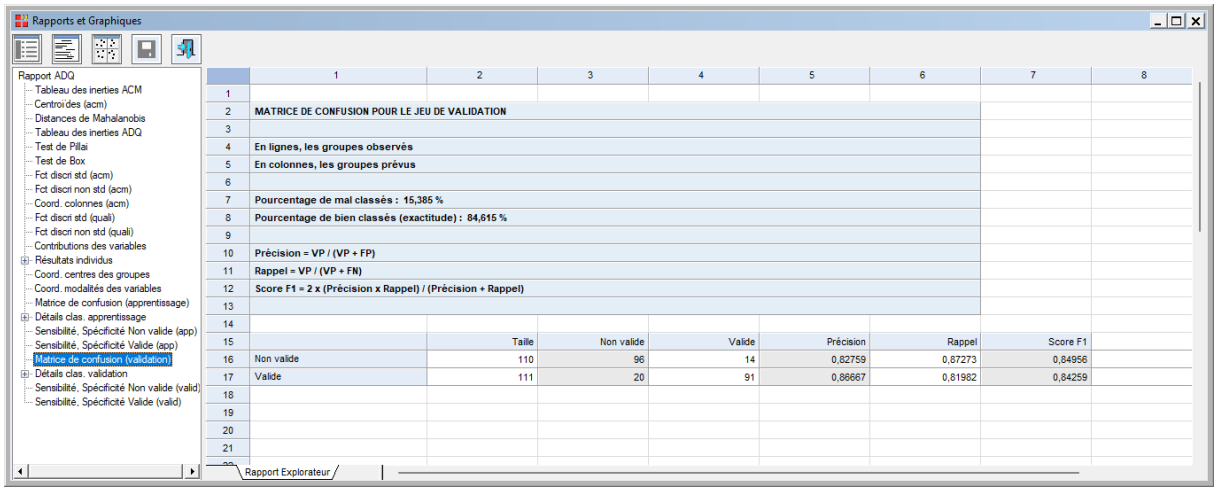

Analyse discriminante qualitative
Client
Jeu
Réclamation
Usage
Type
Langue
Cohorte
Région
BonusMalus
Souscription
Puissance
Construction
Facteur de classement :
Réclamation
Variables explicatives qualitatives :
Usage
Type
Langue
Cohorte
Région
BonusMalus
Souscription
Puissance
Construction
(Libellés des variables explicatives :)
(Libellés des individus :)
Client
Ok
Annuler
Sélection
Supprimer
Aide


Définition de la sélection
Et
Jeu
=
A
Liaison
Variable
Relation
Valeur ou variable
Et
Et non
Ou
Ou non
BonusMalus
Client
Cohorte
Construction
Jeu
Langue
Puissance
=
<>
<
<=
>
>=
débute
BonusMalus
Client
Cohorte
Construction
Jeu
Langue
Puissance
Ok
Annuler
Ajouter
Aide

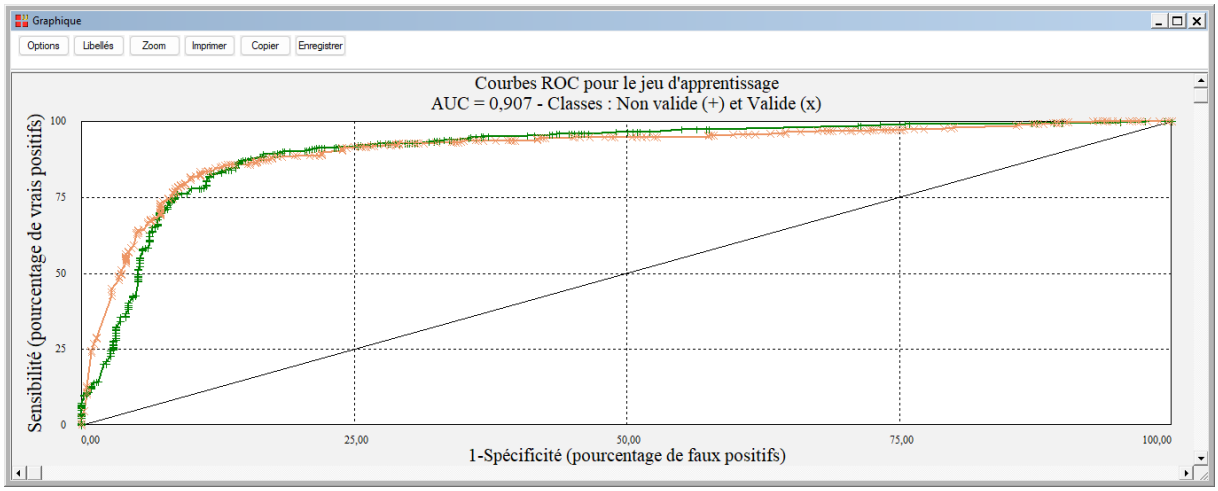
Cliquons sur le bouton Ok pour exécuter le traitement de l'analyse.

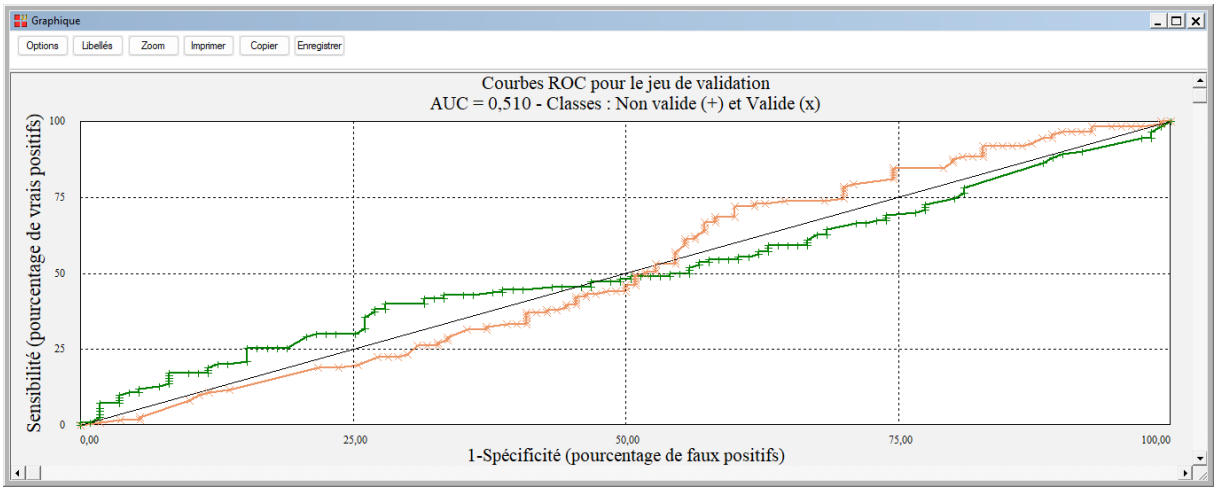
Visualisons les résultats des classements des populations d'apprentissage et de validation.





Le rapport affiche les sensibilités et spécificités qui permettent les tracés des courbes ROC des populations d'apprentissage et de validation.





Calculs de la matrice de confusion et des indicateurs

Dans le cas de deux classes A et B, nous avons le tableau suivant :

	Prévu A	**Prévu B**	**Total**	**% correct**
Observé A	VP	FN	VP + FN	$\frac{100 * \mathrm{VP}}{(\mathrm{VP}+\mathrm{FN})}$
Observé B	FP	VN	FP + VN	$\frac{100 * \mathrm{VN}}{(\mathrm{VN}+\mathrm{FP})}$
Total	VP + FP	FN + VN	VP + FP + VN + FN	
% correct	$\frac{100 * \mathrm{VP}}{(\mathrm{VP}+\mathrm{FP})}$	$\frac{100 * \mathrm{VN}}{(\mathrm{FN}+\mathrm{VN})}$		$\frac{100 * (\mathrm{VP}+\mathrm{VN})}{(\mathrm{VP}+\mathrm{VN}+\mathrm{FP}+\mathrm{FN})}$
				% total correctement prévu

Dans le cas multi-classes (plus de 2 classes), chaque classe est étudiée par rapport une classe virtuelle réunissant l'ensemble des autres classes.

Définition des indicateurs :

- la sensibilité VP / (VP+FN)
- la spécificité VN / (VN+FP)
- l'exactitude (VP+VN) / (VP+VN+FP+FN)
- la précision VP / (VP+FP)
- le rappel VP / (VP+FN)
- le score F1 2 x (précision x rappel) / (précision + rappel)

La sensibilité (ou rappel) indique la capacité du modèle à prévoir les vrais positifs.

La spécificité (ou taux de vrais négatifs) permet de mesurer la capacité du modèle à prévoir les vrais négatifs.

L'exactitude mesure le pourcentage de prévisions correctes par rapport à toutes les prévisions positives et négatives. Elle varie entre 0 et 1 et est sensible aux données déséquilibrées. Plus elle est proche de 1, meilleure est la prévision globale.

Le rappel (ou sensibilité ou taux de vrais positifs) varie entre 0 et 1 et n'est pas sensible aux données déséquilibrées. Un rappel égal à 1 indique une prévision parfaite des positifs.

La précision mesure le pourcentage de prévisions positives correctes. Elle varie entre 0 et 1 et n'est pas sensible aux données déséquilibrées. Une précision égale à 1 indique que tous les positifs sont prédits positifs.

Le score F1 combine la précision et le rappel en utilisant les moyennes harmoniques. Il varie entre 0 et 1. Maximiser ce score revient à maximiser la précision et le rappel. Il n'est pas sensible aux données déséquilibrées.

Les variables internes créées par la procédure

Voici la liste des variables internes créées par la procédure. Ces variables peuvent notamment être utilisées avec l'option 'Sélection'. A noter que certaines des variables mentionnées ci-dessous peuvent ne pas apparaître, en fonction des options choisies.

Variable	*Contenu*
cpcol	Composantes principales colonnes (ACM)
cplig	Composantes principales individus (ACM)
coorcol	Coordonnées des colonnes (ACM)
coorlig	Coordonnées des individus (ACM)
fdstd1	Fonctions discriminantes standardisées (ACM)
fdnstd1	Fonctions discriminantes non standardisées (ACM)
fdstd2	Fonctions discriminantes standardisées (ADQ)
fdnstd2	Fonctions discriminantes non standardisées (ADQ)
cindA	Coordonnées des individus (apprentissage
libindA	Libellés des individus (apprentissage)
clindA	Classes d'origine des individus (apprentissage)
distind	Distances carrées des individus à l'origine (apprentissage)
cosind	Cosinus carrés des individus (apprentissage
conind	Contributions des individus (apprentissage)
cindV	Coordonnées des individus (validation)
libindV	Libellés des individus (validation)
clindV	Classes d'origine des individus (validation)

seuilA	Seuils (apprentissage)
specificiteA	Spécificités (apprentissage)
sensibiliteA	Sensibilités (apprentissage)
aireA	Aire sous les courbes ROC (apprentissage)
seuilV	Seuils (validation)
specificiteV	Spécificités (validation)
sensibiliteV	Sensibilités (validation)
aireV	Aires sous les courbes ROC (validation)
classeA	Classement apprentissage
classeV	Classement validation
classeP	Classement prévision
libindP	Libellés des individus (prévision)