

UNIWIN VERSION 10.4.0

ANALYSE DISCRIMINANTE BAYESIENNE

Révision : 15/09/2025

| | |
|---|----|
| Définition..... | 1 |
| Entrée des données | 2 |
| Données manquantes | 3 |
| Exemple 1 : Fichier IRIS3..... | 3 |
| L'option Rapports | 8 |
| L'option Graphiques | 9 |
| Exemple 2 : Fichier INFARCT2 | 14 |
| L'option Rapports | 17 |
| L'option Graphiques | 18 |
| Exemple 3 : Fichier BORDEAUX..... | 22 |
| Exemple 4 : Fichier TITANIC..... | 25 |
| Calculs de la matrice de confusion et des indicateurs | 27 |
| Les variables internes créées par la procédure | 28 |

Définition

L'Analyse Discriminante Bayésienne (ADB) permet de construire à partir d'un ensemble de variables quantitatives et d'une variable qualitative découpant la population en plusieurs groupes (2 ou plus), des fonctions discriminantes qui définissent une règle de décision optimale à partir de laquelle on peut affecter des individus de validation et de prévision aux différents groupes.

Cette technique suppose que l'on connaisse a priori les probabilités d'appartenance aux différents groupes et que les données suivent une loi multi-normale.

La méthode proposée permet de traiter les cas linéaire (égalité des matrices de variances) et quadratique (non-égalité des matrices de variances).

L'entrée des probabilités a priori est proposée. Par défaut, le système utilise les probabilités issues des fréquences des groupes dans les données entrées.

En fonction des données et des paramètres définis par l'utilisateur, l'analyse ADB 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 ADB dans le ruban Expliquer. La boîte de dialogue montrée ci-après s'affiche :

The dialog box is titled "Analyse discriminante bayésienne". It contains a large empty list box on the left. On the right side, there are several input fields and controls:

- Facteur de classement :** A text input field with a right-pointing arrow icon to its left.
- Variables explicatives quantitatives :** A list box with a right-pointing arrow icon to its left.
- (Libellés des variables explicatives :) :** A text input field with a right-pointing arrow icon to its left.
- (Libellés des individus :) :** A text input field with a right-pointing arrow icon to its left.
- (Probabilités initiales :) :** A text input field with a right-pointing arrow icon to its left.
- Centrage et réduction :** A section with two radio buttons: "Oui" (unselected) and "Non" (selected).

At the bottom of the dialog box, there are five buttons: "Ok", "Annuler", "Sélection", "Supprimer", and "Aide".

Cette boîte de dialogue permet de préciser la variable qualitative définissant facteur de classement, les variables explicatives quantitatives, la variable contenant les libellés des variables explicatives et la variable contenant les libellés des individus.

Elle permet également de définir la variable contenant les probabilités a priori si on ne désire pas qu'UNIWIN les calcule à partir des fréquences des groupes dans les données.

Enfin, l'option de centrage et réduction des données est proposée. Cette option est utile si le classement de certains individus n'est pas possible suite à un dépassement de capacité lors des calculs. Il se peut que cela soit lié aux grandeurs des données et les centrer-réduire peut résoudre ce problème de calcul.

Données manquantes

Les données manquantes ne sont pas autorisées pour le facteur de classement. Elles sont autorisées pour les variables quantitatives.

Exemple 1 : Fichier IRIS3

Nous utiliserons le fichier IRIS3 pour illustrer cette procédure. Ce fichier contient pour 150 iris de trois espèces différentes les mesures des quatre caractéristiques suivantes exprimées en millimètres : longueur du sépale, largeur du sépale, longueur du pétale et largeur du pétale

Les trois espèces sont : Iris Setosa, Iris Versicolor et Iris Virginica

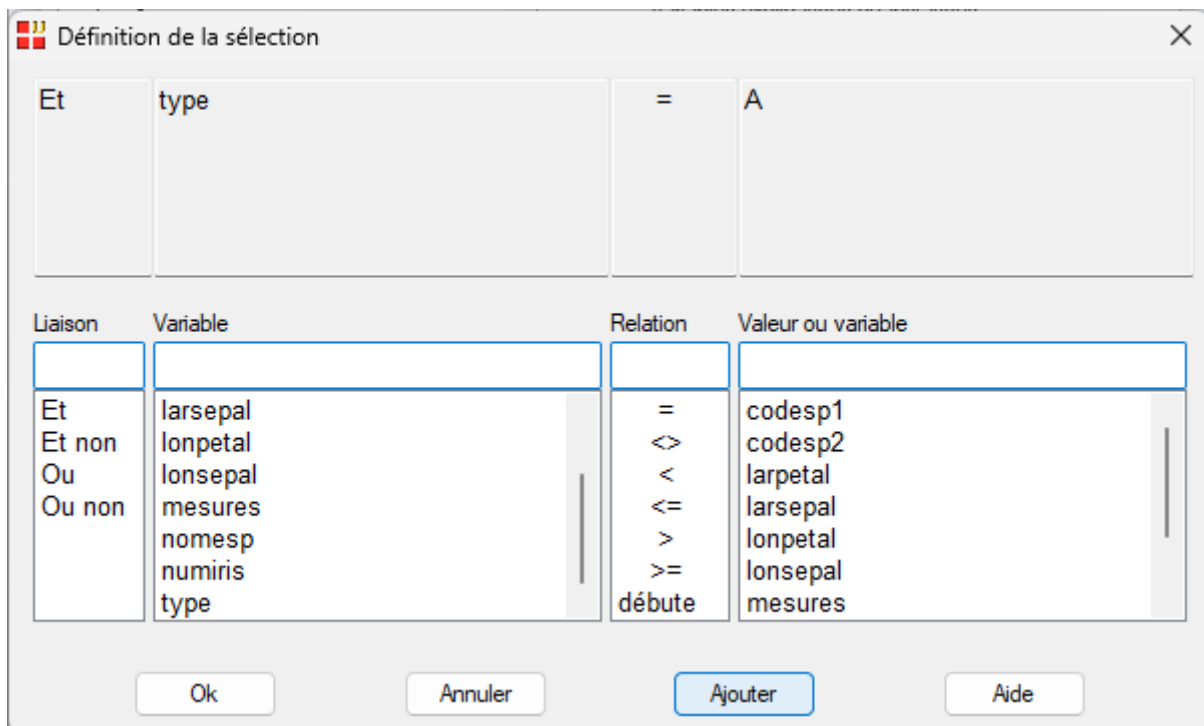
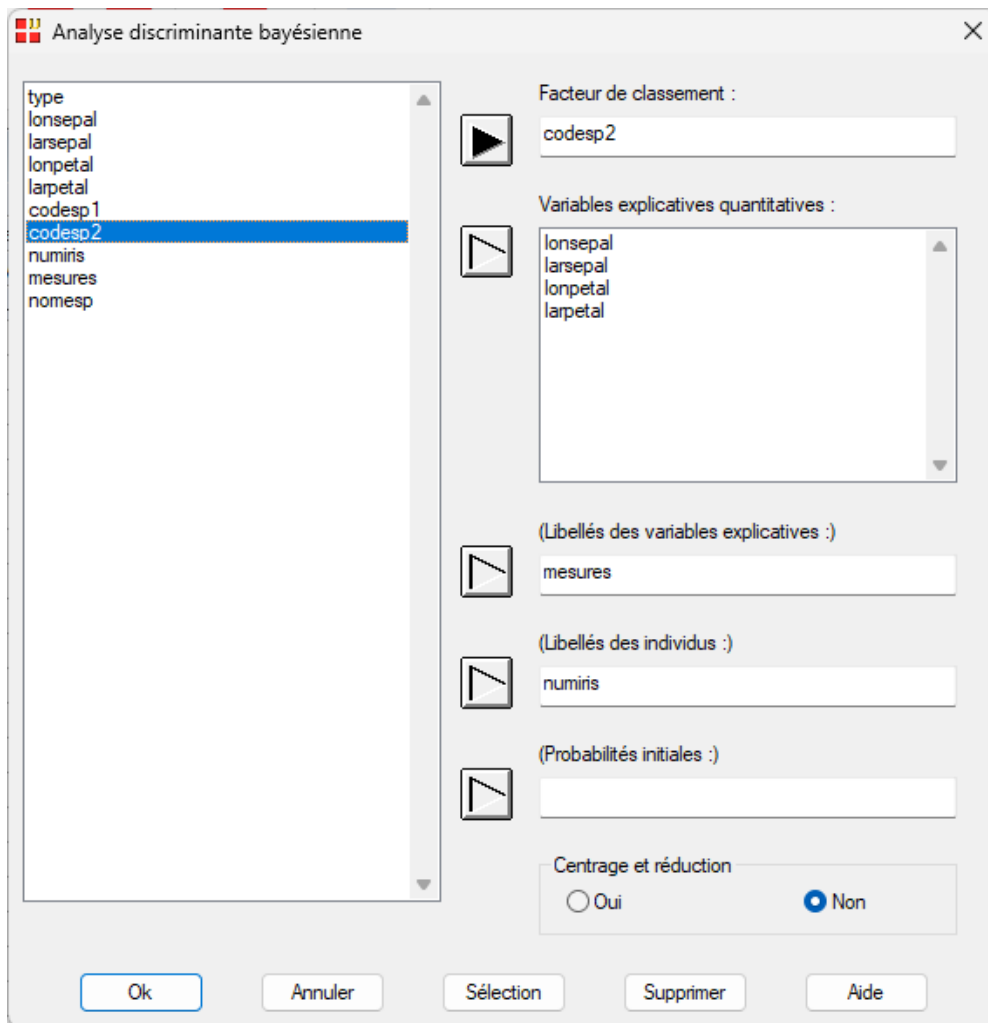
Cliquons sur l'icône ADB dans le ruban Expliquer.

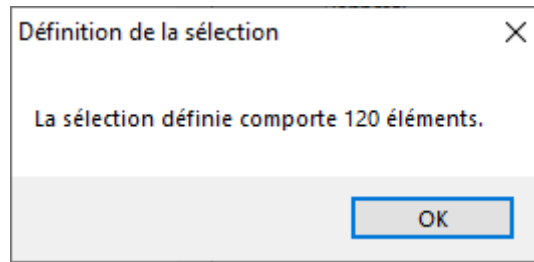
La boîte de dialogue montrée ci-après s'affiche.

Choisissons les variables *lonsepal* à *larpetal* comme variables quantitatives, la variable *codesp2* comme facteur de classement, la variable *mesures* comme variable contenant les libellés des variables quantitatives et la variable *numiris* comme variable contenant les libellés des individus.

Réalisons une analyse non centrée-réduite et laissons à UNIWIN le soin de calculer les probabilités initiales.

Cliquons sur le bouton Sélection pour définir la population d'apprentissage.





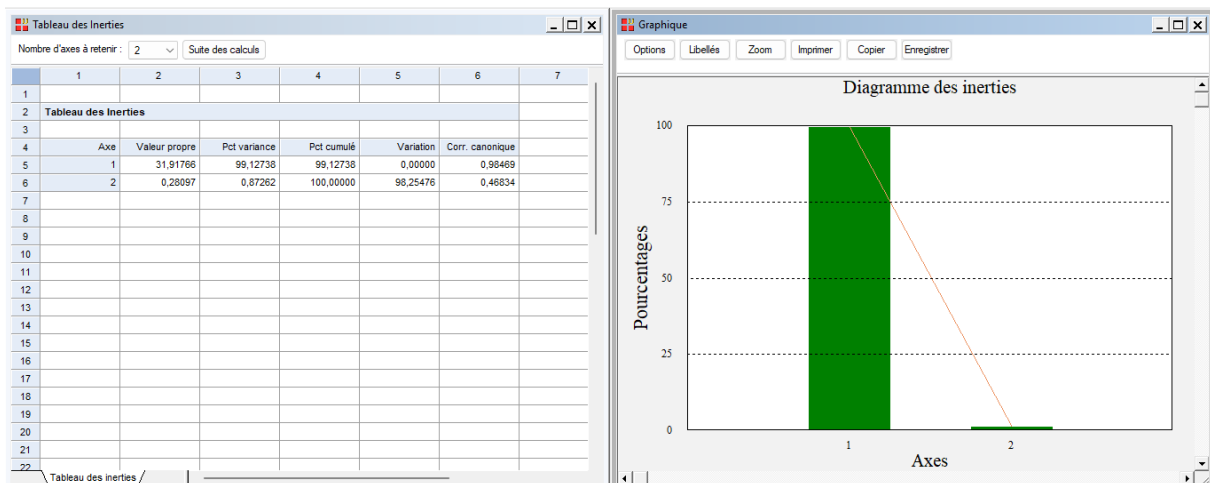
Cette sélection comporte 120 individus qui constituent la population d'apprentissage.

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

Les individus 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 Bayésienne.

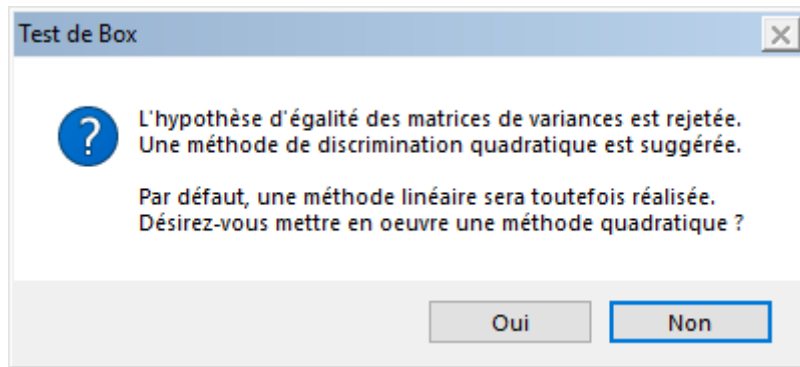
Après quelques instants, un tableau précisant l'inertie expliquée par les différents vecteurs propres issus de l'analyse apparaît ainsi qu'un diagramme des pourcentages d'inertie expliquée par chacun des axes.



L'option 'Nombre d'axes à retenir' permet de préciser le nombre de composantes principales à extraire.

Cliquons sur le bouton 'Suite des calculs'.

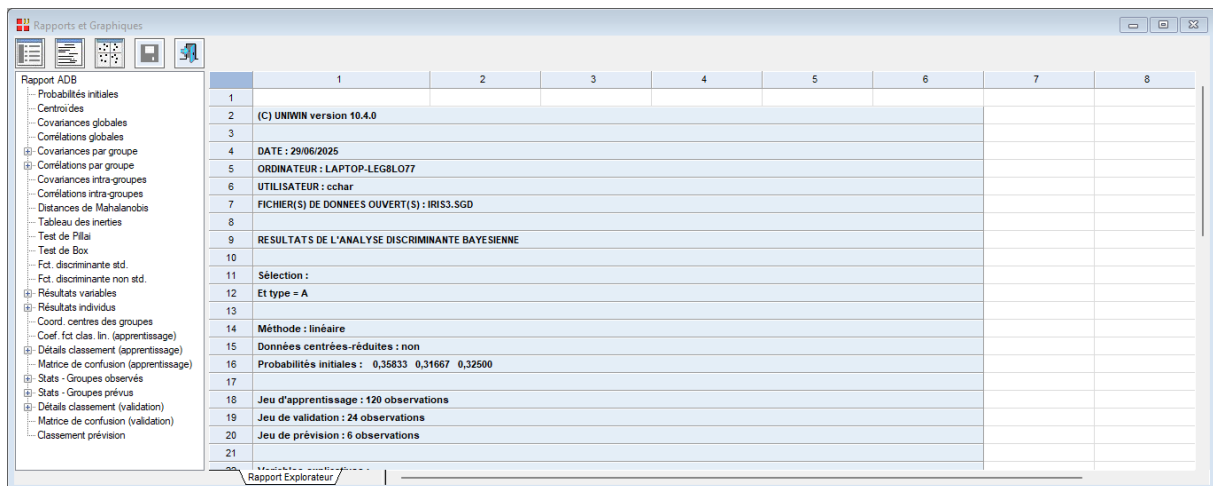
Après quelques instants, UNIWIN affiche une boîte de dialogue vous indiquant si l'hypothèse d'égalité des matrices de variances est vérifiée ou non.





Dans notre exemple, l'hypothèse doit être rejetée et donc une analyse discriminante quadratique est suggérée.

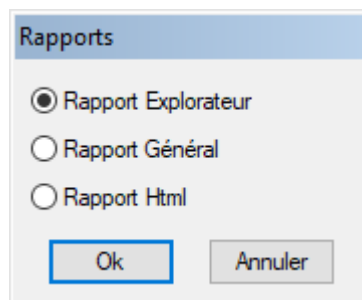
Choisissons cependant de mettre en œuvre une analyse linéaire.


Après quelques instants, l'écran suivant s'affiche :

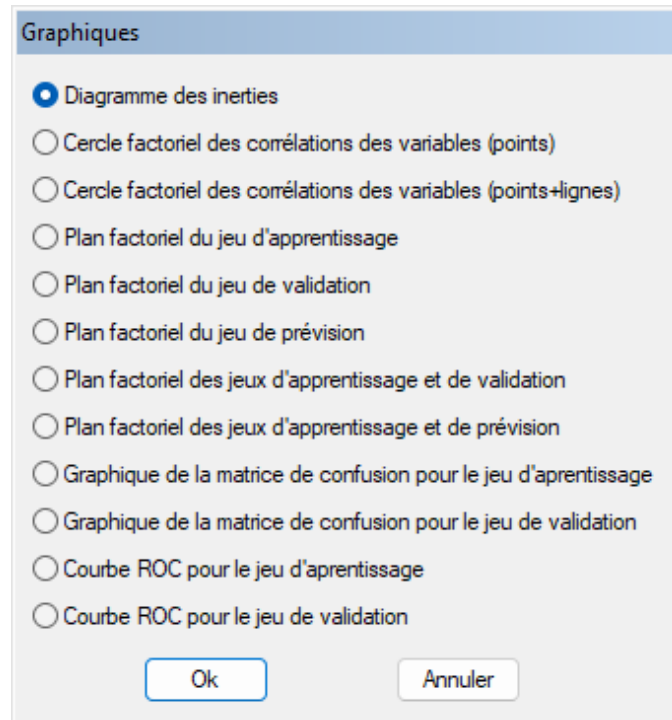


La barre d'outils 'Rapports et Graphiques' permet par l'icône 'Données'  de rappeler la boîte de dialogue d'entrée des données.

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



et l'icône 'Graphiques'  affiche la boîte de dialogue, montrée ci-après, des options pour les graphiques :

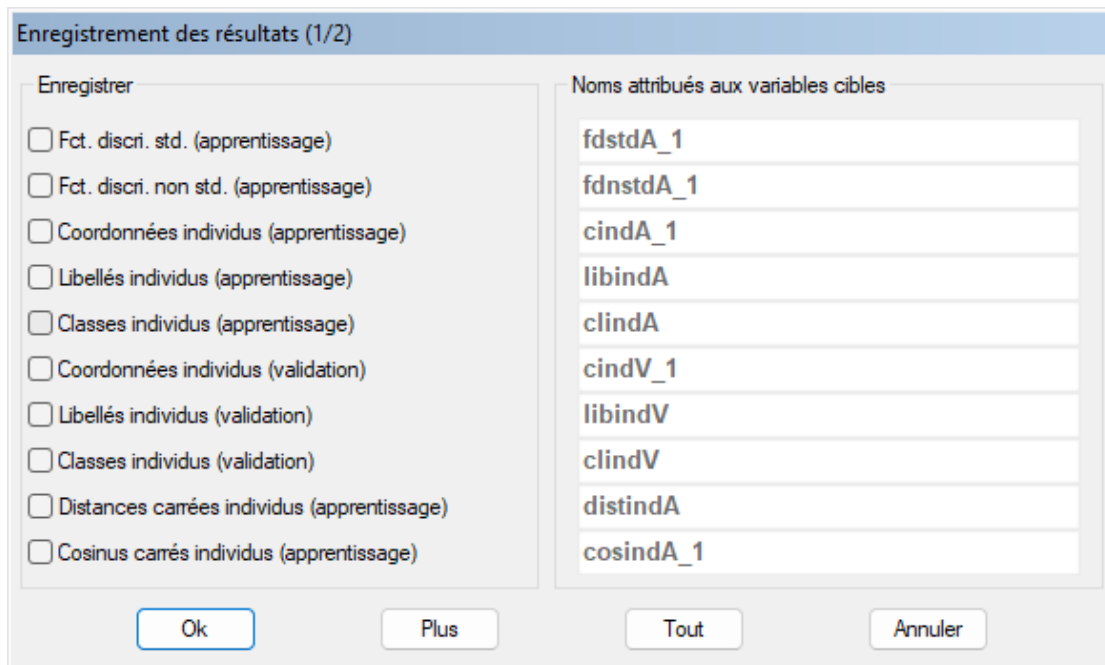


Graphiques

- Diagramme des inerties
- Cercle factoriel des corrélations des variables (points)
- Cercle factoriel des corrélations des variables (points-lignes)
- Plan factoriel du jeu d'apprentissage
- Plan factoriel du jeu de validation
- Plan factoriel du jeu de prévision
- Plan factoriel des jeux d'apprentissage et de validation
- Plan factoriel des jeux d'apprentissage et de prévision
- Graphique de la matrice de confusion pour le jeu d'apprentissage
- Graphique de la matrice de confusion pour le jeu de validation
- Courbe ROC pour le jeu d'apprentissage
- Courbe ROC pour le jeu de validation

Ok Annuler

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



Enregistrement des résultats (1/2)

Enregistrer


- Fct. discri. std. (apprentissage)
- Fct. discri. non std. (apprentissage)
- Coordonnées individus (apprentissage)
- Libellés individus (apprentissage)
- Classes individus (apprentissage)
- Coordonnées individus (validation)
- Libellés individus (validation)
- Classes individus (validation)
- Distances carrées individus (apprentissage)
- Cosinus carrés individus (apprentissage)

Noms attribués aux variables cibles

- fdstdA_1
- fdnstdA_1
- cindA_1
- libindA
- clindA
- cindV_1
- libindV
- clindV
- distindA
- cosindA_1

Ok Plus Tout Annuler

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.

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

Rapports et Graphiques

Rapport ADB

- Probabilités initiales
- Centroides
- Covariances globales
- Corrélations globales
- Covariances par groupe
- Corrélations par groupe
- Covariances intra-groupes
- Corrélations intra-groupes
- Distances de Mahalanobis
- Tableau des inerties
- Test de Pillai
- Test de Box
- Fct. discriminante std.
- Fct. discriminante non std.
- Résultats variables
- Résultats individus
 - Facteur 1
 - Facteur 2
 - Points multiples individus
 - Coord. centres des groupes
 - Coef. fct. clas. lin. (apprentissage)
- Détails classement (apprentissage)
 - Matrice de confusion (apprentissage)
- Stats - Groupes observés
- Stats - Groupes prévus
- Détails classement (validation)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | |
|----|--|---|-------|----------|------------|---------|----------|-----------|------------|
| 1 | | | | | | | | | |
| 2 | RESULTATS INDIVIDUS POUR LE FACTEUR : 1 | | | | | | | | |
| 3 | | | | | | | | | |
| 4 | DISTANCE*2 = CARRÉS DES DISTANCES A L'ORIGINE OU AU BARYCENTRE | | | | | | | | |
| 5 | COORD. = COORDONNÉES DES INDIVIDUS | | | | | | | | |
| 6 | CONTRIB. = CONTRIBUTIONS A L'INERTIE | | | | | | | | |
| 7 | COSINUS*2 = COSINUS CARRÉS | | | | | | | | |
| 8 | COS*2 CUM. = SOMMES CUMULÉES DES COSINUS CARRÉS | | | | | | | | |
| 9 | | | | | | | | | |
| 10 | | | | | | | | | |
| 11 | | | GRUPE | INDIVIDU | DISTANCE*2 | COORD. | CONTRIB. | COSINUS*2 | COS*2 CUM. |
| 12 | 1 | | 1 | 1 | 57,32730 | 7,58828 | 1,48724 | 0,99916 | 0,9991 |
| 13 | 2 | | 1 | 2 | 45,52882 | 6,95660 | 1,16403 | 0,98467 | 0,9846 |
| 14 | 4 | | 1 | 3 | 40,65299 | 6,34050 | 1,04384 | 0,98890 | 0,9889 |
| 15 | 5 | | 1 | 4 | 58,20534 | 7,61605 | 1,50607 | 0,99655 | 0,9965 |
| 16 | 6 | | 1 | 5 | 53,47025 | 7,10429 | 1,34015 | 0,96528 | 0,9652 |
| 17 | 7 | | 1 | 6 | 45,09110 | 6,70541 | 1,16744 | 0,99715 | 0,9971 |
| 18 | 8 | | 1 | 7 | 50,66633 | 7,11765 | 1,31540 | 0,99989 | 0,9998 |
| 19 | 9 | | 1 | 8 | 38,24358 | 6,10453 | 0,96759 | 0,97442 | 0,9744 |
| 20 | 11 | | 1 | 9 | 62,60347 | 7,89480 | 1,61833 | 0,99560 | 0,9956 |
| 21 | 12 | | 1 | 10 | 45,10741 | 6,71478 | 1,17071 | 0,99958 | 0,9995 |

Rapport Explorateur

Rapports et Graphiques

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|-----|---|-----------|---------------|--------------|---|---|---|---|---|----|----|----|----|
| 513 | | | | | | | | | | | | | |
| 514 | ANALYSE LINEAIRE - COEFFICIENTS DES FONCTIONS DE CLASSEMENT | | | | | | | | | | | | |
| 515 | | | | | | | | | | | | | |
| 516 | | | | | | | | | | | | | |
| 517 | | Setosa | Versicolor | Virginica | | | | | | | | | |
| 518 | Constante | -85,34195 | -72,20247 | -103,06840 | | | | | | | | | |
| 519 | lonsepal | 22,72256 | 14,66467 | 10,78294 | | | | | | | | | |
| 520 | larsepal | 23,46884 | 8,91552 | 5,92113 | | | | | | | | | |
| 521 | lonpetal | -14,98624 | 5,89627 | 13,90215 | | | | | | | | | |
| 522 | larpetal | -17,29620 | 4,45679 | 19,29284 | | | | | | | | | |
| 523 | | | | | | | | | | | | | |
| 524 | DISCRIMINATION LINEAIRE | | | | | | | | | | | | |
| 525 | | | | | | | | | | | | | |
| 526 | GRUPE OBSERVE : Setosa | | | | | | | | | | | | |
| 527 | | | | | | | | | | | | | |
| 528 | JEU D'APPRENTISSAGE | | | | | | | | | | | | |
| 529 | | | | | | | | | | | | | |
| 530 | L'INDIVIDU EST AFFECTE AU GROUPE DE PLUS FORTE PROBABILITE | | | | | | | | | | | | |
| 531 | | | | | | | | | | | | | |
| 532 | | | | | | | | | | | | | |
| 533 | INDIVIDU-GROUPE | P(Setosa) | P(Versicolor) | P(Virginica) | | | | | | | | | |

Rapport Général

Rapports et Graphiques

ANALYSE LINEAIRE - MATRICE DE CONFUSION POUR LE JEU D'APPRENTISSAGE

En lignes, les groupes observés
En colonnes, les groupes prévus

Pourcentage de mal classés : 1,667 %
Pourcentage de bien classés (exactitude) : 98,333 %

Précision = VP / (VP + FP)
Rappel = VP / (VP + FN)
Score F1 = 2 x (Précision x Rappel) / (Précision + Rappel)

| | Taille | Setosa | Versicolor | Virginica | Précision | Rappel | Score F1 |
|------------|--------|--------|------------|-----------|-----------|---------|----------|
| Setosa | 43 | 43 | 0 | 0 | 1,00000 | 1,00000 | 1,00000 |
| Versicolor | 38 | 0 | 37 | 1 | 0,97368 | 0,97368 | 0,97368 |
| Virginica | 39 | 0 | 1 | 38 | 0,97436 | 0,97436 | 0,97436 |

STATISTIQUES GROUPE OBSERVE : Setosa

| | lonsepal | larsepal | lonpetal | larpetal |
|----------|----------|----------|----------|----------|
| Effectif | 43,00000 | 43,00000 | 43,00000 | 43,00000 |
| Moyenne | 5,02326 | 3,44419 | 1,46512 | 0,25349 |

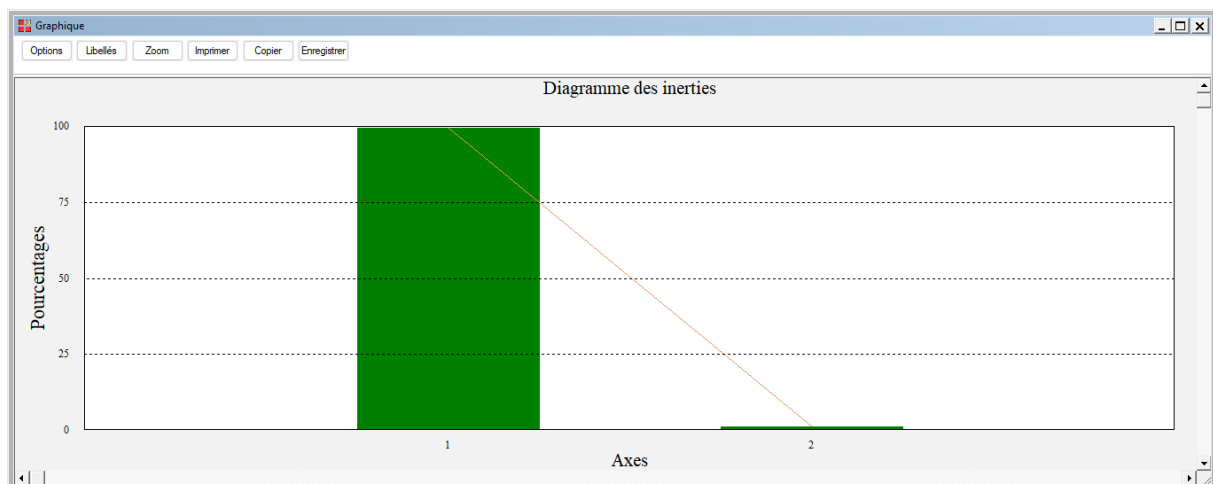
Ces rapports nous fournissent les renseignements suivants :

- Probabilités initiales
- Centroïdes des groupes et global
- Matrices des covariances et corrélations globales et des groupes
- Matrice des covariances et des corrélations intra-groupes
- Distances de Mahalanobis entre les groupes, Fishers, niveaux de signification
- Tableau des inerties (avec corrélation canonique, lambda de Wilks, Khi-2, degrés de liberté et niveau de signification)
- Test de Pillai
- Test de Box
- Fonctions discriminantes standardisées et non standardisées
- Résultats pour les variables et pour les individus
- Coordonnées des centres des groupes
- Détails du classement de la population d'apprentissage
- Matrice de confusion de la population d'apprentissage
- Statistiques pour les groupes observés et prévus (apprentissage)
- Détails du classement de la population de validation
- Matrice de confusion de la population de validation
- Classement de la population de prévision

L'option Graphiques

- Diagramme des inerties

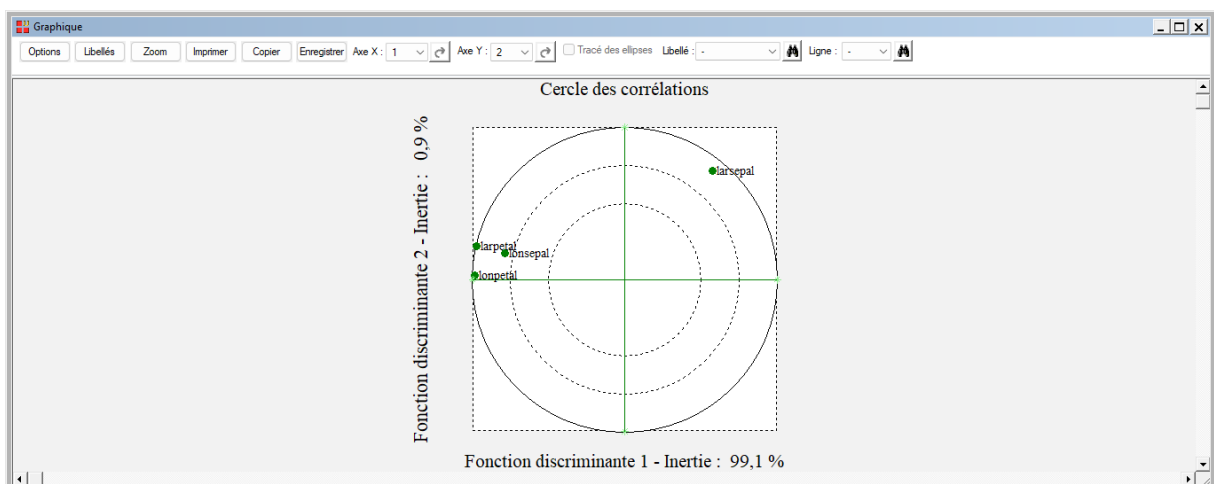
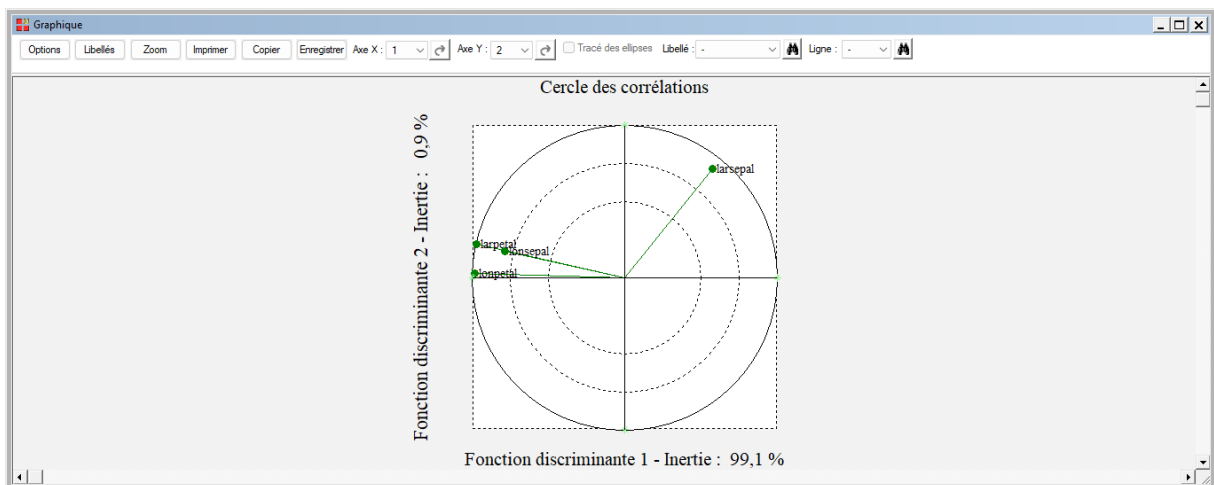
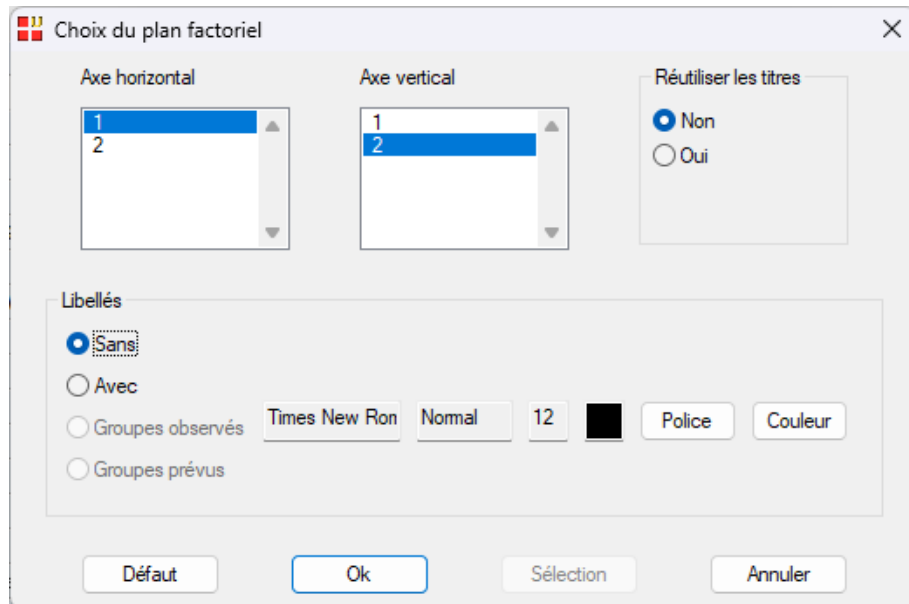
Ce graphique affiche les pourcentages d'inertie pour chacun des axes factoriels.



- Cercle factoriel des corrélations des variables

Ces options permettent d'afficher le cercle de corrélations des variables et de choisir si on désire tracer les lignes reliant les points à l'origine du cercle. L'option sans ces lignes est utile lorsqu'il y a un grand nombre de variables représentées.

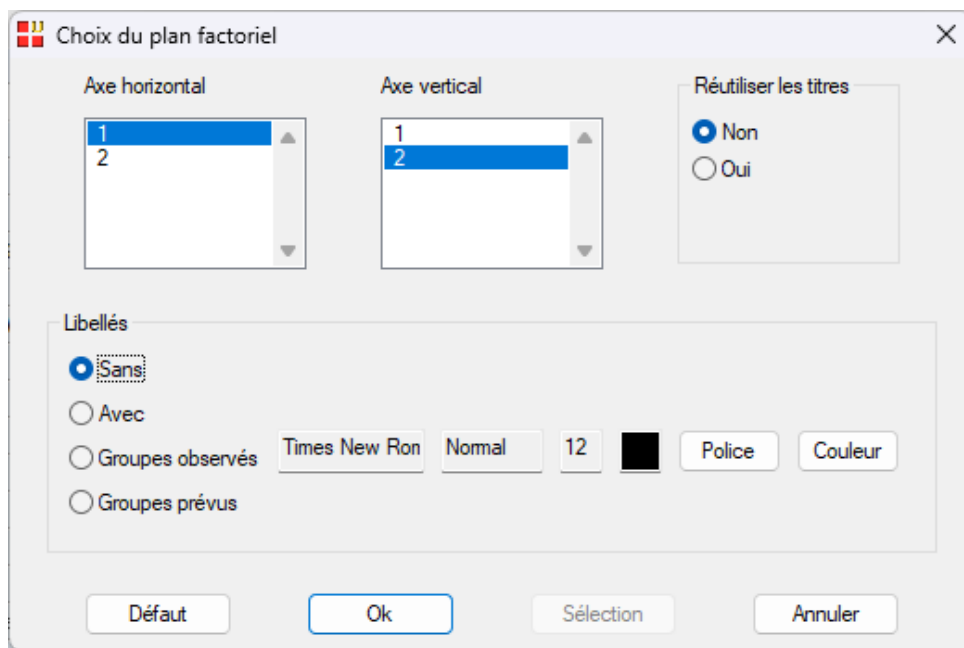
Choisissons les variables avec lignes puis sans lignes. Une boîte de dialogue permettant de choisir le plan factoriel s'affiche. Elle permet également de préciser si l'on désire afficher les libellés des variables, de choisir la couleur et la police et d'indiquer si les titres du graphique (titre 1, titre 2), doivent être conservés pour être réutilisés ultérieurement dans d'autres graphiques créés lors de cette même session de travail.



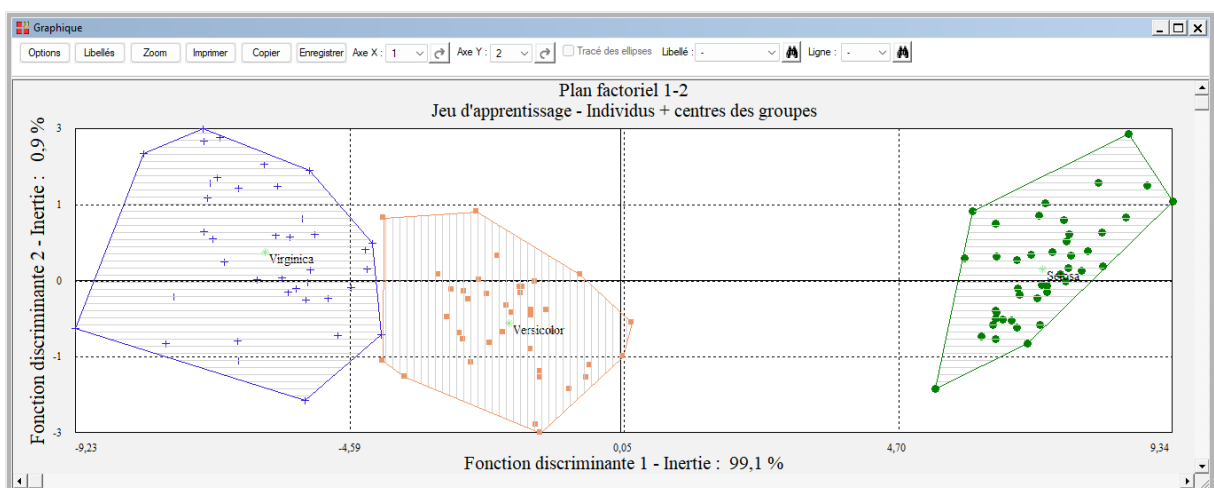
- Plan factoriel des individus et centres des groupes

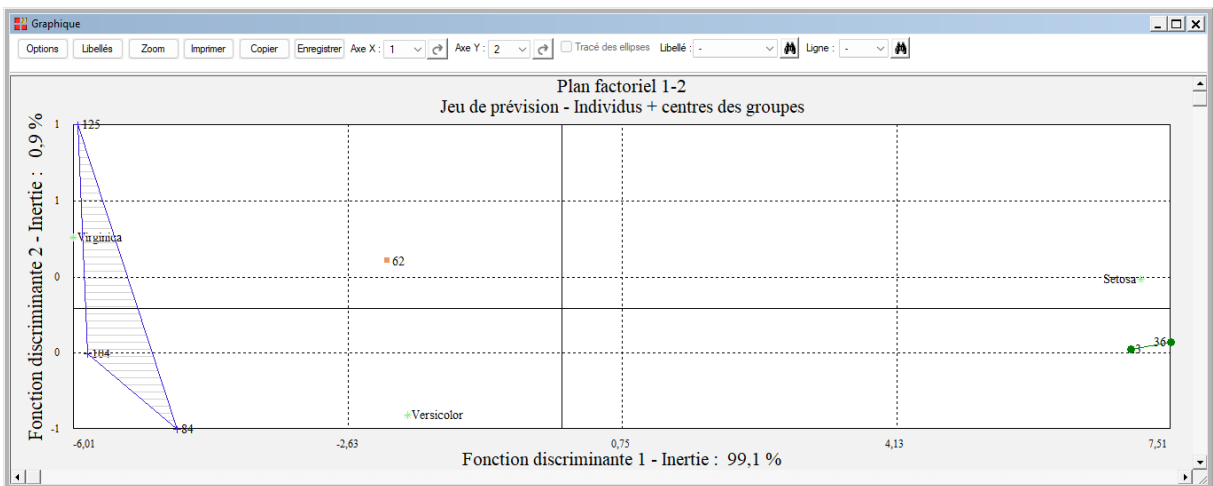
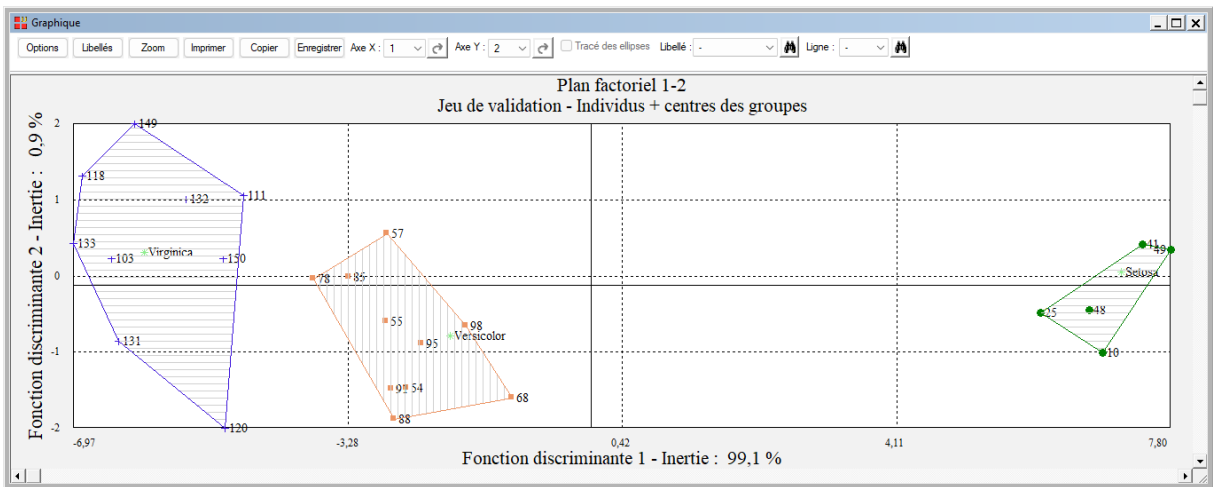
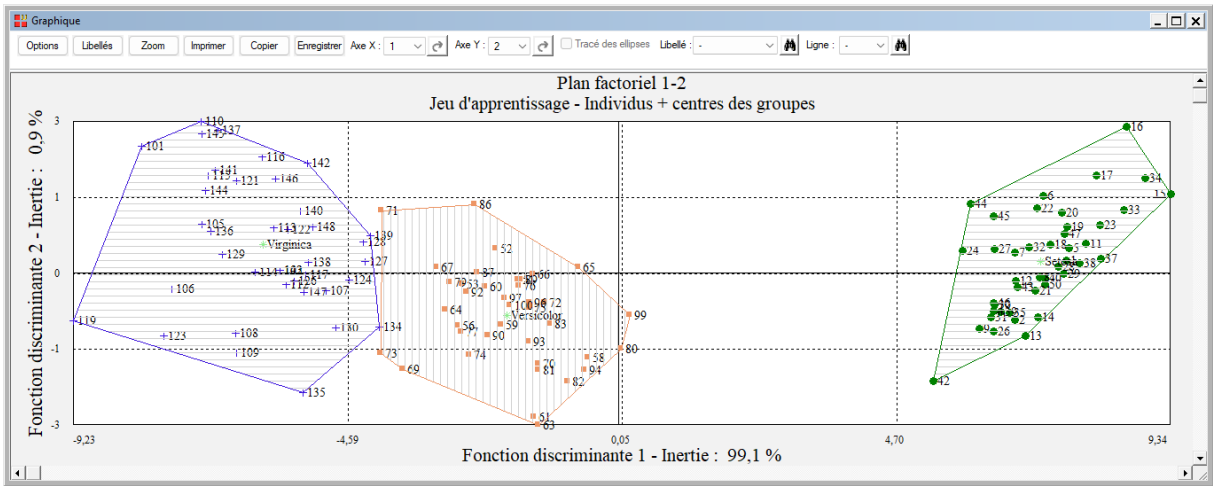
Ces options permettent d'afficher des plans factoriels des individus et des centres des groupes pour les populations d'apprentissage, de validation et de prévision. Une boîte de dialogue permettant de choisir le plan factoriel s'affiche.

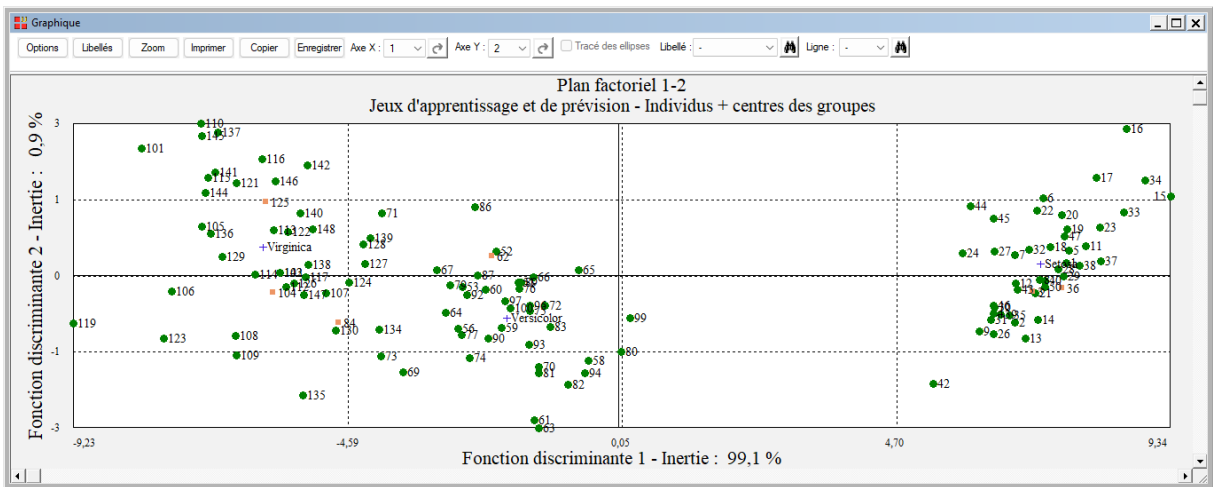
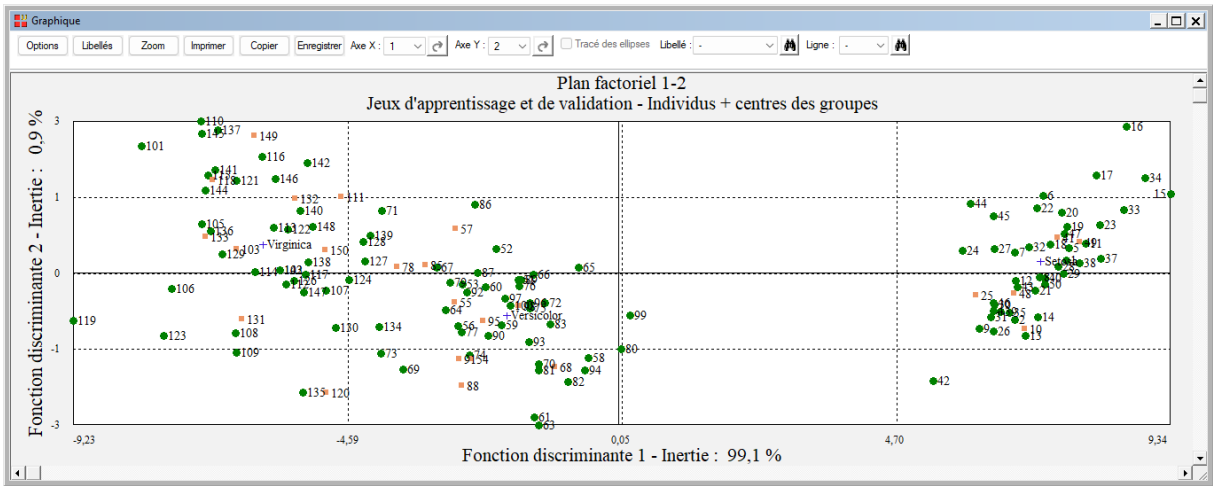
Elle permet de préciser si l'on désire afficher ou non les libellés des individus, de préciser si ces libellés sont les codes des groupes observés ou les codes des groupes prévus, de choisir la couleur et la police pour ces libellés. Il est également possible d'indiquer si les titres du graphique (titre 1, titre 2), doivent être conservés pour être réutilisés ultérieurement dans d'autres graphiques créés lors de cette même session de travail.



Des exemples de plans factoriels sont montrés ci-après.

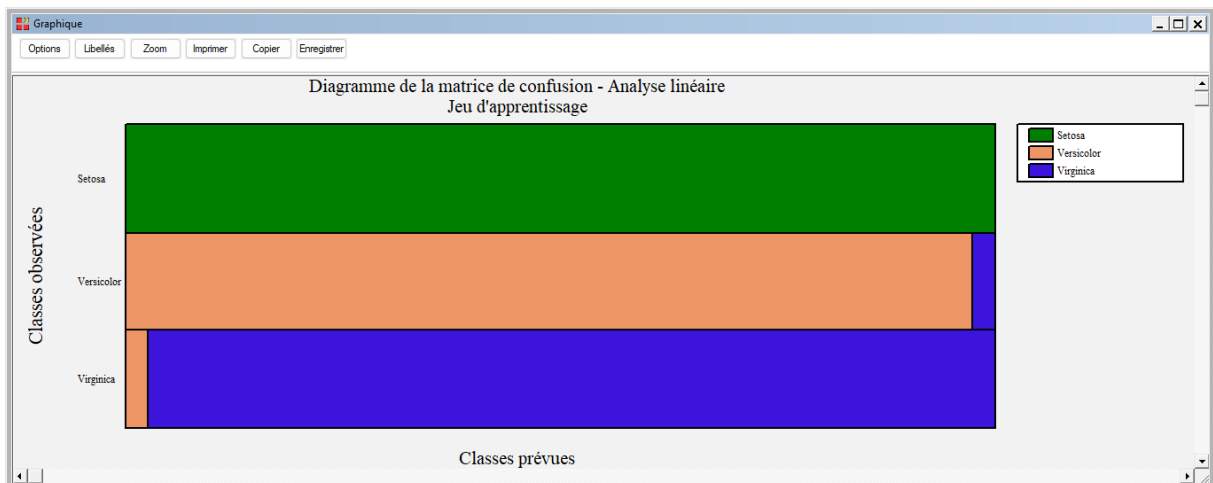


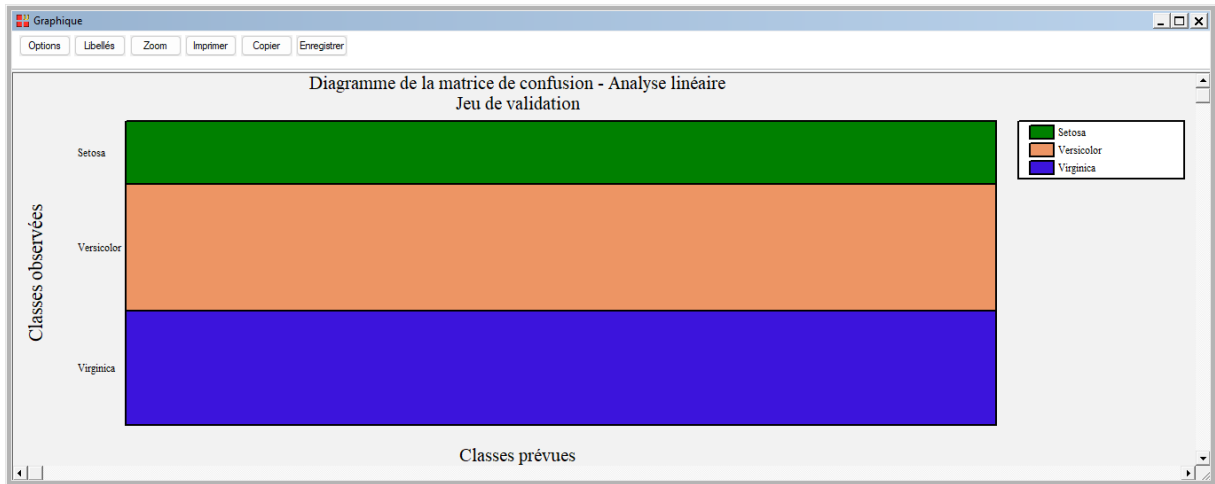




- Graphiques des matrices de confusion

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 disponibles dans cette analyse car la variable à expliquer a plus de deux modalités.

Exemple 2 : Fichier INFARCT2

Pour ce deuxième exemple, nous utiliserons le fichier INFARCT2.

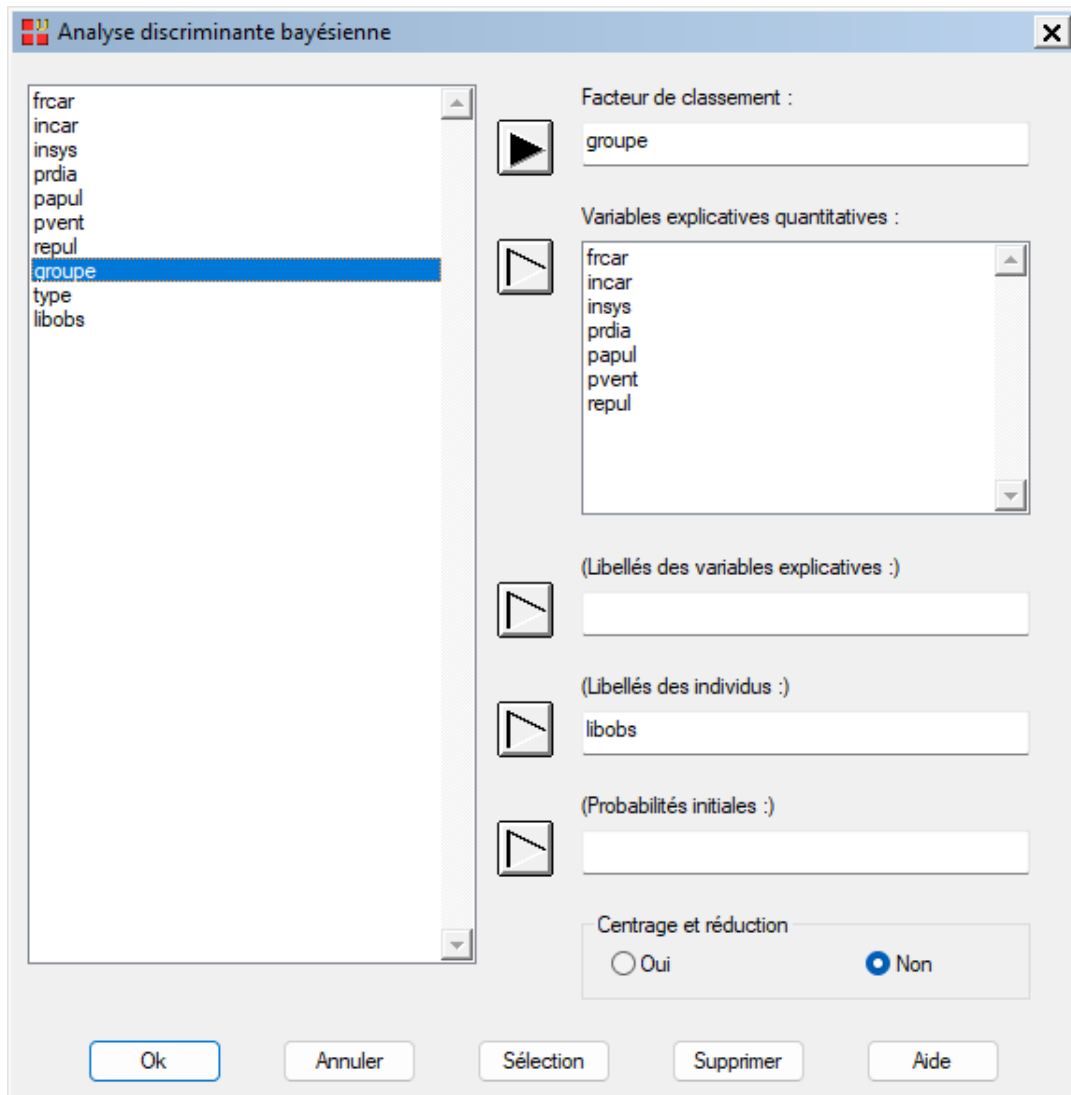
Ce fichier contient des informations concernant 101 victimes d'un infarctus du myocarde. Cette population est constituée d'une population d'apprentissage de 81 individus et d'une population de validation de 20 individus. La population d'apprentissage comprend les individus 11 à 51 (groupe Décès) et les individus 52 à 91 (groupe Survie). La population de validation comprend les individus 1 à 10 (groupe Décès) et les individus 92 à 101 (groupe Survie).

Les variables mesurées sont :

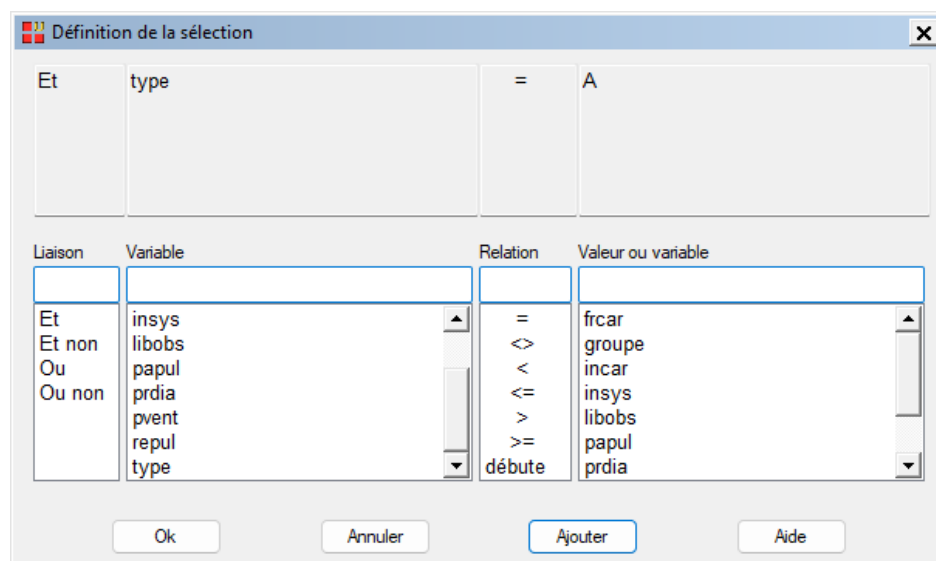
| Variable | Mesure |
|--------------|--------------------------------|
| <i>frcar</i> | fréquence cardiaque |
| <i>incar</i> | index cardiaque |
| <i>insys</i> | index systolique |
| <i>prdia</i> | pression diastolique |
| <i>papul</i> | pression artérielle pulmonaire |
| <i>pvent</i> | pression ventriculaire |
| <i>repul</i> | résistance pulmonaire |

La variable *groupe* indique le groupe d'appartenance de chaque individu (Décès ou Survie). La variable *type* précise la population d'appartenance de chaque individu (A si apprentissage, V si validation). La variable *libobs* contient les libellés des individus des populations d'apprentissage et de validation.

Cliques sur l'icône ADB dans le ruban Expliquer. La boîte de dialogue montrée ci-dessous s'affiche.



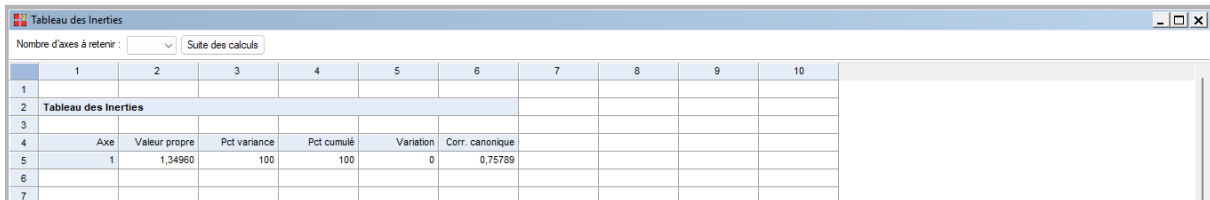
Cliques sur le bouton Sélection pour définir la population d'apprentissage.



Un message nous indique que 81 individus sont sélectionnés.

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

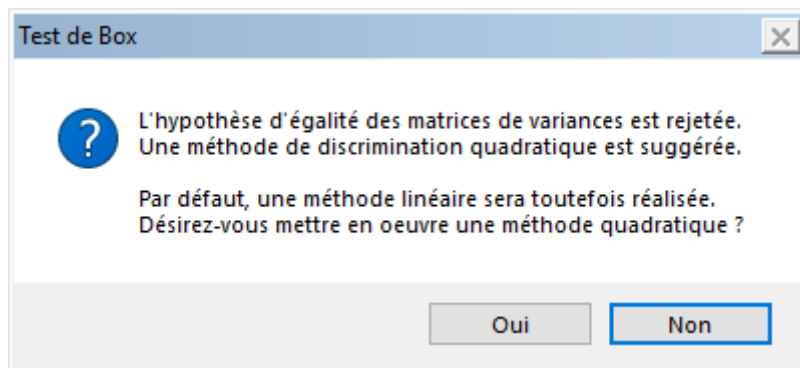
Après quelques instants, un tableau précisant l'inertie expliquée par l'unique vecteur propre issu de l'analyse s'affiche.



| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|-----------------------------|---------------|--------------|------------|-----------|-----------------|---|---|---|----|
| 1 | | | | | | | | | | |
| 2 | Tableau des Inerties | | | | | | | | | |
| 3 | | | | | | | | | | |
| 4 | Axe | Valeur propre | Pct variance | Pct cumulé | Variation | Corr. canonique | | | | |
| 5 | 1 | 1,34960 | 100 | 100 | 0 | 0,75789 | | | | |
| 6 | | | | | | | | | | |
| 7 | | | | | | | | | | |

Aucun graphique n'est proposé, car il n'y a qu'une unique valeur propre.

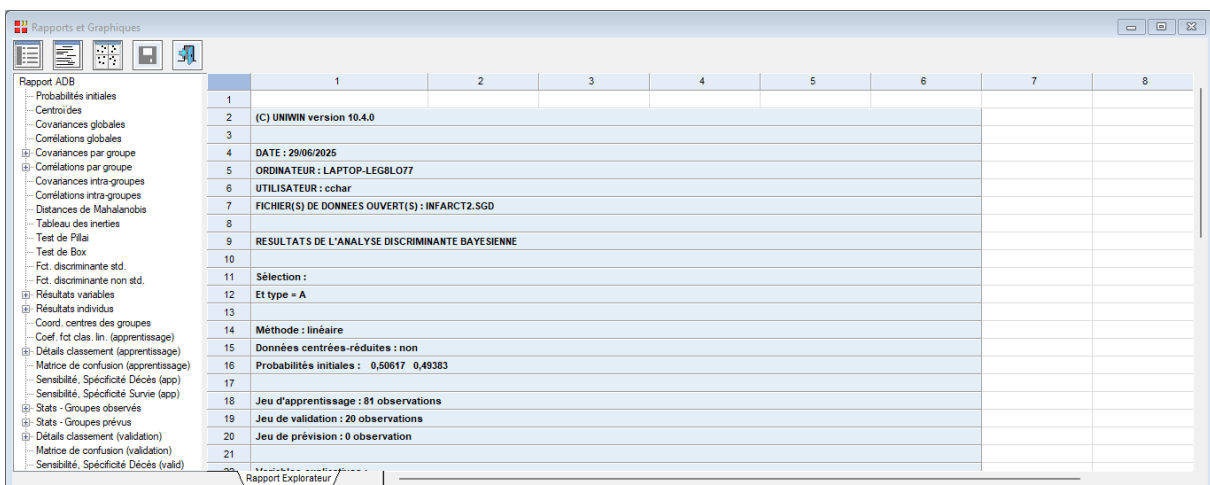
Après quelques instants, UNIWIN affiche une boîte de dialogue vous indiquant si l'hypothèse d'égalité des matrices de variances est vérifiée ou non.



Dans notre exemple, l'hypothèse doit être rejetée et donc une analyse discriminante quadratique est suggérée.

Choisissons donc de mettre en œuvre une analyse quadratique.

Après quelques instants, l'écran suivant s'affiche :



| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|----|--|----------------|---|---|---|---|---|---|
| 1 | | | | | | | | |
| 2 | (C) UNIWIN version 10.4.0 | | | | | | | |
| 3 | | | | | | | | |
| 4 | DATE : | 29/06/2025 | | | | | | |
| 5 | ORDINATEUR : | LAPTOP-LEGL077 | | | | | | |
| 6 | UTILISATEUR : | cchar | | | | | | |
| 7 | FICHIER(S) DE DONNEES OUVERT(S) : | INFARCT2.SGD | | | | | | |
| 8 | | | | | | | | |
| 9 | RESULTATS DE L'ANALYSE DISCRIMINANTE BAYESIENNE | | | | | | | |
| 10 | | | | | | | | |
| 11 | Sélection : | | | | | | | |
| 12 | Et type = A | | | | | | | |
| 13 | | | | | | | | |
| 14 | Méthode : linéaire | | | | | | | |
| 15 | Données centrées-réduites : non | | | | | | | |
| 16 | Probabilités initiales : 0,50617 0,49383 | | | | | | | |
| 17 | | | | | | | | |
| 18 | Jeu d'apprentissage : 81 observations | | | | | | | |
| 19 | Jeu de validation : 20 observations | | | | | | | |
| 20 | Jeu de prévision : 0 observation | | | | | | | |
| 21 | | | | | | | | |

La barre d'outils 'Rapports et Graphiques' permet par l'icône 'Données' de rappeler la boîte de dialogue d'entrée des données, par l'icône 'Rapports' d'afficher la boîte de dialogue des options pour les rapports et par et par l'icône 'Graphiques' d'afficher la boîte de dialogue des options pour les graphiques.

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.

Ces rapports nous fournissent les renseignements suivants :

- Probabilités initiales
- Centroïdes des groupes et global
- Matrices des covariances et corrélations globales et des groupes
- Matrice des covariances et des corrélations intra-groupes
- Distances de Mahalanobis entre les groupes, Fishers, niveaux de signification
- Tableau des inerties (avec corrélation canonique, lambda de Wilks, Khi-2, degrés de liberté et niveau de signification)
- Tests de Pillai et de Box
- Fonctions discriminantes standardisées et non standardisées
- Résultats pour les variables et pour les individus
- Coordonnées des centres des groupes
- Coefficients des fonctions de classement
- Détails du classement de la population d'apprentissage
- Matrice de confusion de la population d'apprentissage
- Seuils, spécificités, sensibilités (apprentissage)
- Statistiques pour les groupes observés et prévus (apprentissage)
- Détails du classement de la population de validation
- Matrice de confusion de la population de validation
- Seuils, spécificités, sensibilités (validation)
- Classement de la population de prévision

The screenshot shows the 'Rapports et Graphiques' software interface. The 'Rapport Explorateur' window displays a table with 8 columns and 21 rows. The table contains discriminant analysis results for 20 observations across 8 variables. The first 8 rows are headers for various statistical measures, and the remaining 12 rows provide data for individual observations, including group assignments, distances, coordinates, contributions, and cosinus values.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|----|--|-------|----------|------------|----------|----------|-----------|------------|
| 1 | | | | | | | | |
| 2 | RESULTATS INDIVIDUS POUR LE FACTEUR : 1 | | | | | | | |
| 3 | | | | | | | | |
| 4 | DISTANCE*2 = CARRÉS DES DISTANCES A L'ORIGINE OU AU BARYCENTRE | | | | | | | |
| 5 | COORD. = COORDONNÉES DES INDIVIDUS | | | | | | | |
| 6 | CONTRIB. = CONTRIBUTIONS A L'INERTIE | | | | | | | |
| 7 | COSINUS*2 = COSINUS CARRÉS | | | | | | | |
| 8 | COS*2 CUM. = SOMMES CUMULÉES DES COSINUS CARRÉS | | | | | | | |
| 9 | | | | | | | | |
| 10 | | | | | | | | |
| 11 | | GRUPE | INDIV/DU | DISTANCE*2 | COORD. | CONTRIB. | COSINUS*2 | COS*2 CUM. |
| 12 | Obs011 | 1 | 1 | 0,12121 | -0,34815 | 0,06531 | 1 | |
| 13 | Obs012 | 1 | 2 | 2,89426 | 1,69631 | 1,55397 | 1 | |
| 14 | Obs013 | 1 | 3 | 0,00002 | -0,00455 | 0,00001 | 1 | |
| 15 | Obs014 | 1 | 4 | 4,63268 | 2,19833 | 2,60372 | 1 | |
| 16 | Obs015 | 1 | 5 | 1,90155 | 1,37897 | 1,02451 | 1 | |
| 17 | Obs016 | 1 | 6 | 2,85242 | 1,68891 | 1,53681 | 1 | |
| 18 | Obs017 | 1 | 7 | 1,29329 | 1,13723 | 0,69679 | 1 | |
| 19 | Obs018 | 1 | 8 | 0,22116 | 0,47028 | 0,11915 | 1 | |
| 20 | Obs019 | 1 | 9 | 3,80729 | 1,95123 | 2,05127 | 1 | |
| 21 | Obs020 | 1 | 10 | 1,19438 | 1,09288 | 0,64350 | 1 | |

Rapports et Graphiques

Rapport ADB

- Probabilités initiales
- Centroids
- Covariances globales
- Corrélations globales
- Covariances par groupe
- Corrélations par groupe
- Covariances intra-groupes
- Corrélations intra-groupes
- Distances de Mahalanobis
- Tableau des inerties
- Test de Pits
- Test de Box
- Fct. discriminante std.
- Fct. discriminante non std.
- Résultats variables
- Résultats individus
- Coord. centres des groupes
- Coef. fct. clas. quad. (apprentissage)
- Détails classement (apprentissage)
- Matrice de confusion (apprentissage)**
- Sensibilité, Spécificité Décès (app)
- Sensibilité, Spécificité Survie (app)
- Stats - Groupes observés
- Stats - Groupes prévus
- Détails classement (validation)
- Matrice de confusion (validation)
- Sensibilité, Spécificité Décès (valid)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|----|--|----|--------|-------|--------|-----------|---------|----------|
| 1 | | | | | | | | |
| 2 | ANALYSE QUADRATIQUE - MATRICE DE CONFUSION POUR LE JEU D'APPRENTISSAGE | | | | | | | |
| 3 | | | | | | | | |
| 4 | En lignes, les groupes observés | | | | | | | |
| 5 | En colonnes, les groupes prévus | | | | | | | |
| 6 | | | | | | | | |
| 7 | Pourcentage de mal classés : 11,111 % | | | | | | | |
| 8 | Pourcentage de bien classés (exacttude) : 88,889 % | | | | | | | |
| 9 | | | | | | | | |
| 10 | Précision = VP / (VP + FP) | | | | | | | |
| 11 | Rappel = VP / (VP + FN) | | | | | | | |
| 12 | Score F1 = 2 x (Précision x Rappel) / (Précision + Rappel) | | | | | | | |
| 13 | | | | | | | | |
| 14 | | | | | | | | |
| 15 | | | Taille | Décès | Survie | Précision | Rappel | Score F1 |
| 16 | Décès | 41 | | 36 | 5 | 0,90000 | 0,87805 | 0,88889 |
| 17 | Survie | | 40 | 4 | 36 | 0,87805 | 0,90000 | 0,88889 |
| 18 | | | | | | | | |
| 19 | | | | | | | | |
| 20 | | | | | | | | |
| 21 | | | | | | | | |

Rapport Explorateur /

L'option Graphiques

- Diagramme des inerties

Ce diagramme n'est pas disponible dans cet exemple car il n'y a qu'une seule composante.

- Cercle factoriel des corrélations des variables

Ces options permettent d'afficher le cercle de corrélations des variables et de choisir si on désire tracer les lignes reliant les points à l'origine du cercle. L'option sans ces lignes est utile lorsqu'il y a un grand nombre de variables représentées. Choisissons les variables avec lignes puis sans lignes. 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.

Choix du plan factoriel

Axe horizontal: 1

Axe vertical: 1

Réutiliser les titres: Non Oui

Libellés: Sans Avec

Groupes observés: Times New Ron, Normal, 12, [Noir], Police, Couleur

Groupes prévus

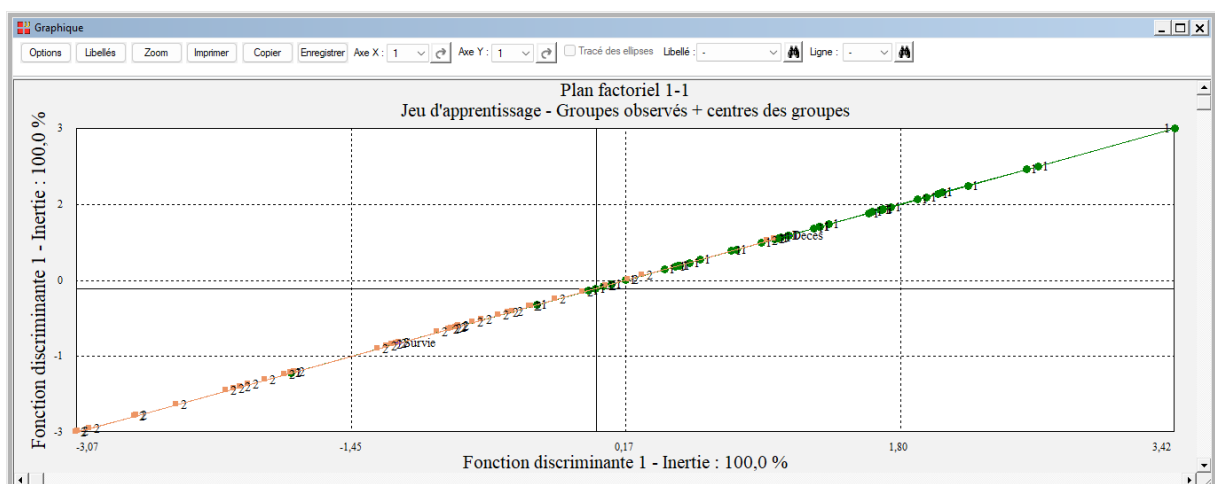
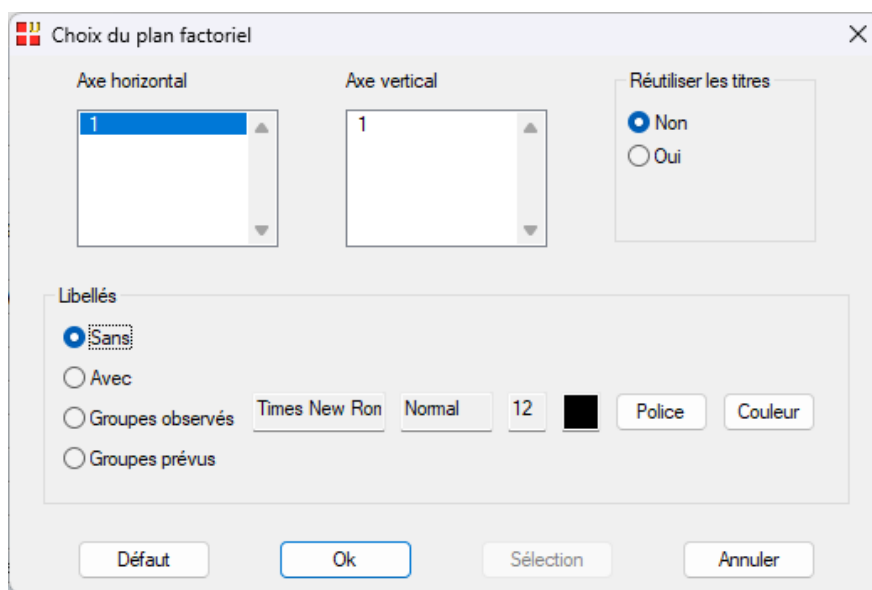
Défaut Ok Sélection Annuler

A noter que dans notre exemple, il n'y a qu'un axe factoriel et donc le tracé n'est pas proposé.

- Plan factoriel des individus et centres des groupes

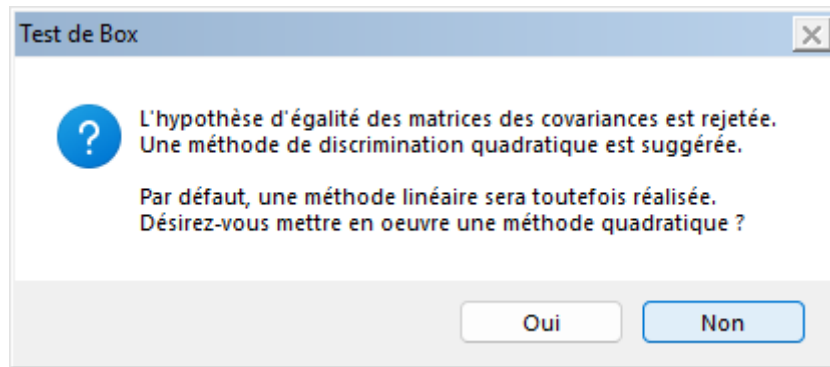
Ces options permettent d'afficher des plans factoriels des individus et des centres des groupes pour les populations d'apprentissage, de validation et de prévision.

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.



Refaisons maintenant cette même analyse en choisissant une méthode linéaire.

Pour cela, il faut refaire l'analyse actuelle en choisissant 'Non' lorsque la boîte de dialogue 'Test de Box' s'affiche.

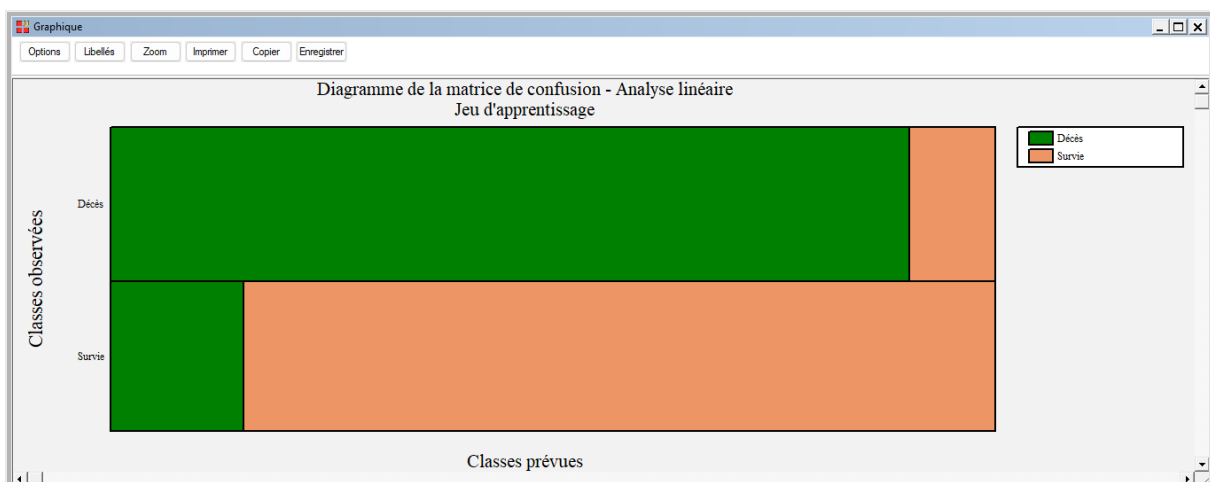


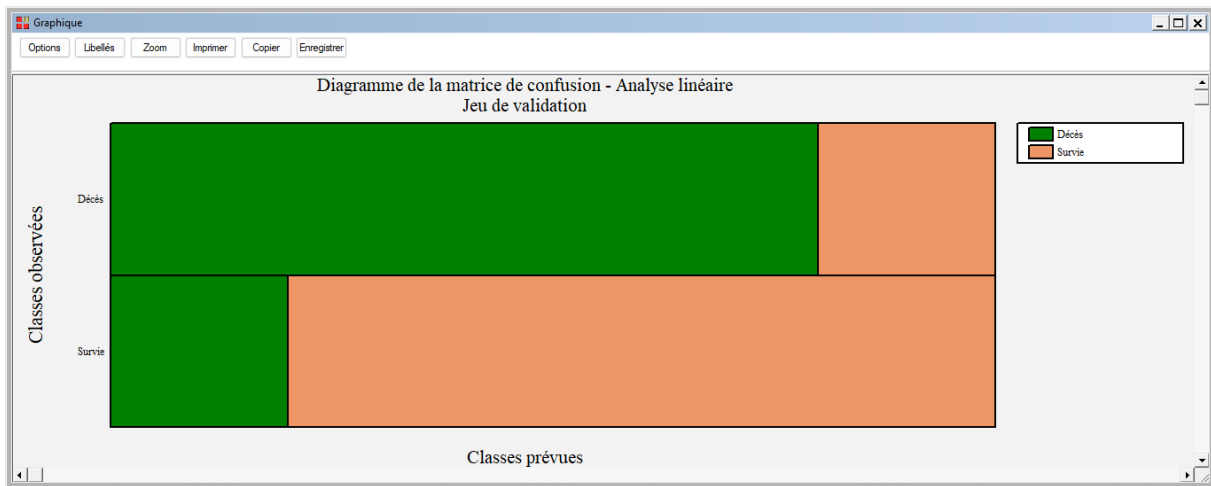
Rapports et Graphiques

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|----|---|----|--------|-------|---------|-----------|---------|----------|
| 1 | | | | | | | | |
| 2 | ANALYSE LINEAIRE - MATRICE DE CONFUSION POUR LE JEU D'APPRENTISSAGE | | | | | | | |
| 3 | | | | | | | | |
| 4 | En lignes, les groupes observés | | | | | | | |
| 5 | En colonnes, les groupes prévus | | | | | | | |
| 6 | | | | | | | | |
| 7 | Pourcentage de mal classés : 9,877 % | | | | | | | |
| 8 | Pourcentage de bien classés (exactitude) : 90,123 % | | | | | | | |
| 9 | | | | | | | | |
| 10 | Précision = VP / (VP + FP) | | | | | | | |
| 11 | Rappel = VP / (VP + FN) | | | | | | | |
| 12 | Score F1 = 2 x (Précision x Rappel) / (Précision + Rappel) | | | | | | | |
| 13 | | | | | | | | |
| 14 | | | | | | | | |
| 15 | | | Taille | Décès | Survie | Précision | Rappel | Score F1 |
| 16 | Décès | 41 | 38 | 3 | 0,88372 | 0,92883 | 0,90476 | |
| 17 | Survie | 40 | 5 | 35 | 0,92105 | 0,87500 | 0,89744 | |
| 18 | | | | | | | | |
| 19 | | | | | | | | |
| 20 | | | | | | | | |
| 21 | | | | | | | | |

Le pourcentage d'erreur de classement (9,88 %) est meilleur que pour l'analyse quadratique (11,11 %) mais l'hypothèse d'égalité des matrices des variances, non observée, y est supposée.

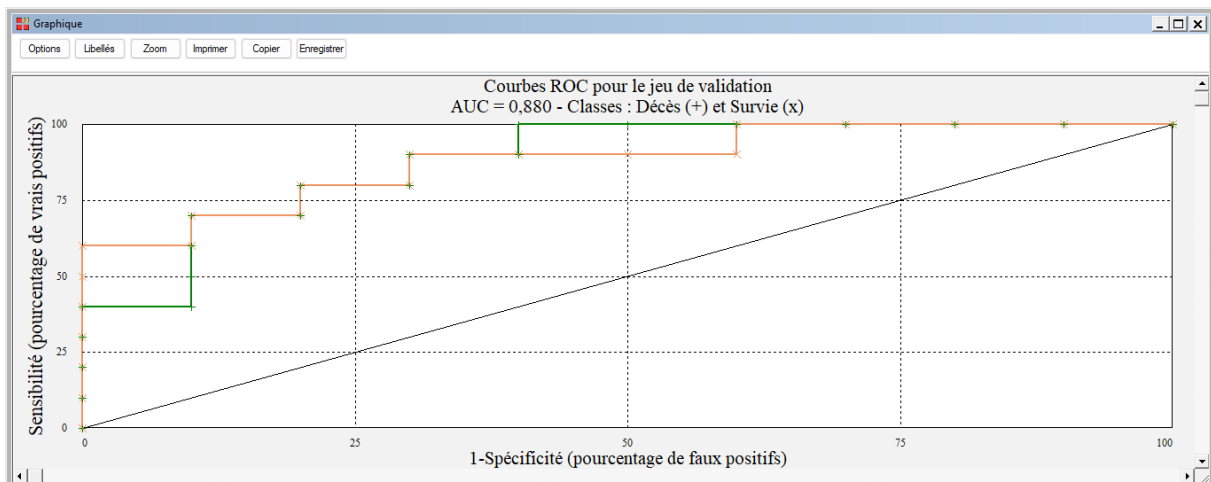
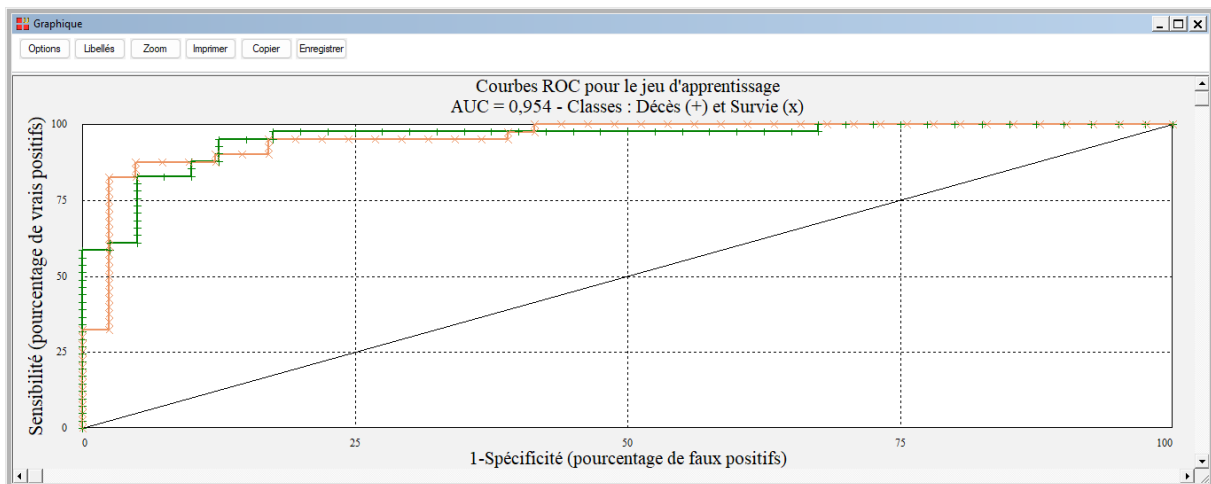
- Graphiques des matrices de confusion





- Courbes ROC

Le tracé de la courbe ROC et le calcul de l'aire sous la courbe (AUC) sont possibles car la variable à expliquer possède deux modalités. Visualisons les courbes pour le jeu d'apprentissage et le jeu de validation.



Exemple 3 : Fichier BORDEAUX

Pour ce troisième exemple, nous utiliserons le fichier BORDEAUX.

Ce fichier contient des informations sur la qualité de vins de Bordeaux en relation avec les conditions météorologiques.

La variable qualitative *Qualité*, facteur de classement, prend trois modalités : Bon, Moyen et Médiocre.

Les variables explicatives sont :

- *Temp* somme des températures moyennes journalières (° C)
- *Insol* durée d'insolation (heures)
- *Chaleur* nombre de jours de grande chaleur
- *Pluie* hauteur des pluies (millimètres)

La variable *Année* fournit les libellés des individus.

Cliquons sur l'icône ADB dans le ruban Expliquer pour afficher la boîte de dialogue d'entrée des données.

Nous sélectionnons une analyse centrée et réduite.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|-----------------------------|---------------|--------------|------------|-----------|-----------------|---|
| 1 | | | | | | | |
| 2 | Tableau des Inerties | | | | | | |
| 3 | | | | | | | |
| 4 | Axe | Valeur propre | Pct variance | Pct cumulé | Variation | Corr. canonique | |
| 5 | 1 | 3,27886 | 95,94508 | 95,94508 | 0,00000 | 0,87538 | |
| 6 | 2 | 0,13857 | 4,05492 | 100,00000 | 91,89017 | 0,34887 | |
| 7 | | | | | | | |

La fenêtre 'Rapports et Graphiques' s'affiche pour la méthode linéaire.

Visualisons les résultats du classement.

Les vins mal classés sont les vins des années :

- 1953, 1955 bons classés moyens
- 1933, 1950 moyens classés bons
- 1926 moyen classé médiocre
- 1935, 1956 médiocres classés moyens

Analyse discriminante bayésienne

Année
Temp
Insol
Chaleur
Pluie
Qualité

Facteur de classement :
Qualité

Variables explicatives quantitatives :
Temp
Insol
Chaleur
Pluie

(Libellés des variables explicatives :)

(Libellés des individus :)
Année

(Probabilités initiales :)

Centrage et réduction
 Oui Non

Ok
Annuler
Sélection
Supprimer
Aide

Rapports et Graphiques

Rapport ADB

- Probabilités initiales
- Centroides
- Covariances globales
- Corrélations globales
- (+) Covariances par groupe
- (+) Corrélations par groupe
- Covariances intra-groupes
- Corrélations intra-groupes
- Distances de Mahalanobis
- Tableau des inerties
- Test de Pillai
- Test de Box
- Fct. discriminante std.
- Fct. discriminante non std.
- (+) Résultats variables
- (+) Résultats individus
- Coord. centres des groupes
- Coef. fct. clas. lin. (apprentissage)
- (+) Détails classement (apprentissage)
- (+) Matrice de confusion (apprentissage)
- (+) Stats - Groupes observés
- (+) Stats - Groupes prévus

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|----|---|--------|-----|-------|----------|-----------|---------|----------|---|
| 1 | | | | | | | | | |
| 2 | ANALYSE LINEAIRE - MATRICE DE CONFUSION POUR LE JEU D'APPRENTISSAGE | | | | | | | | |
| 3 | | | | | | | | | |
| 4 | En lignes, les groupes observés | | | | | | | | |
| 5 | En colonnes, les groupes prévus | | | | | | | | |
| 6 | | | | | | | | | |
| 7 | Pourcentage de mal classés : 20,588 % | | | | | | | | |
| 8 | Pourcentage de bien classés (exactitude) : 79,412 % | | | | | | | | |
| 9 | | | | | | | | | |
| 10 | Précision = VP / (VP + FP) | | | | | | | | |
| 11 | Rappel = VP / (VP + FN) | | | | | | | | |
| 12 | Score F1 = 2 x (Précision x Rappel) / (Précision + Rappel) | | | | | | | | |
| 13 | | | | | | | | | |
| 14 | | | | | | | | | |
| 15 | | Taille | Bon | Moyen | Médiocre | Précision | Rappel | Score F1 | |
| 16 | Bon | 11 | 9 | 2 | 0 | 0,81818 | 0,81818 | 0,81818 | |
| 17 | Moyen | 11 | 2 | 8 | 1 | 0,66667 | 0,72727 | 0,69565 | |
| 18 | Médiocre | 12 | 0 | 2 | 10 | 0,90909 | 0,83333 | 0,86957 | |
| 19 | | | | | | | | | |
| 20 | | | | | | | | | |
| 21 | | | | | | | | | |

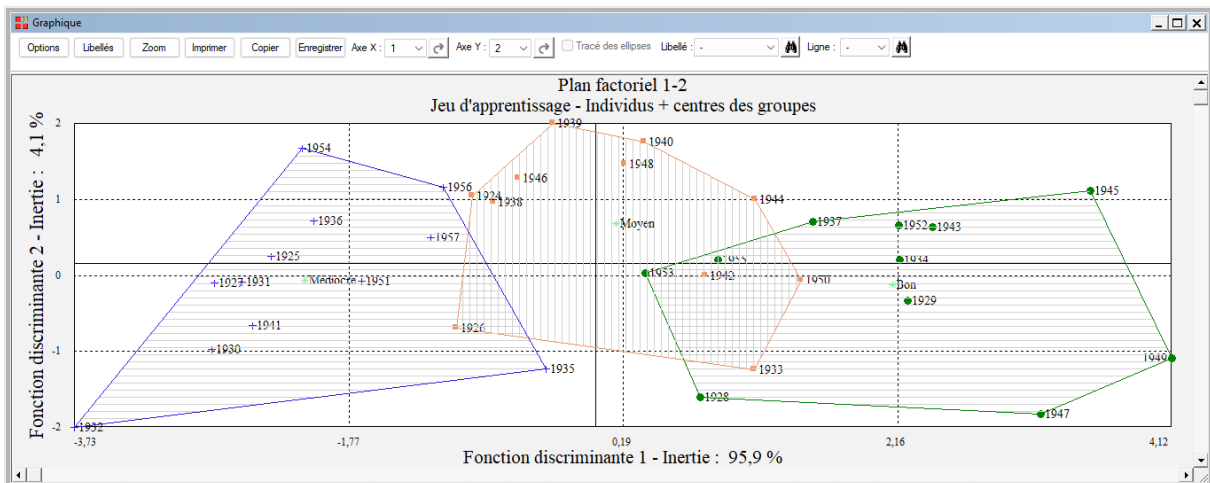
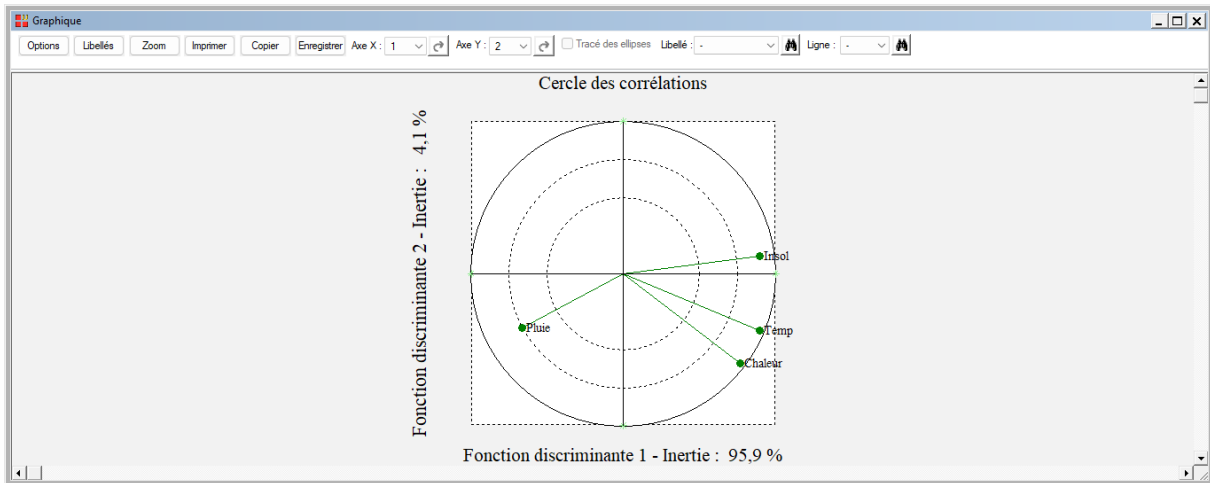
Rapports et Graphiques

Rapport ADB

- Probabilités initiales
- Centroids
- Covariances globales
- Corrélations globales
- Covariances par groupe
- Corrélations par groupe
- Covariances intra-groupes
- Corrélations intra-groupes
- Distances de Mahalanobis
- Tableau des inerties
- Test de Pits
- Test de Box
- Fct. discriminante std.
- Fct. discriminante non std.
- Résultats variables
- Résultats individus
- Coord. centres des groupes
- Coef. fct. clas. lin. (apprentissage)
- Détails classement (apprentissage)
 - Bon
 - Moyen
 - Médiocre
- Matrice de confusion (apprentissage)
- Stats - Groupes observés
- Stats - Groupes prévus

| INDIVIDU-GROUPE | P(Bon) | P(Moyen) | P(Médiocre) |
|-----------------|---------|----------|-------------|
| 1928 - Bon | 0,64171 | 0,32703 | 0,03126 |
| 1929 - Bon | 0,93341 | 0,06650 | 0,00009 |
| 1934 - Bon | 0,89244 | 0,10745 | 0,00012 |
| 1937 - Bon | 0,62214 | 0,37672 | 0,00114 |
| 1943 - Bon | 0,90475 | 0,09520 | 0,00004 |
| 1945 - Bon | 0,98380 | 0,01620 | 0,00000 |
| 1947 - Bon | 0,99656 | 0,00344 | 0,00000 |
| 1949 - Bon | 0,99904 | 0,00096 | 0,00000 |
| 1952 - Bon | 0,85207 | 0,14781 | 0,00012 |
| 1953 - Moyen * | 0,19487 | 0,75191 | 0,05322 |

Demandons également le cercle factoriel et le plan factoriel pour les axes 1 et 2.



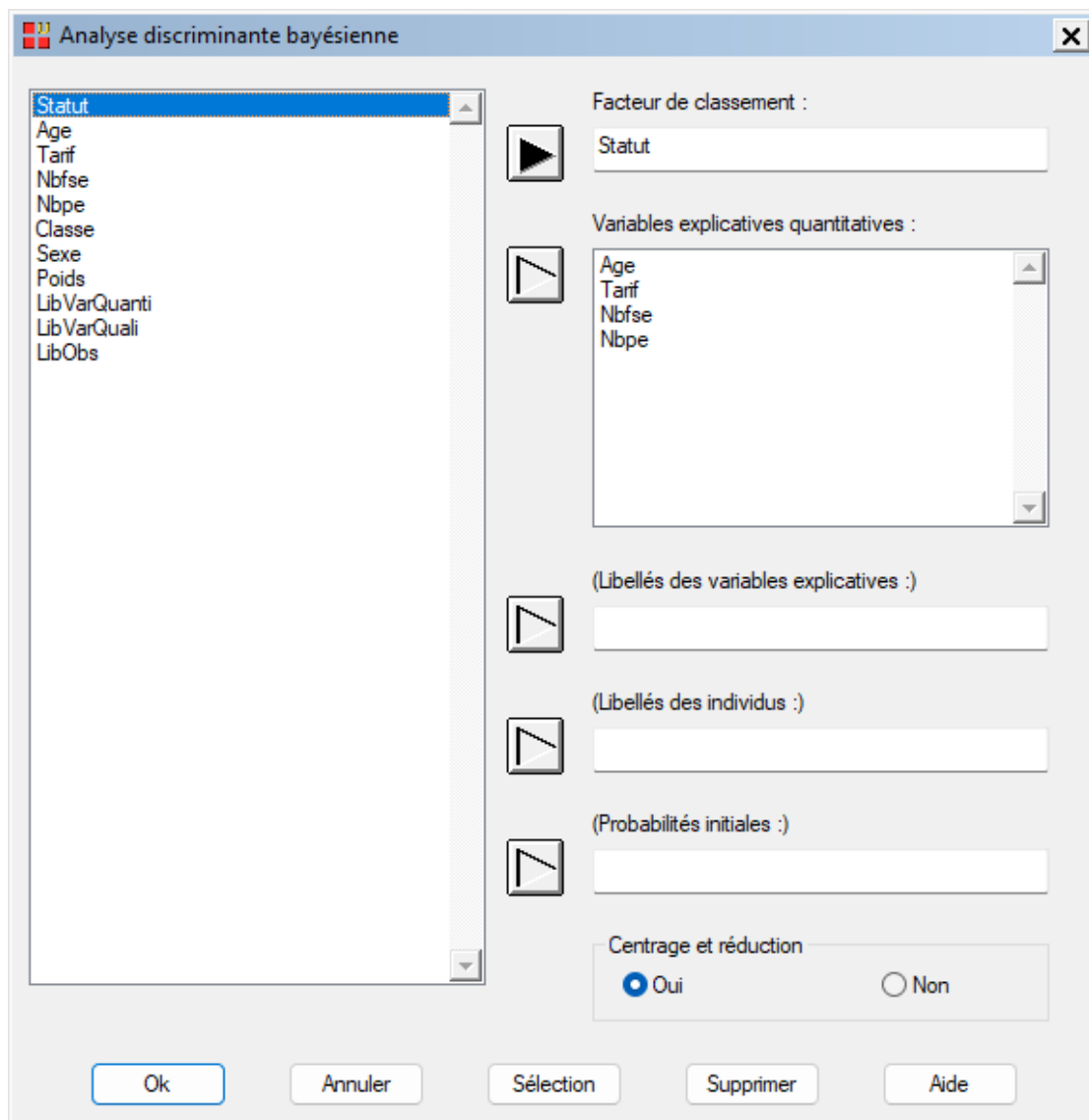
Exemple 4 : Fichier TITANIC

Pour ce quatrième exemple, nous utiliserons le fichier TITANIC.

Ce fichier contient des informations concernant 714 passagers :

| | |
|--------|---|
| Statut | Survie ou Décès |
| Classe | Classe du passager (1 ^{ère} , 2 ^{ème} ou 3 ^{ème}) |
| Sexe | Homme ou Femme |
| Age | Age du passager |
| Nbfse | Nombre de frères, sœurs ou époux, épouses à bord |
| Nbpe | Nombre de parents ou enfants à bord |
| Tarif | Tarif passager (en £) |

Cliquons sur l'icône ADB dans le ruban Expliquer et renseignons la boîte de dialogue comme montré ci-dessous.



Après exécution de la procédure par une analyse quadratique, visualisons la matrice de confusion des données d'apprentissage et la courbe ROC associée.

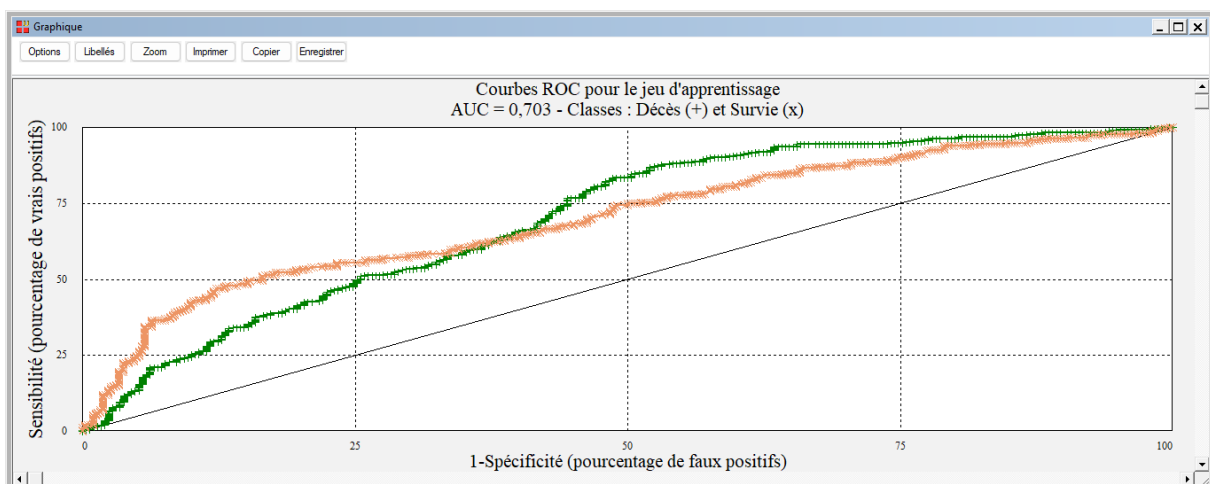
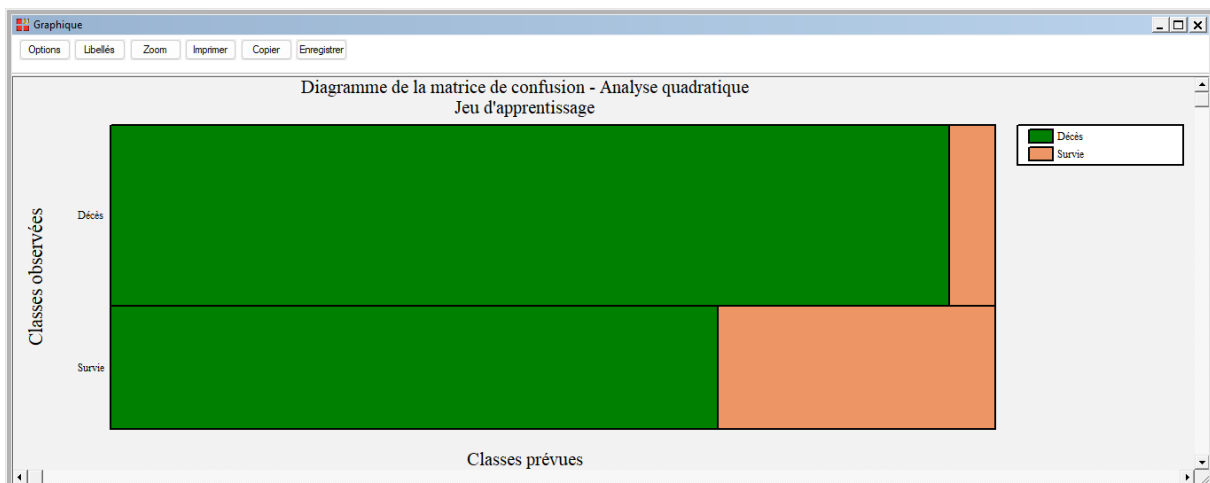
Rapports et Graphiques

Rapport ADB

- Probabilités initiales
- Centroides
- Covariances globales
- Corrélations globales
- Covariances par groupe
- Corrélations par groupe
- Covariances intra-groupes
- Corrélations intra-groupes
- Distances de Mahalanobis
- Tableau des inerties
- Test de Pillai
- Test de Box
- Fct. discriminante std.
- Fct. discriminante non std.
- Résultats variables
- Résultats individus
- Coord. centres des groupes
- Coeff. fct. clas. quad. (apprentissage)
- Détails classement (apprentissage)
- Matrice de confusion (apprentissage)**
- Sensibilité, Spécificité Décès (app)
- Sensibilité, Spécificité Survie (app)
- Stats - Groupes observés
- Stats - Groupes prévus

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|----|--|--------|-------|--------|-----------|---------|----------|---|
| 1 | | | | | | | | |
| 2 | ANALYSE QUADRATIQUE - MATRICE DE CONFUSION POUR LE JEU D'APPRENTISSAGE | | | | | | | |
| 3 | | | | | | | | |
| 4 | En lignes, les groupes observés | | | | | | | |
| 5 | En colonnes, les groupes prévus | | | | | | | |
| 6 | | | | | | | | |
| 7 | Pourcentage de mal classés : 30,952 % | | | | | | | |
| 8 | Pourcentage de bien classés (exactitude) : 69,048 % | | | | | | | |
| 9 | | | | | | | | |
| 10 | Précision = VP / (VP + FP) | | | | | | | |
| 11 | Rappel = VP / (VP + FN) | | | | | | | |
| 12 | Score F1 = 2 x (Précision x Rappel) / (Précision + Rappel) | | | | | | | |
| 13 | | | | | | | | |
| 14 | | | | | | | | |
| 15 | | Taille | Décès | Survie | Précision | Rappel | Score F1 | |
| 16 | Décès | 424 | 402 | 22 | 0,66889 | 0,94811 | 0,78439 | |
| 17 | Survie | 290 | 199 | 91 | 0,80531 | 0,31379 | 0,45161 | |
| 18 | | | | | | | | |
| 19 | | | | | | | | |
| 20 | | | | | | | | |
| 21 | | | | | | | | |

Rapport Explorateur /



Environ 69 % des passagers sont bien classés par cette analyse et l'aire sous la courbe ROC est proche de 0,7.

Note : Pour comparer les performances de plusieurs méthodes d'analyse, cet exemple est traité dans les six analyses AFD, ADB, KNN, BAYES, ANN et ARBRE.

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 | $\frac{100 * VP}{(VP + FN)}$ |
| Observé B | FP | VN | FP + VN | $\frac{100 * VN}{(VN + FP)}$ |
| Total | VP + FP | FN + VN | VP + FP + VN + FN | |
| % correct | $\frac{100 * VP}{(VP + FP)}$ | $\frac{100 * VN}{(FN + VN)}$ | | $\frac{100 * (VP + VN)}{(VP + VN + FP + FN)}$ |
| | | | | % 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 $2 \times (\text{précision} \times \text{rappel}) / (\text{précision} + \text{rappel})$

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.

| <i>Variable</i> | <i>Contenu</i> |
|----------------------|--|
| fdstdA | Fonctions discriminantes standardisées (apprentissage) |
| fdnstdA | Fonctions discriminantes non standardisées (apprentissage) |
| cindA | Coordonnées des individus (apprentissage) |
| libindA | Libellés des individus (apprentissage) |
| clindA | Classes des individus (apprentissage) |
| cindV | Coordonnées des individus (validation) |
| libindV | Libellés des individus (validation) |
| clindV | Classes des individu (validation) |
| distindA | Distances carrées à l'origine pour les individus (apprentissage) |
| cosindA | Cosinus carrés pour les individus (apprentissage) |
| conindA | Contributions pour les individus (apprentissage) |
| cvarA | Coordonnées des variables (apprentissage) |
| disvarA | Distances carrées à l'origine des variables (apprentissage) |
| cosvarA | Cosinus carrés des variables (apprentissage) |
| convarA | Contributions des variables (apprentissage) |
| seuilA | Seuils (apprentissage) |
| specificiteA | Spécificité (apprentissage) |
| sensibiliteA | Sensibilité (apprentissage) |
| aireA | Aires sous les courbes ROC (apprentissage) |
| seuilV | Seuils (validation) |
| specificiteV | Spécificité (validation) |
| sensibiliteV | Sensibilité (validation) |
| aireV | Aires sous les courbes ROC (validation) |
| si modèle linéaire : | |
| coefcl | Fonctions de classement linéaire (apprentissage) |
| classA | Classement linéaire (apprentissage) |
| classV | Classement linéaire (validation) |
| classP | Classement linéaire (prévision) |
| libindP | Libellés des individus (prévision) |

si modèle quadratique :

| | |
|-----------|--|
| ccq(i) | Constantes classement quadratique groupe i (apprentissage) |
| coefcq(i) | Fonctions de classement quadratique groupe i (apprentissage) |
| classqA | Classement quadratique (apprentissage) |
| classqV | Classement quadratique (validation) |
| classqP | Classement quadratique (prévision) |
| libindP | Libellés des individus (prévision) |