UNIWIN VERSION 10.1.0

REGRESSION SUR COMPOSANTES PRINCIPALES

Révision : 23/11/2024

Définition

La méthode de Régression sur Composantes Principales (RCP) est une technique de régression utile lorsque de fortes colinéarités entre les variables explicatives sont présentes et que l'on ne désire pas utiliser les algorithmes de régression pas à pas pour éliminer les variables corrélées entre elles ou les régressions Ridge ou PLS.

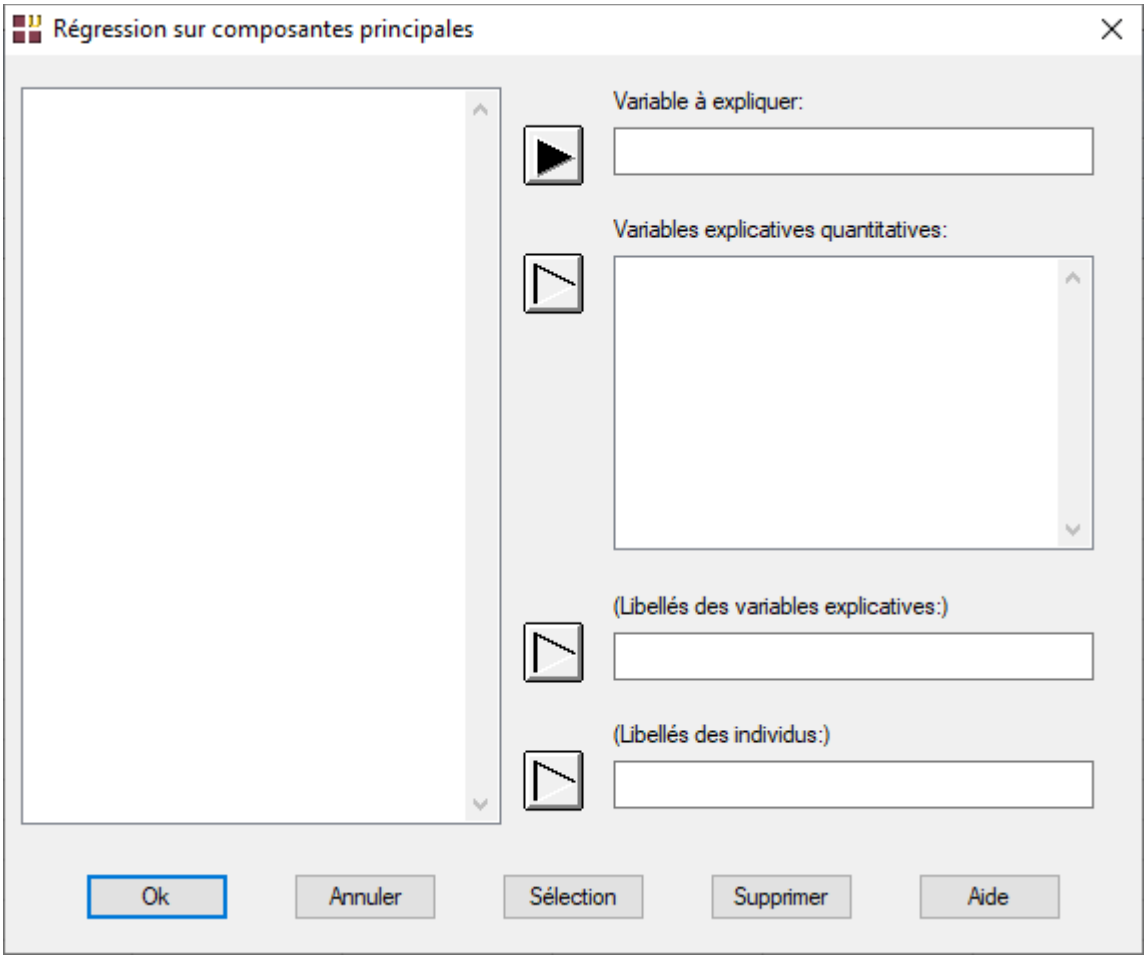
Cette technique utilise à la fois l'Analyse en Composantes Principales (ACP) et la Régression Multiple pour élaborer un modèle dont les coefficients sont stables.

Après l'affichage du tableau et de l'histogramme des inerties, un rapport général de synthèse est proposé contenant notamment les résultats de l'Analyse en Composantes Principales (ACP), les descriptions des différents modèles de régression et le tableau de l'analyse de la variance.

Les graphiques des cercles factoriels, des plans factoriels, des régressions, des composants et des résidus sont également disponibles.

Entrée des données

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



Cette boîte de dialogue permet de définir la variable à expliquer, les variables explicatives quantitatives et les libellés associés ainsi que les libellés des individus.

Les zones de libellés sont optionnelles. Si elles ne sont pas renseignées, UNIWIN génère automatiquement des libellés.

Données manquantes

Dans cette procédure les données manquantes ne sont pas permises pour les variables explicatives mais le sont pour la variable à expliquer.

Les individus pour lesquels la valeur de la variable à expliquer est manquante définissent le jeu de prévision.

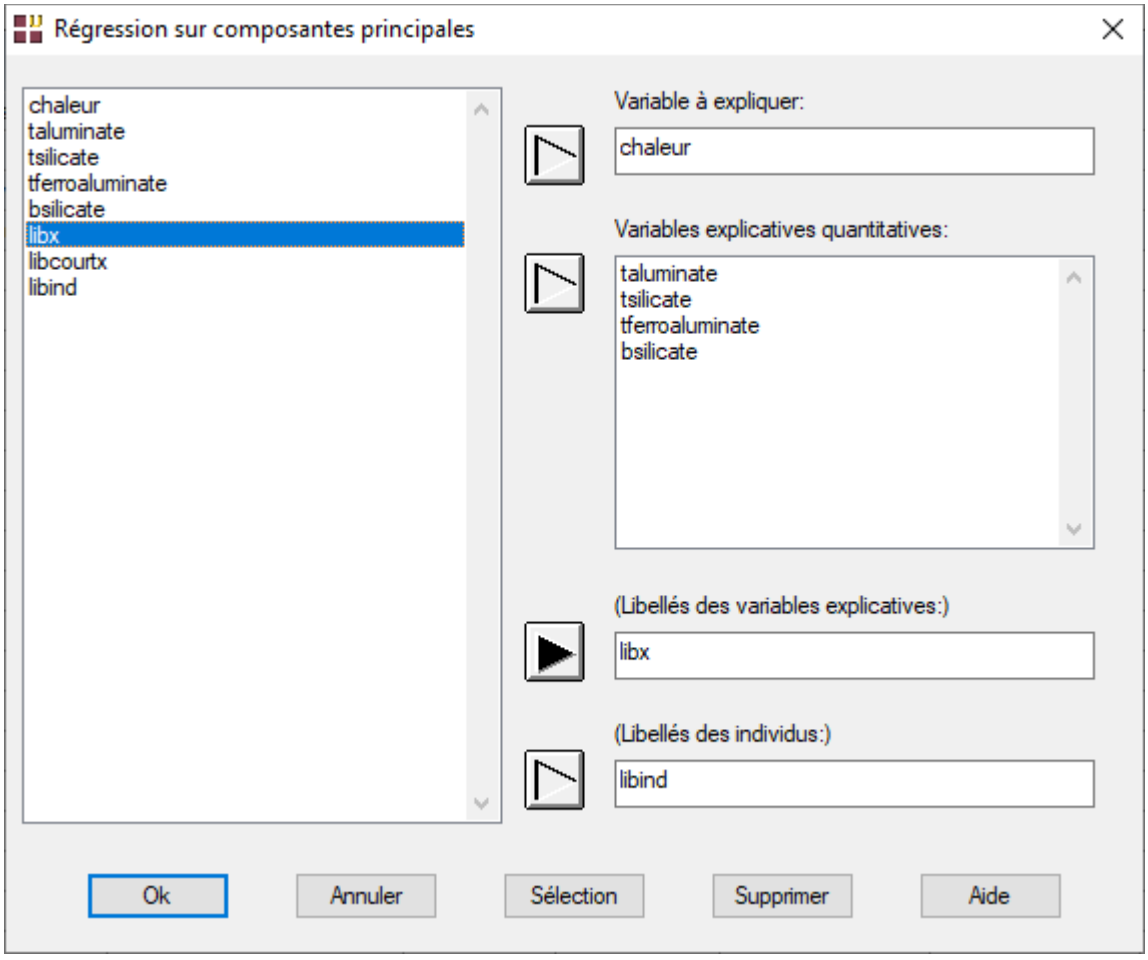
Exemple 1 : Fichier Ciment

Pour illustrer cette procédure, nous utiliserons le fichier CIMENT contenant les données utilisées par Hald et publiées dans 'Industrial and Engineering Chemistry' (24, 1932, 1207-14, Table I, par H. Woods, H. H. Seymour et H. R. Starke : « Effect of Composition of Portland Cement on Heat Evolved during Hardening »).

Ce fichier contient les informations suivantes :

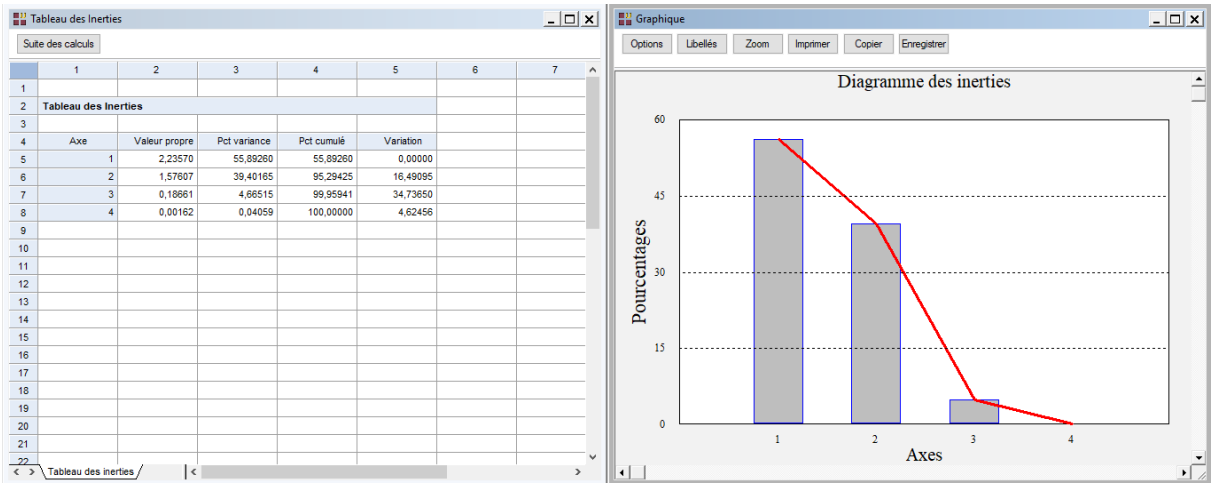
chaleur	Chaleur en calories par gramme
taluminate (aluminate tricalcique)	Quantité en %
tsilicate (silicate tricalcique)	Quantité en %
tferroaluminate (ferroaluminate tetracalcique)	Quantité en %
bsilicate (silicate bicalcique)	Quantité en %
libx	Libellés longs des variables explicatives (X)
libcourtx	Libellés courts des variables explicatives (X)
libind	Libellés des individus

Renseignons la boîte de dialogue comme montré ci-dessous.

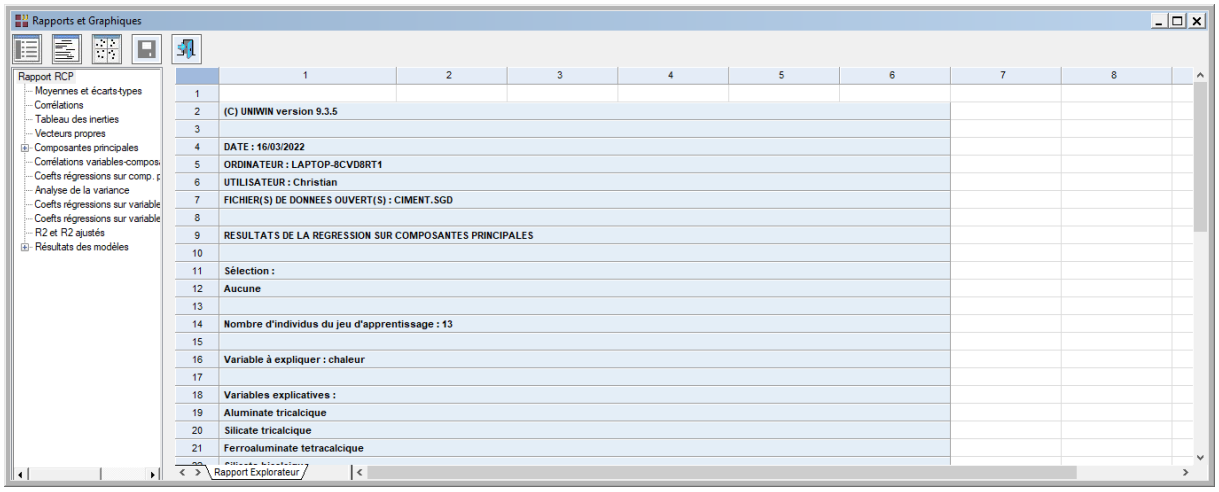


Cliquons sur le bouton Ok pour exécuter le traitement de l'analyse.

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.

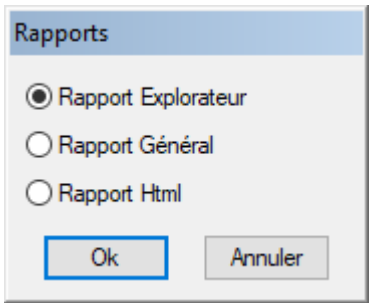


Cliquons sur le bouton 'Suite des calculs'. Après quelques instants, l'écran suivant s'affiche :

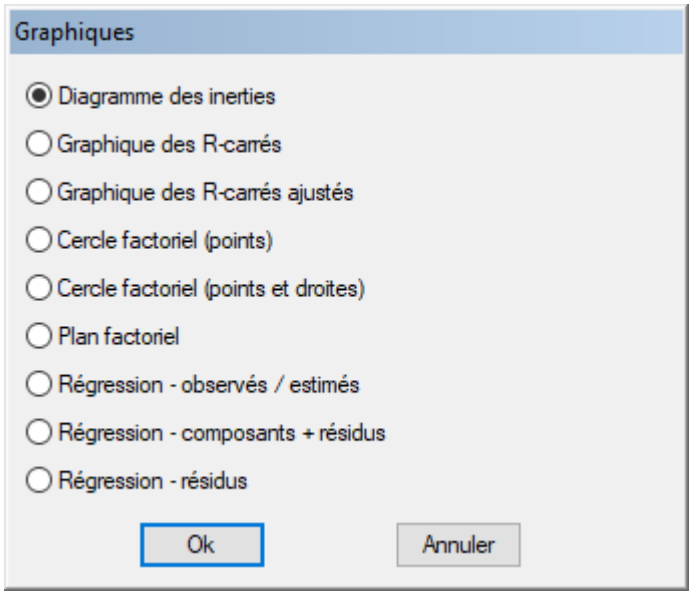


La barre d'outils 'Rapports et Graphiques' permet par l'icône 'Données'  de rappeler la boîte de dialogue d'entrée des données.

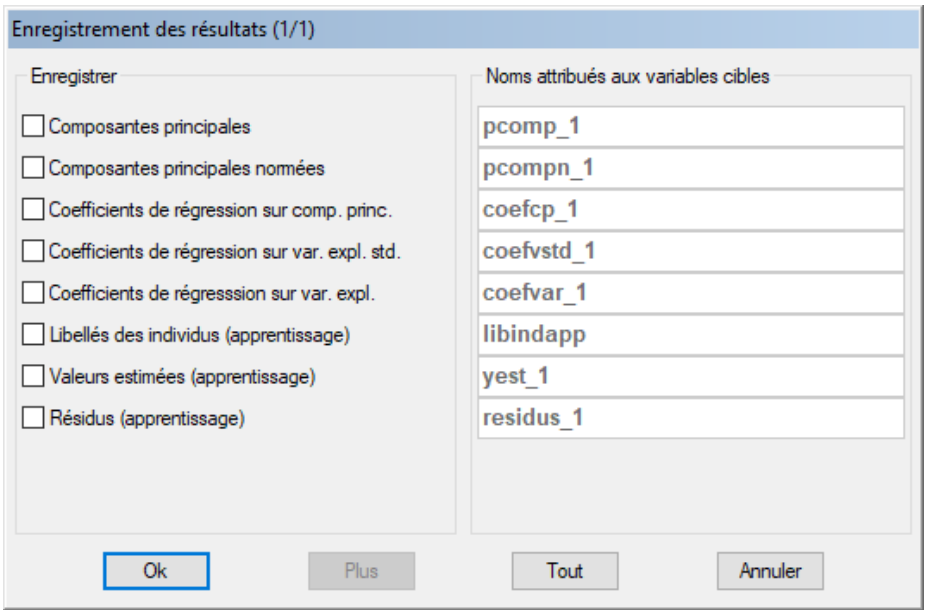
L'icône 'Rapports'  affiche la boîte de dialogue des options pour les rapports :



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



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



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

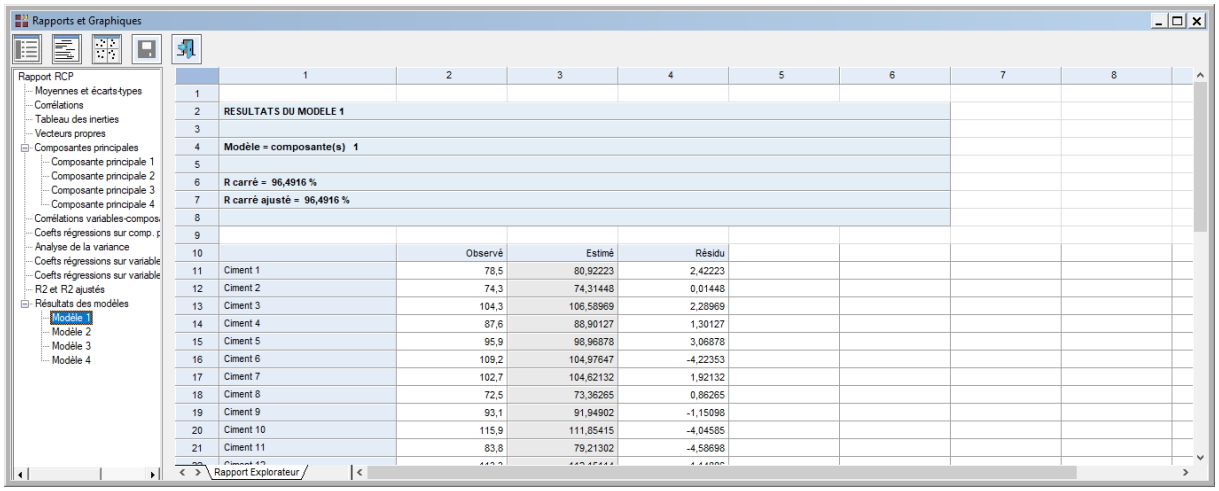
L’icône ‘Quitter’  permet de quitter l’analyse.

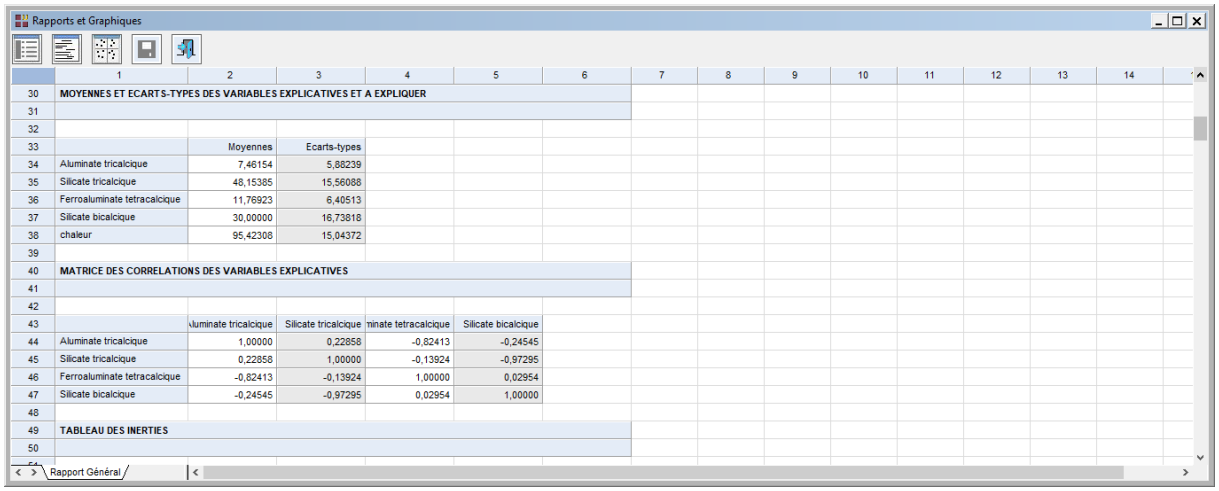
L’option Rapports

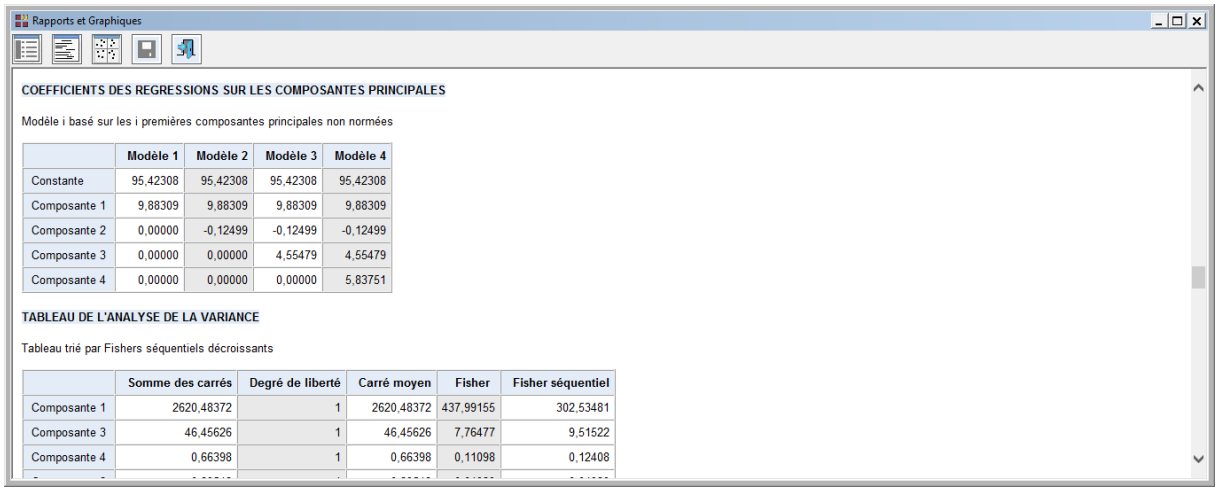
Cette option permet d’obtenir le rapport à l’écran sous la forme d’un explorateur, d’un tableur ou au format HTML.

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 analyse : Explorateur, Général, HTML.







Ce rapport nous donne les informations suivantes :

◊ Moyennes et écarts-types des variables explicatives et de la variable à expliquer

◊ Matrice des corrélations entre les variables explicatives

◊ Tableau des inerties issu de l'analyse en composantes principales des variables explicatives

◊ Vecteurs propres issus de l'analyse en composantes principales

◊ Coordonnées des individus sur les composantes principales et composantes principales normées

◊ Corrélations des variables explicatives et à expliquer avec les composantes principales normées

◊ Coefficients des régressions sur les composantes principales (modèle à 1 composante, modèle à 2 composantes, …)

◊ Tableau de l'analyse de la variance trié par valeurs décroissantes des Fishers séquentiels (sommes des carrés, degrés de liberté, carrés moyens, Fishers, Fishers séquentiels).

Les Fishers séquentiels sont les Fishers obtenus lorsque les composantes principales sont ajoutées une à une dans l'ordre des contributions respectives.

◊ Coefficients des régressions sur les variables explicatives centrées et réduites.

Ces coefficients sont issus du modèle utilisant la composante ayant le plus grand Fisher séquentiel, puis du modèle utilisant les deux composantes ayant les deux plus grands Fishers séquentiels, etc….

◊ Coefficients des régressions sur les variables.

Ces coefficients sont issus du modèle utilisant la composante ayant le plus grand Fisher séquentiel, puis du modèle utilisant les deux composantes ayant les deux plus grands Fishers séquentiels, etc….

◊ R-carrés et R-carrés ajustés.

◊ Résultats des modèles (R-carré, Valeurs observées, Valeurs estimées, Résidus)

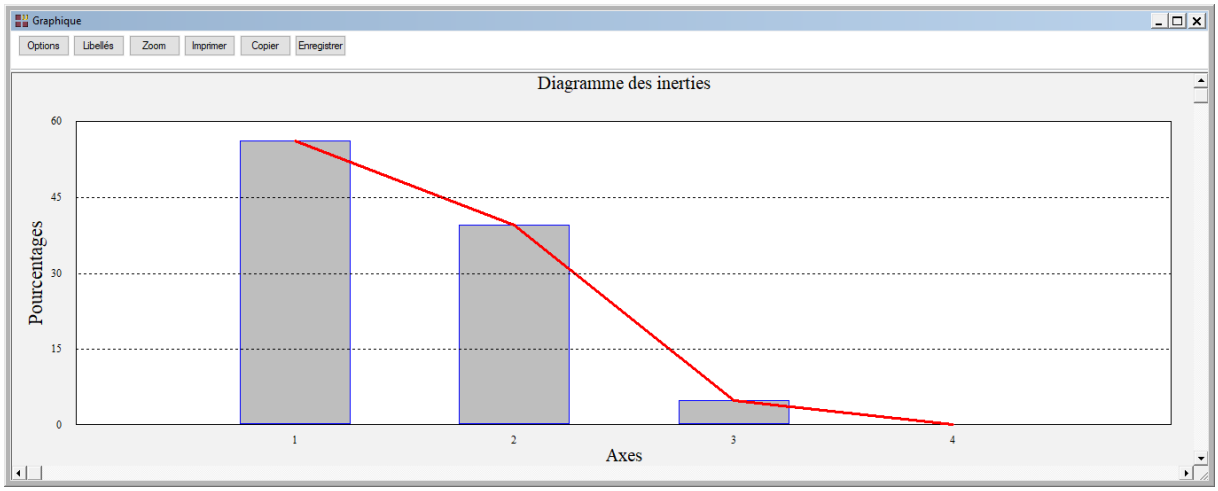
Ces résultats sont issus du modèle utilisant la composante ayant le plus grand Fisher séquentiel, puis du modèle utilisant les deux composantes ayant les deux plus grands Fishers séquentiels, etc….

L'option Graphiques

Cette option permet d'obtenir divers graphiques pour notre analyse.

- L'option Diagramme des inerties

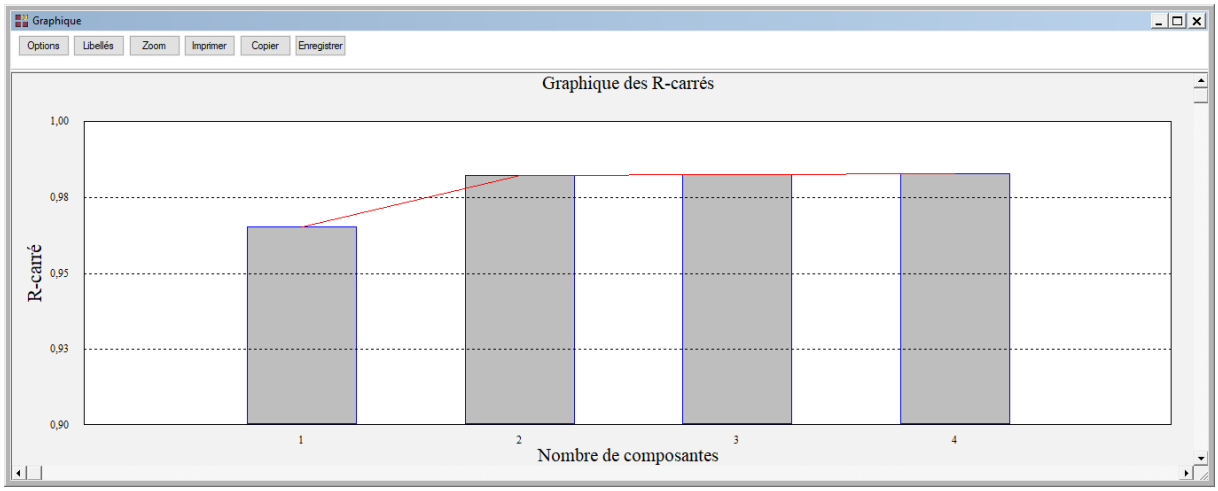
Ce graphique affiche les pourcentages d'inertie pour chacune des composantes principales.

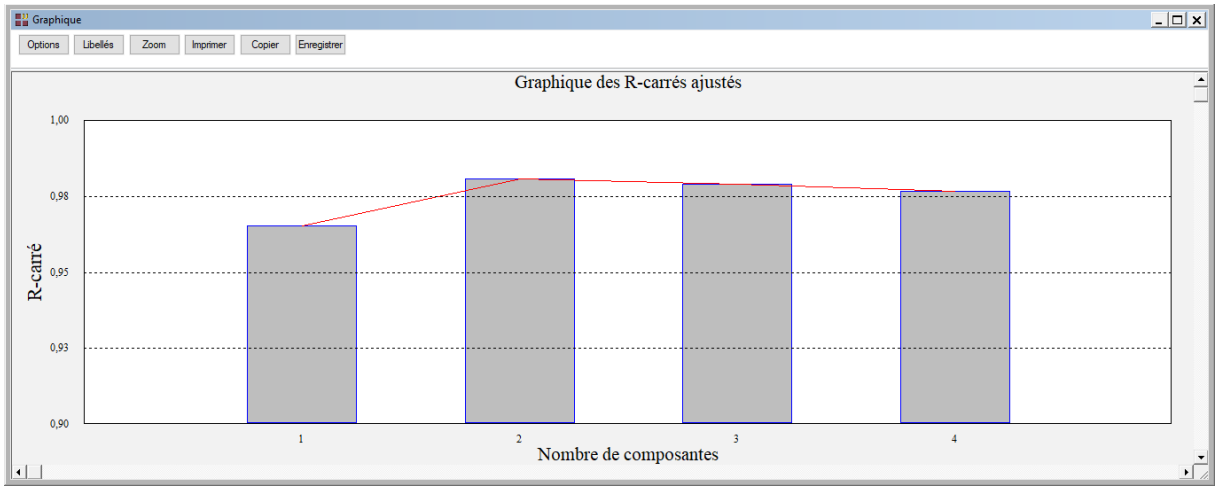


- Les options Graphique des R-carrés et Graphique des R-carrés ajustés

Ces deux graphiques montrent les évolutions des R2 et R2 ajustés en fonction du nombre de composantes dans le modèle.

Le R2 ajusté prenant en compte la complexité du modèle, il est préférable de l'utiliser pour comparer des modèles de régression dont les nombres de variables explicatives (ici nos composantes principales) sont différents.





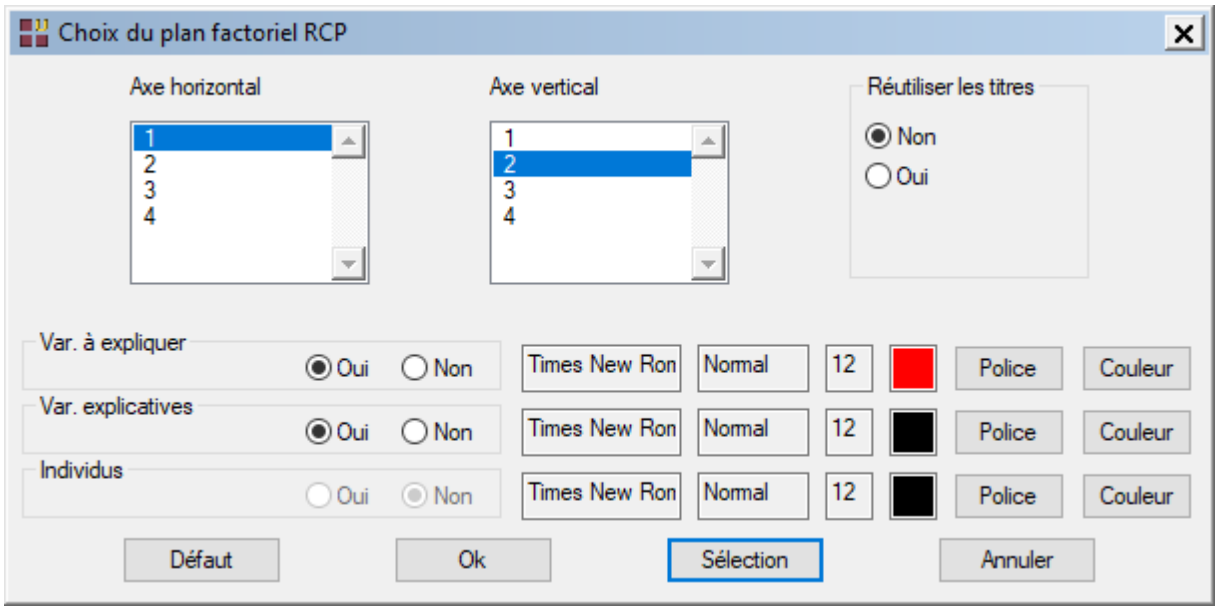
- Les options Cercle factoriel (points) et Cercle factoriel (points et droites)

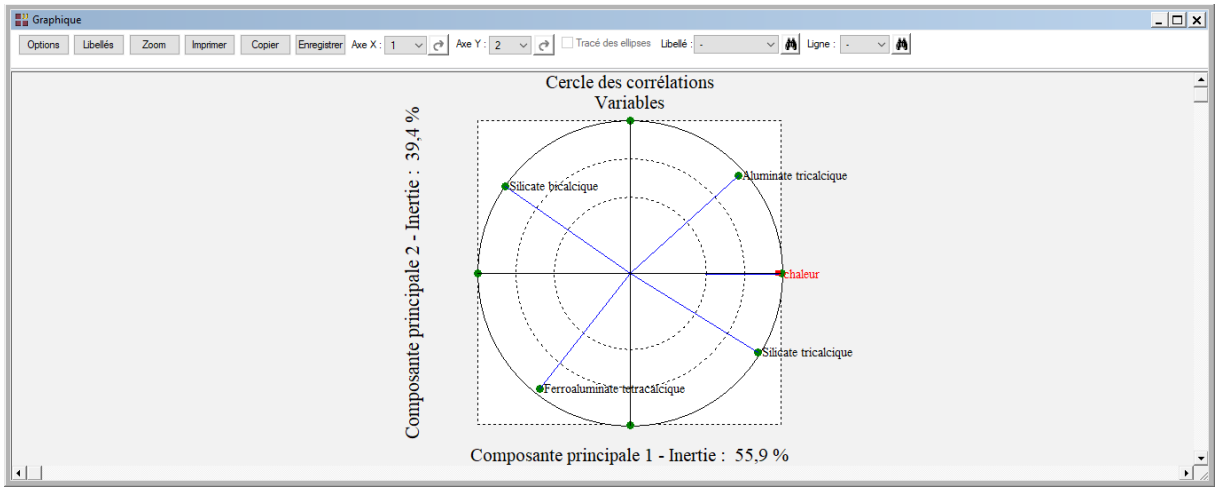
Ces options permettent d'afficher le cercle de corrélations des variables X et de la variable Y et de choisir si on désire tracer les droites reliant les points à l'origine du cercle.

L'option sans ces droites (points) est utile lorsqu'il y a un grand nombre de variables représentées. Choisissons les variables de base avec droites.

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.

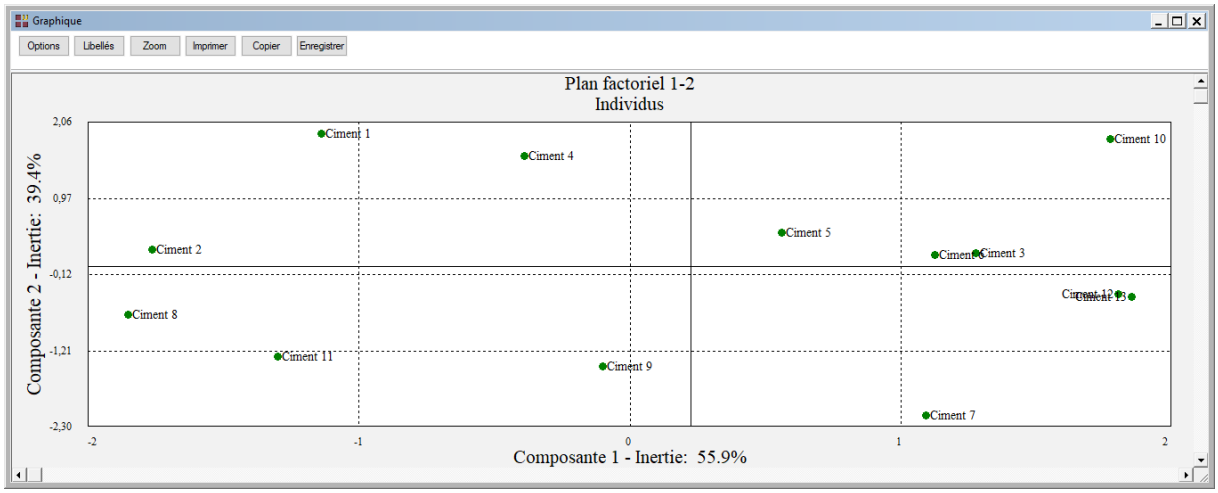




- L'option Plan factoriel

Cette option permet d'afficher des plans factoriels des individus.

Une boîte de dialogue permettant de choisir le plan factoriel s'affiche alors. Elle permet également de préciser si l'on désire afficher les libellés des individus, de choisir la couleur et la police, 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.

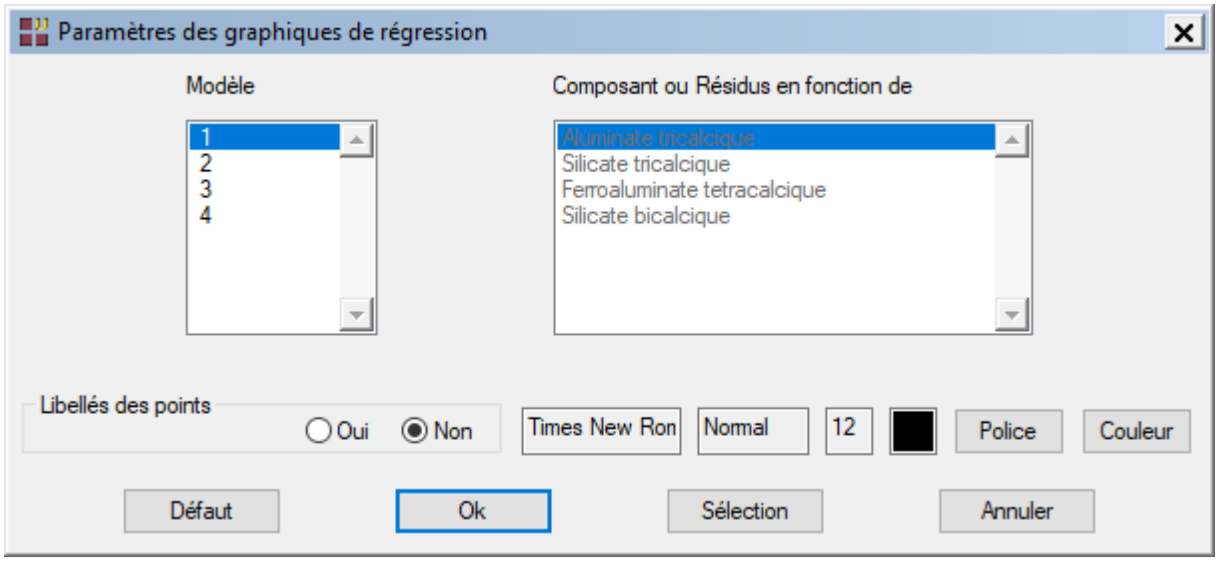


- Les options Graphiques de Régression

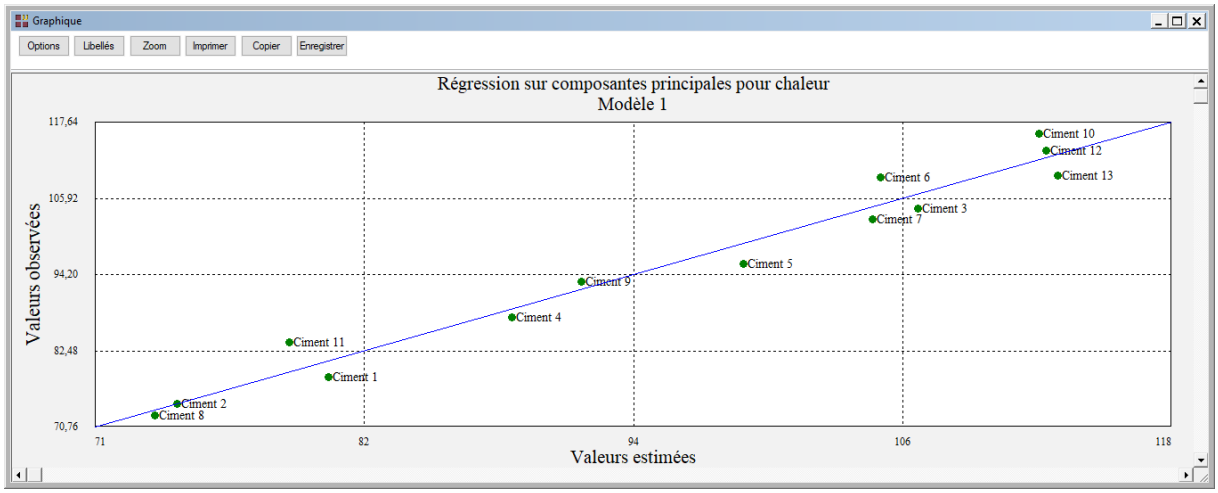
Ces options permettent d'afficher trois graphiques :

◊ Valeurs observées et valeurs estimées
◊ Composants + Résidus (variable explicative et résidus)
◊ Résidus

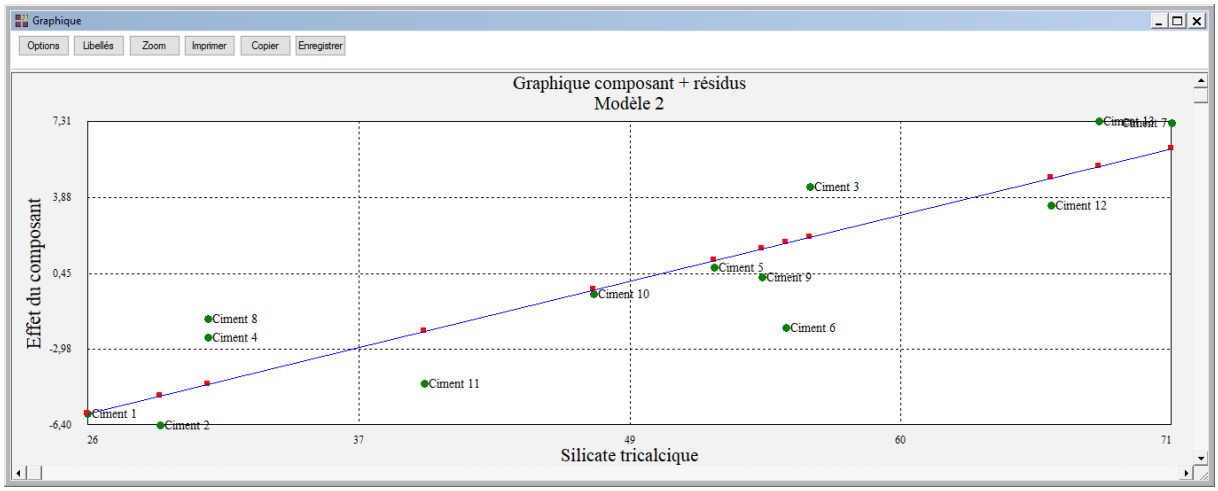
Pour chacune des options, une boîte de dialogue s'affiche.



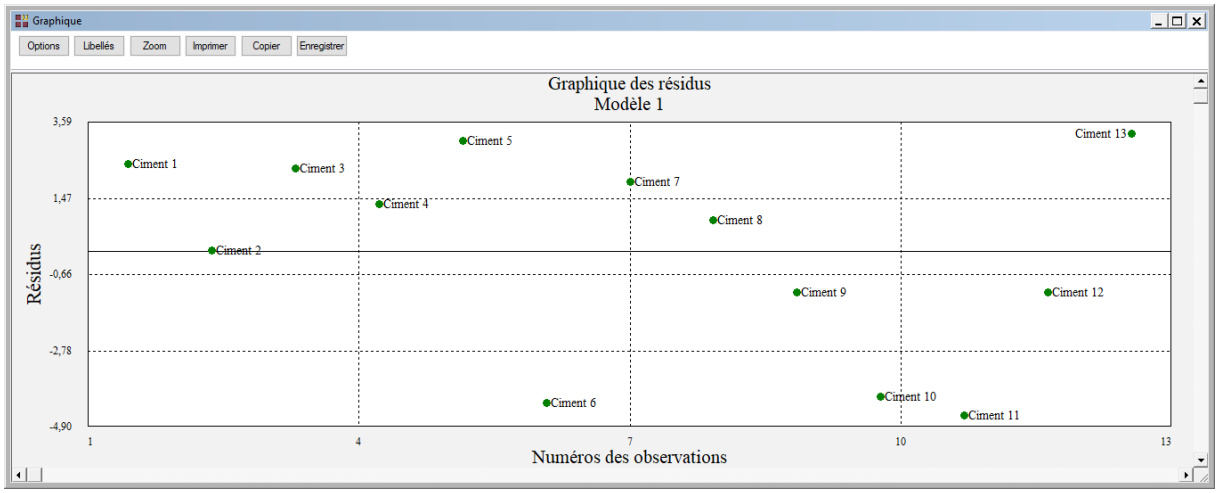
Pour la première option, elle permet de sélectionner le modèle à représenter graphiquement et d'indiquer si on désire afficher les libellés des individus.



Elle permet également pour la deuxième option, de préciser par rapport à quelle variable explicative on désire afficher les composants.



Pour la troisième option, elle permet de préciser si on désire afficher les résidus par rapport aux numéros des individus (observations), aux valeurs estimées ou par rapport à une variable explicative.



Une rapide interprétation des résultats

Les résultats de cette analyse montrent clairement que seules deux des quatre composantes principales sont utiles pour bâtir le modèle de régression.

A noter que dans toutes les équations, toutes les variables explicatives sont présentes. Aucune des variables n'a été éliminée par la procédure à aucun moment.

Cette technique est fréquemment utilisée notamment en physique, chimie, ingénierie, biologie et en sciences sociales.

Exemple 2 : Fichier Octane

Pour illustrer ce deuxième exemple, nous utiliserons le fichier OCTANE (données de Cornell).

Ce fichier a été complété, pour les besoins de cet exemple, par deux lignes supplémentaires pour lesquelles les valeurs de la variable à expliquer sont à prévoir.

Ce fichier contient les données suivantes :

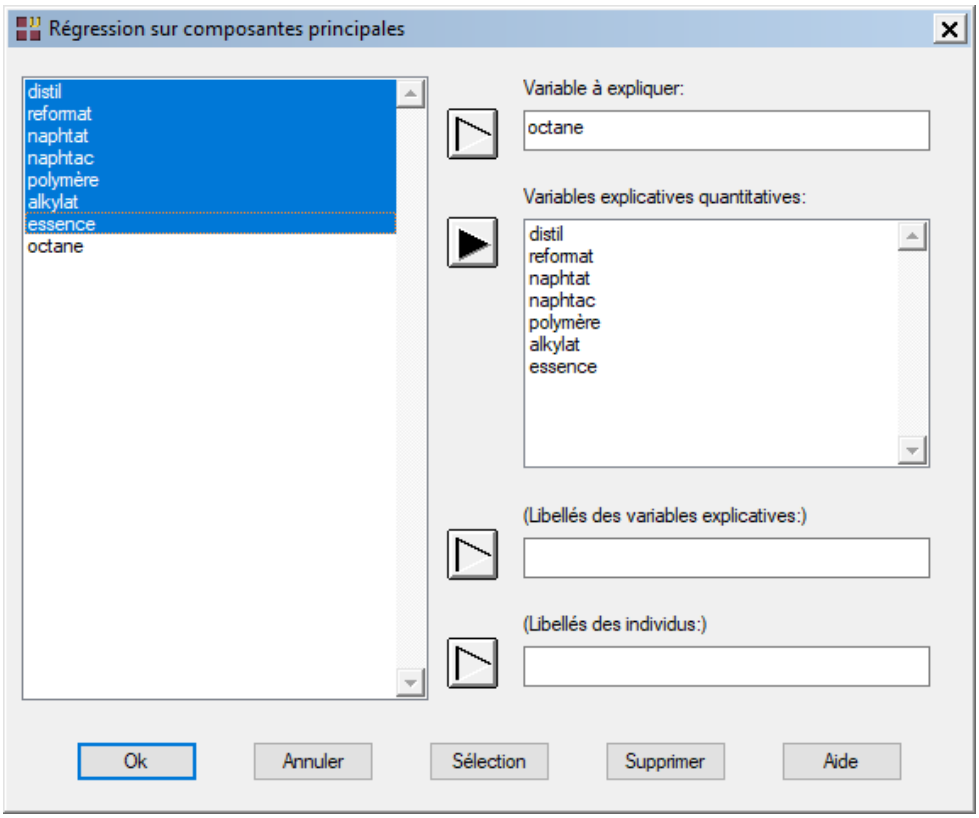
distil	Distillation directe
reformat	Réformat
naphtat	Naphta de craquage thermique
naphtac	Naphta de craquage catalytique
polymère	Polymère
alkylat	Alkylat
essence	Essence naturelle
octane	indice d'octane

La variable octane est la variable à expliquer.

Les lignes pour lesquelles la variable à expliquer n'est pas renseignée définissent le jeu de prévision.

Renseignons la boîte de dialogue comme montré ci-dessous.

Des exemples des résultats obtenus sont montrés ci-après.


Régression sur composantes principales
distil
reformat
naphtat
naphtac
polymère
alkylat
essence
octane
Variable à expliquer:
octane
Variables explicatives quantitatives:
distil
reformat
naphtat
naphtac
polymère
alkylat
essence
(Libellés des variables explicatives:)
(Libellés des individus:)
Ok
Annuler
Sélection
Supprimer
Aide

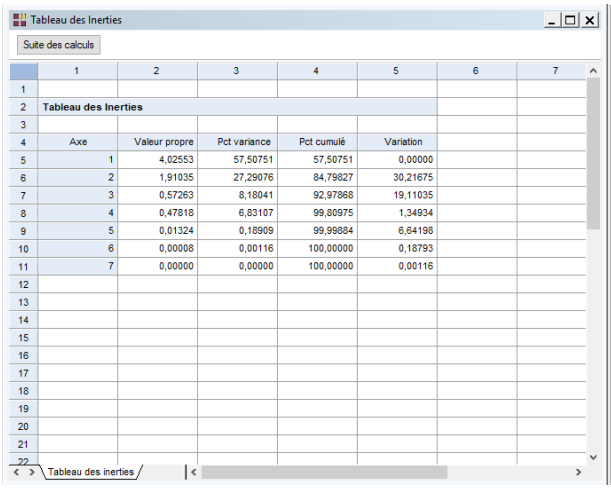
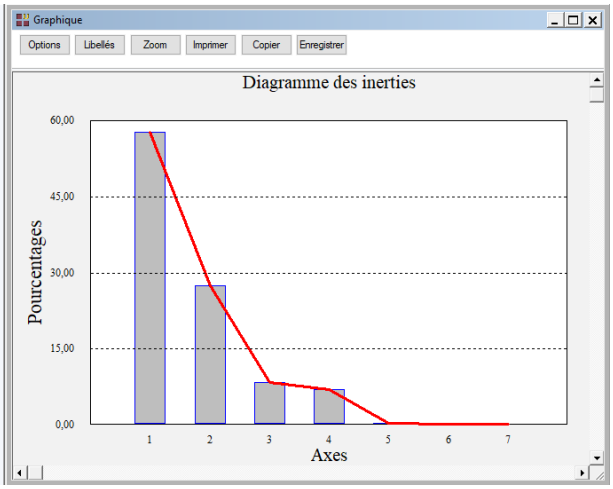
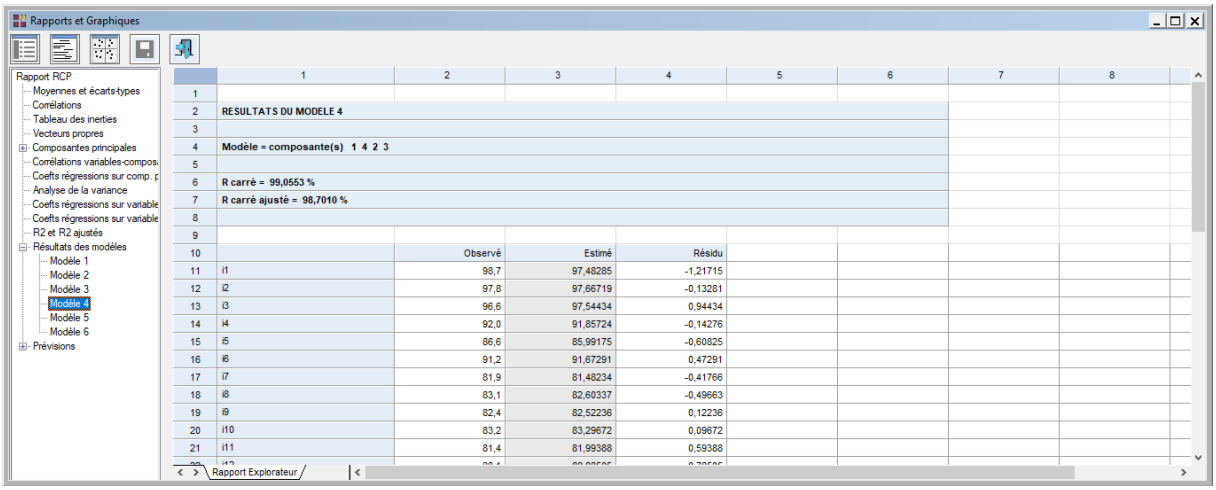
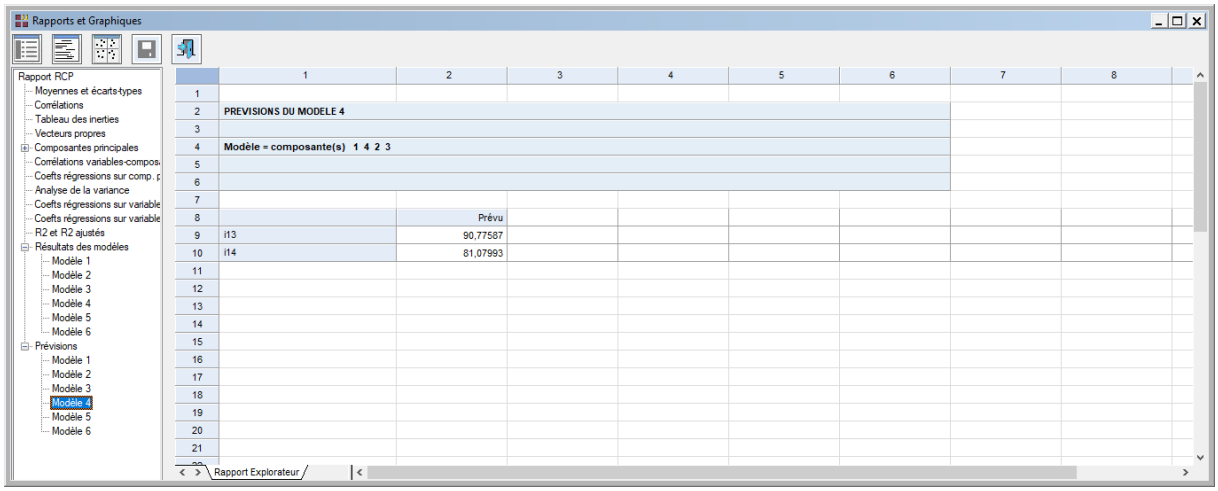
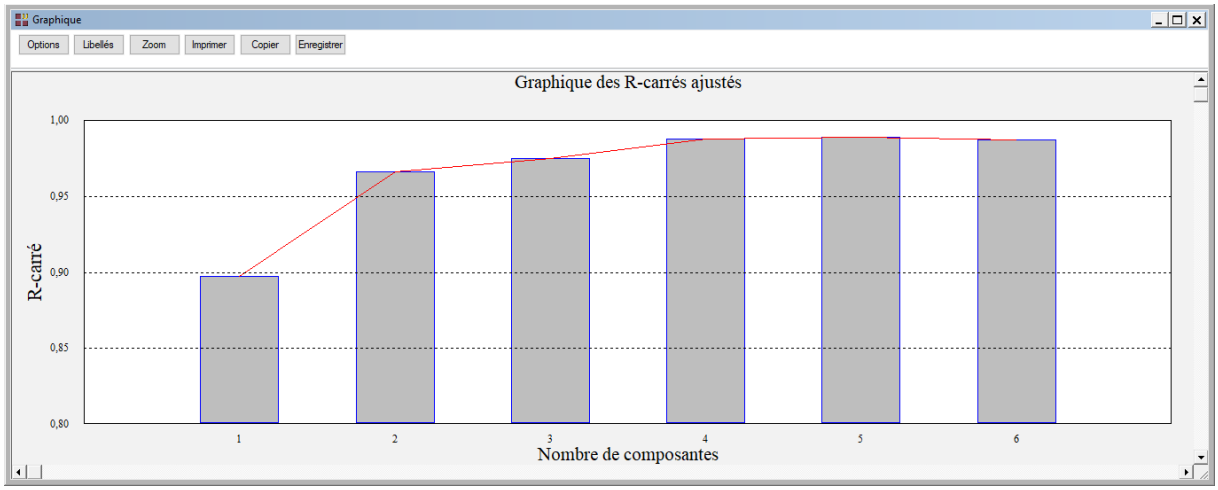


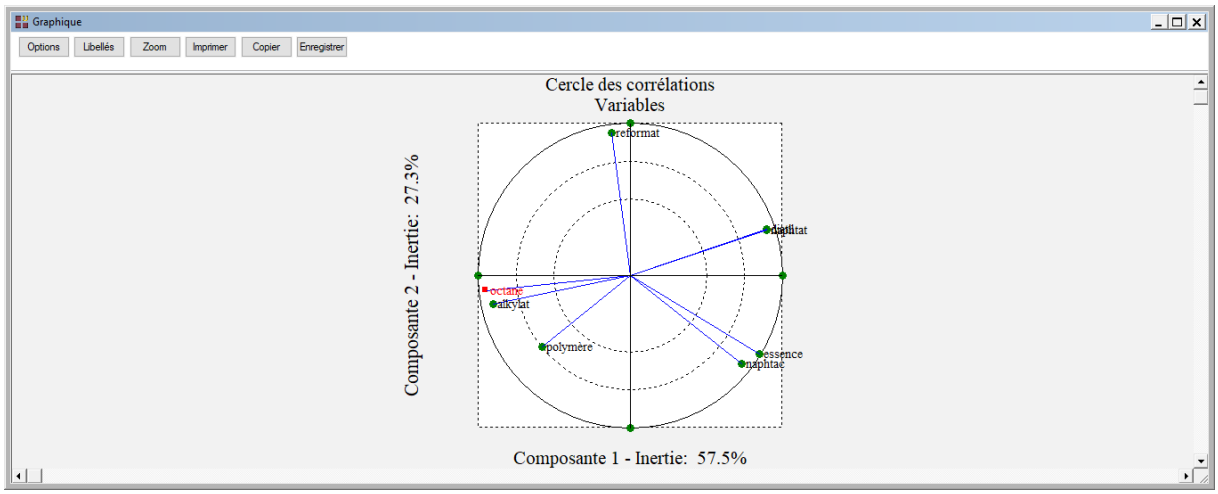
Tableau des Inerties
Suite des calculs
Tableau des Inerties
Axe | Valeur propre | Pct variance | Pct cumulé | Variation
1 | 4,02553 | 57,50751 | 57,50751 | 0,00000
2 | 1,91035 | 27,29076 | 84,79827 | 30,21675
3 | 0,57263 | 8,18041 | 92,97868 | 19,11035
4 | 0,47818 | 6,83107 | 99,80975 | 1,34934
5 | 0,01324 | 0,18909 | 99,99884 | 6,64198
6 | 0,00008 | 0,00116 | 100,00000 | 0,18793
7 | 0,00000 | 0,00000 | 100,00000 | 0,00116
Tableau des inerties

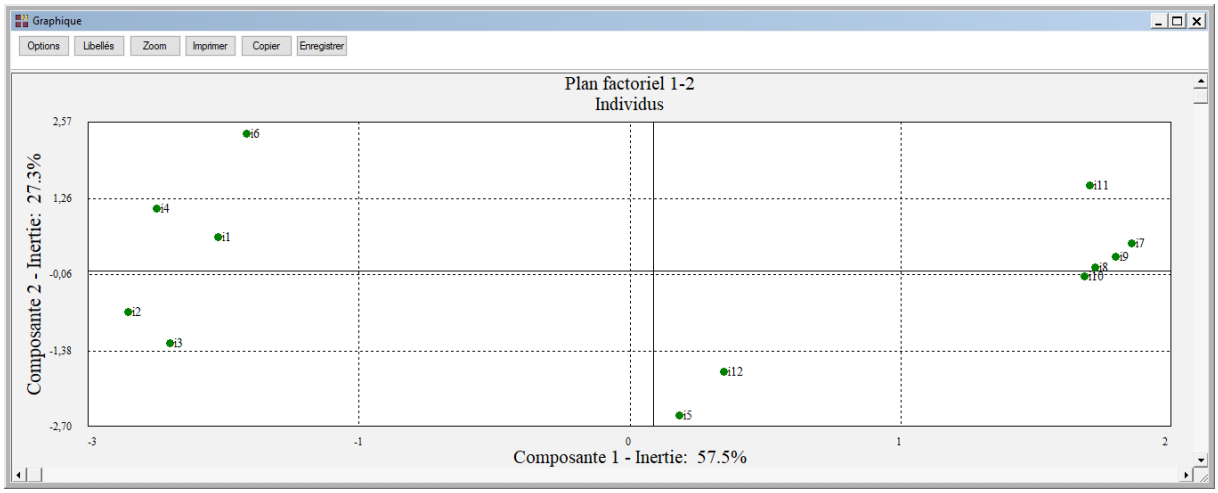

Graphique
Options
Libellés
Zoom
Imprimer
Copier
Enregistrer
Diagramme des inerties
Pourcentages
60,00
45,00
30,00
15,00
0,00
1
2
3
4
5
6
7
Axes

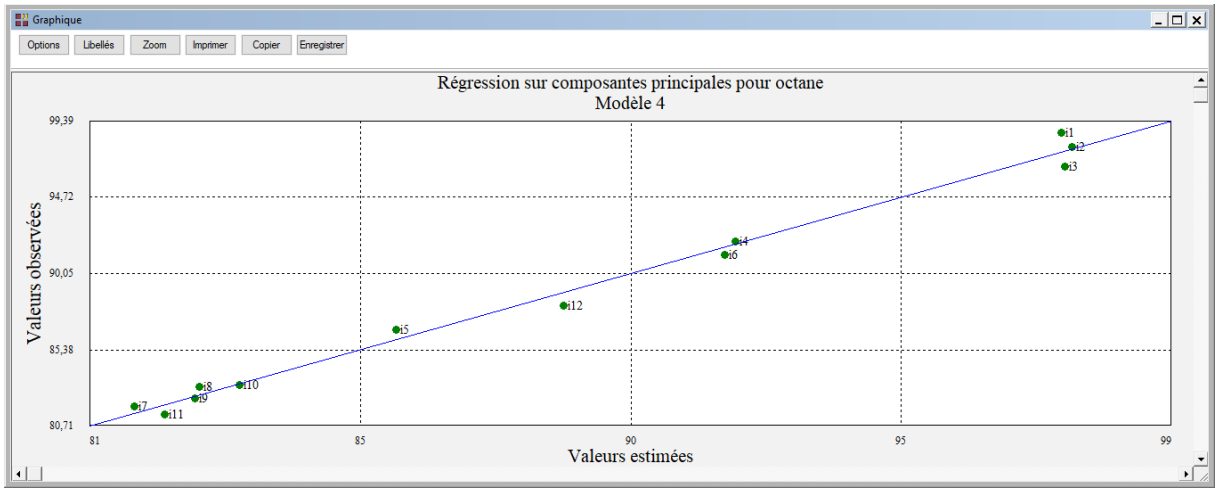

Rapports et Graphiques
Rapport RCP
Moyennes et écarts-types
Corrélations
Tableau des inerties
Vecteurs propres
Composantes principales
Corrélations variables-compos.
Coefts régressions sur comp. p
Analyse de la variance
Coefts régressions sur variable
Coefts régressions sur variable
R2 et R2 ajustés
Résultats des modèles
Modèle 1
Modèle 2
Modèle 3
Modèle 4
Modèle 5
Modèle 6
Prévisions
RESULTATS DU MODELE 4
Modèle = composante(s) 1 4 2 3
R carré = 99,0553 %
R carré ajusté = 98,7010 %
Observé | Estimé | Résidu
i1 | 98,7 | 97,48285 | -1,21715
i2 | 97,8 | 97,66719 | -0,13281
i3 | 96,6 | 97,54434 | 0,94434
i4 | 92,0 | 91,85724 | -0,14276
i5 | 86,6 | 85,99175 | -0,60825
i6 | 91,2 | 91,67291 | 0,47291
i7 | 81,9 | 81,48234 | -0,41766
i8 | 83,1 | 82,60337 | -0,49663
i9 | 82,4 | 82,52236 | 0,12236
i10 | 83,2 | 83,29672 | 0,09672
i11 | 81,4 | 81,99388 | 0,59388
Rapport Explorateur

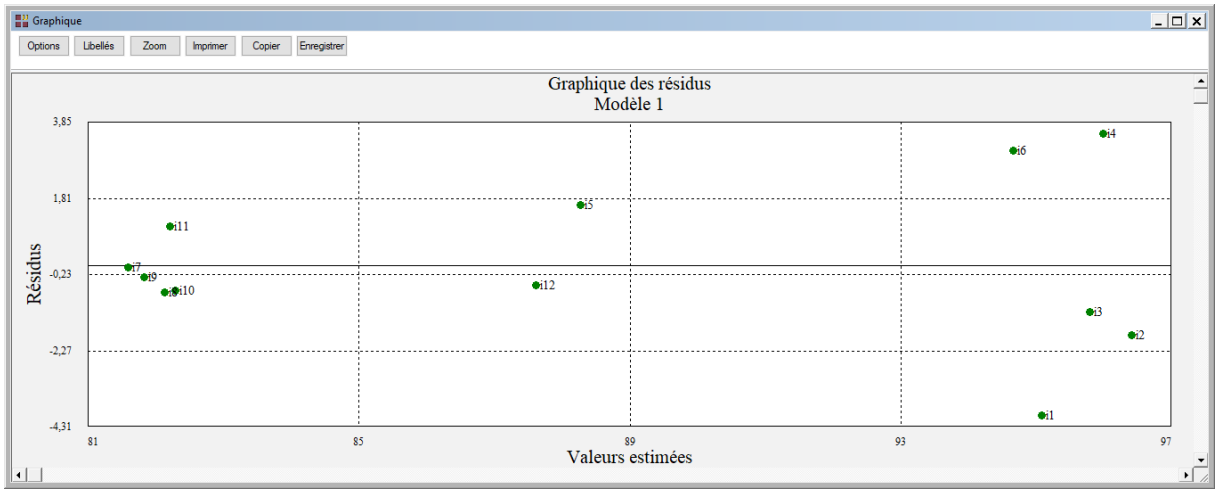

Rapports et Graphiques
Rapport RCP
Moyennes et écarts-types
Corrélations
Tableau des inerties
Vecteurs propres
Composantes principales
Corrélations variables-compos.
Coefts régressions sur comp. p
Analyse de la variance
Coefts régressions sur variable
Coefts régressions sur variable
R2 et R2 ajustés
Résultats des modèles
Modèle 1
Modèle 2
Modèle 3
Modèle 4
Modèle 5
Modèle 6
Prévisions
Modèle 1
Modèle 2
Modèle 3
Modèle 4
Modèle 5
Modèle 6
PREVISIONS DU MODELE 4
Modèle = composante(s) 1 4 2 3
Prévu
i13 90,77587
i14 81,07993
Rapport Explorateur


Graphique
Options
Libellés
Zoom
Imprimer
Copier
Enregistrer
Graphique des R-carrés ajustés
R-carré
1,00
0,95
0,90
0,85
0,80
1
2
3
4
5
6
Nombre de composantes


Graphique
Options
Libellés
Zoom
Imprimer
Copier
Enregistrer
Cercle des corrélations
Variables
reformat
octane
alkylat
polymère
essence
naphtae
Composante 2 - Inertie: 27.3%
Composante 1 - Inertie: 57.5%


Graphique
Options
Libellés
Zoom
Imprimer
Copier
Enregistrer
Plan factoriel 1-2
Individus
Composante 2 - Inertie: 27.3%
Composante 1 - Inertie: 57.5%
2,57
1,26
-0,06
-1,38
-2,70
-3
-1
0
1
2
i1
i2
i3
i4
i5
i6
i7
i8
i9
i10
i11
i12


Graphique
Options
Libellés
Zoom
Imprimer
Copier
Enregistrer
Régression sur composantes principales pour octane
Modèle 4
Valeurs observées
Valeurs estimées
99,39
94,72
90,05
85,38
80,71
81
85
90
95
99
i1
i2
i3
i4
i5
i6
i7
i8
i9
i10
i11
i12


Graphique
Options
Libellés
Zoom
Imprimer
Copier
Enregistrer
Graphique des résidus
Modèle 1
Résidus
Valeurs estimées
3,85
1,81
-0,23
-2,27
-4,31
81
85
89
93
97
i1
i2
i3
i4
i5
i6
i7
i8
i9
i10
i11
i12

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*
coefcp	Coefficients des régressions sur les composantes principales
coefvar	Coefficients des régressions avec constante sur les variables
coefvstd	Coefficients des régressions avec constante sur les variables explicatives centrées et réduites
libindapp	Libellés des individus (jeu d'apprentissage)
libindprev	Libellés des individus (jeu de prévision)
pcomp	Composantes principales
pcompn	Composantes principales normées
residus	Résidus des régressions
yest	Valeurs estimées de la variable à expliquer
yprev	Valeurs prévues de la variable à expliquer