UNIWIN VERSION 9.7.0

REGRESSION PLS

Révision : 02/09/2023

Définition

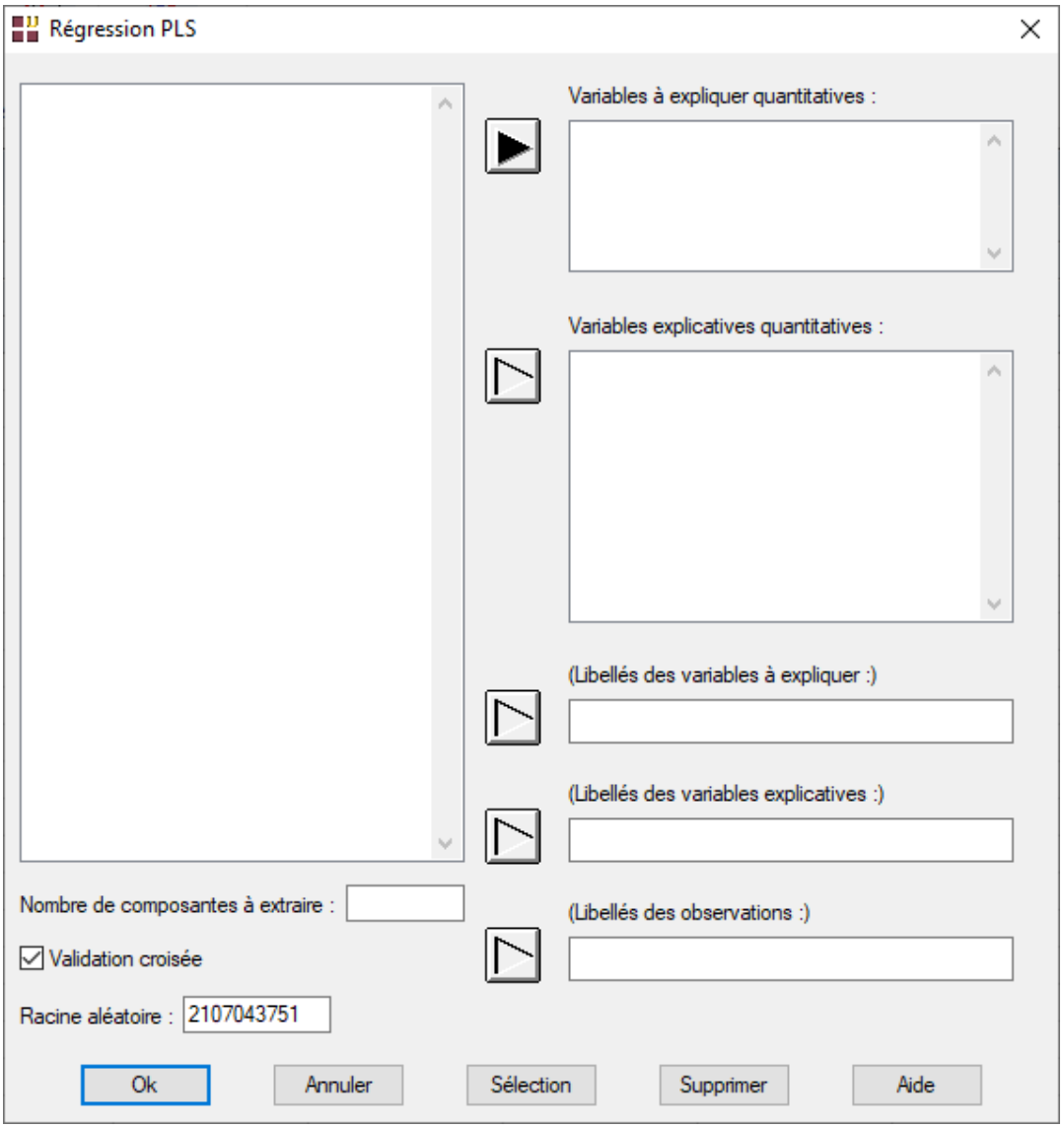
La méthode Régression PLS (partial least squares ou moindres carrés partiels) est conçue pour ajuster un modèle statistique reliant un ensemble de variables explicatives X à une variable à expliquer Y (PLS1) ou à plusieurs variables à expliquer Y (PLS2). La procédure est utile lorsqu'il y a de nombreux X et que le but principal est de prévoir simultanément les variables Y. Elle est recommandée dans le cas où un grand nombre de variables X est utilisé, lorsqu'il y a de fortes colinéarités entre ces variables X, lorsque le nombre de variables X est supérieur au nombre d'observations et lorsqu'il y a des données manquantes. La méthode PLS est notamment largement utilisée par les chimistes et les chimiométriciens pour l'étalonnage en spectrométrie.

Un rapport général de synthèse est proposé contenant notamment les PRESS, R2, Q2, R2X, R2Y, les poids w et w*, les scores t et u, les poids des variables X et Y, les corrélations des variables avec les composantes, les valeurs observées, ajustées et résidus du modèle, les distances au modèle en X et Y, les T2 de Hotelling et les VIP. Les graphiques des R2X, R2Y, Q2, des coefficients standardisés, des cercles des corrélations, des plans factoriels, des poids des variables, des T2, des distances au modèle en X et Y, des VIP, des valeurs observées vs estimées et des résidus sont également disponibles.

La procédure implémentée est basée sur le package R 'plsdepot'.

Entrée des données

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



Cette boîte de dialogue permet de définir la ou les variables à expliquer, les variables explicatives quantitatives, les libellés optionnels des variables à expliquer et explicatives, les libellés optionnels des observations, le nombre de composantes à extraire, si la validation croisée est mise en œuvre on non et la racine aléatoire pour la validation croisée.

Les données à expliquer et explicatives sont automatiquement centrées et réduites.

Si le nombre de composantes à extraire n'est pas précisé, il est déterminé par la validation croisée. Un minimum de deux composantes est extrait. A noter que la validation croisée n'est pas mise en œuvre s'il y a des données manquantes dans les variables explicatives ou si le nombre d'observations est inférieur à 10.

Données manquantes ou non sélectionnées

Les données manquantes ne sont pas autorisées dans les variables à expliquer. Les lignes contenant une ou des données manquantes ne participent donc pas aux calculs et définissent le jeu de prévision. Les lignes non sélectionnées définissent le jeu de validation. Les données manquantes ne sont pas autorisées dans les variables explicatives en régression PLS2.

Exemple 1 : Fichier Octane (PLS1)

Pour illustrer ce premier exemple, nous utiliserons le fichier OCTANE (données de Cornell augmentées de deux observations pour lesquelles la valeur d'octane est non présente). 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 Y à expliquer.

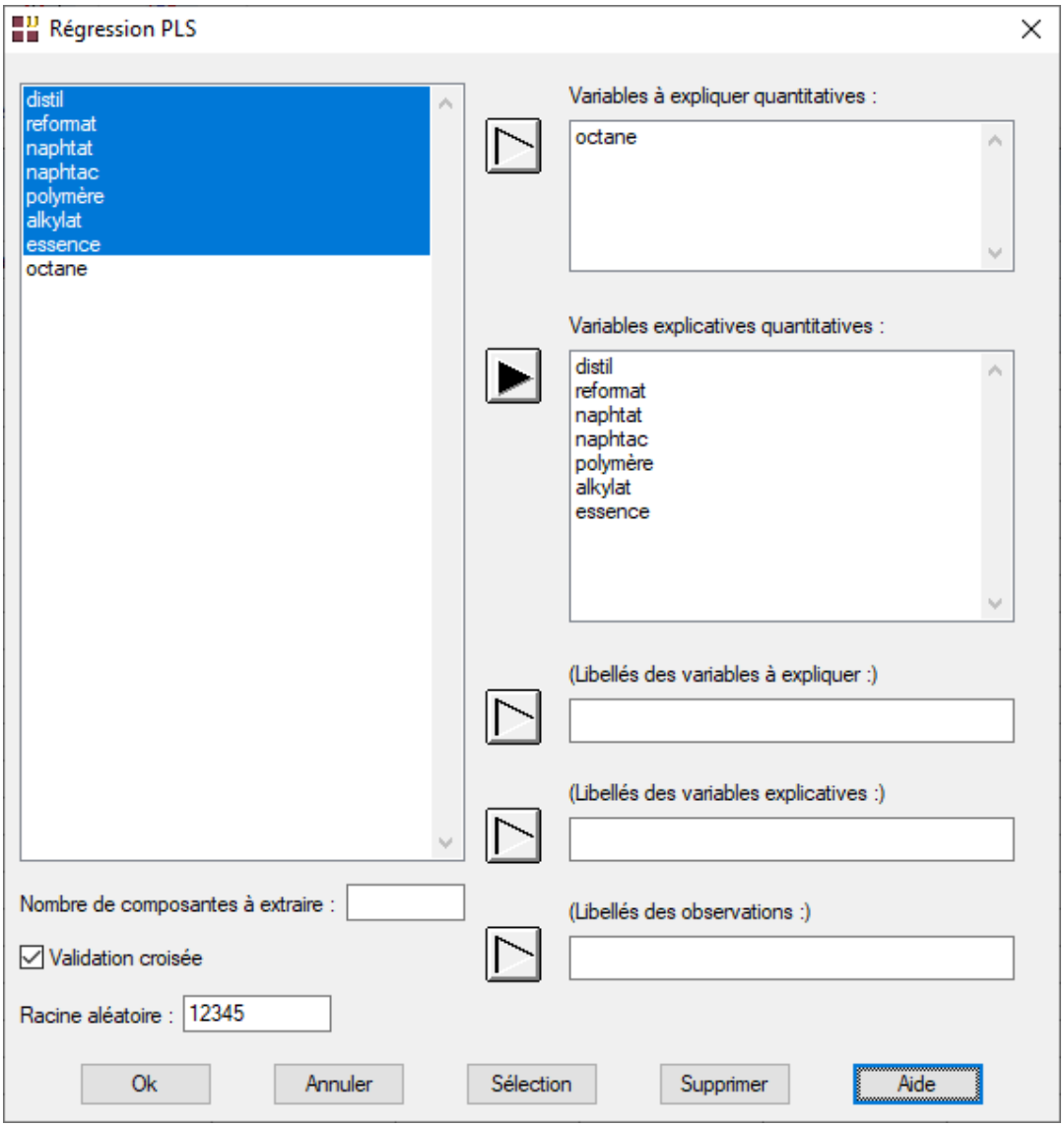
Données

	distil	reformat	naphtat	naphtac	polymère	alkylat	essence	octane
	Distillation directe	Réformat	Naphta craquage thermique	Naphta craquage cataly.	Polymère	Alkylat	Essence naturelle	
	Type = Numérique	Type = Numérique	Type = Numérique	Type = Numérique	Type = Numérique	Type = Numérique	Type = Numérique	Type = Numérique
	Longueur = 14	Longueur = 14	Longueur = 14	Longueur = 14	Longueur = 14	Longueur = 14	Longueur = 14	Longueur = 12
1	0,00	0,23	0,00	0,00	0,00	0,74	0,03	98,7
2	0,00	0,10	0,00	0,00	0,12	0,74	0,04	97,8
3	0,00	0,00	0,00	0,10	0,12	0,74	0,04	96,6
4	0,00	0,49	0,00	0,00	0,12	0,37	0,02	92,0
5	0,00	0,00	0,00	0,62	0,12	0,18	0,08	86,6
6	0,00	0,62	0,00	0,00	0,00	0,37	0,01	91,2
7	0,17	0,27	0,10	0,38	0,00	0,00	0,08	81,9
8	0,17	0,19	0,10	0,38	0,02	0,06	0,08	83,1
9	0,17	0,21	0,10	0,38	0,00	0,06	0,08	82,4
10	0,17	0,15	0,10	0,38	0,02	0,10	0,08	83,2
11	0,21	0,36	0,12	0,25	0,00	0,00	0,06	81,4
12	0,00	0,00	0,00	0,55	0,00	0,37	0,08	88,1
13	0,00	0,20	0,00	0,40	0,00	0,50	0,06	
14	0,17	0,40	0,10	0,40	0,10	0,10	0,08	

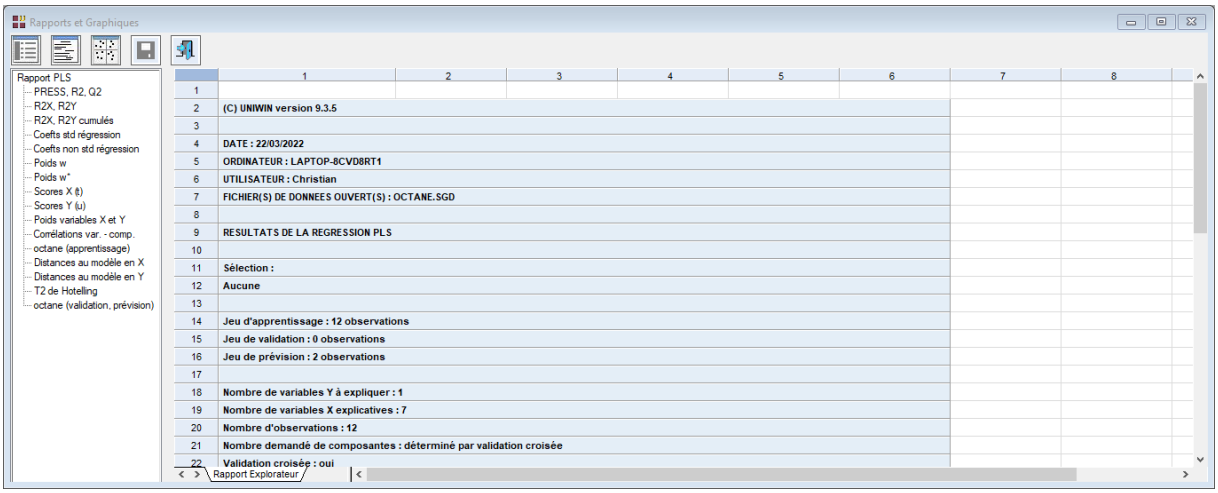
Renseignons la boîte de dialogue comme montré ci-dessous.

Nous ne précisons pas le nombre de composantes à extraire et laissons la validation croisée le déterminer.

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

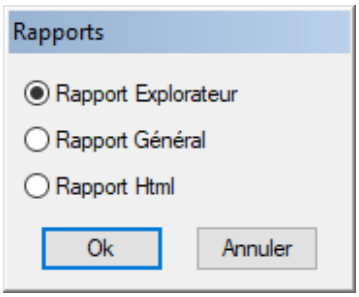


Après quelques instants, la fenêtre Rapports et Graphiques s'affiche :

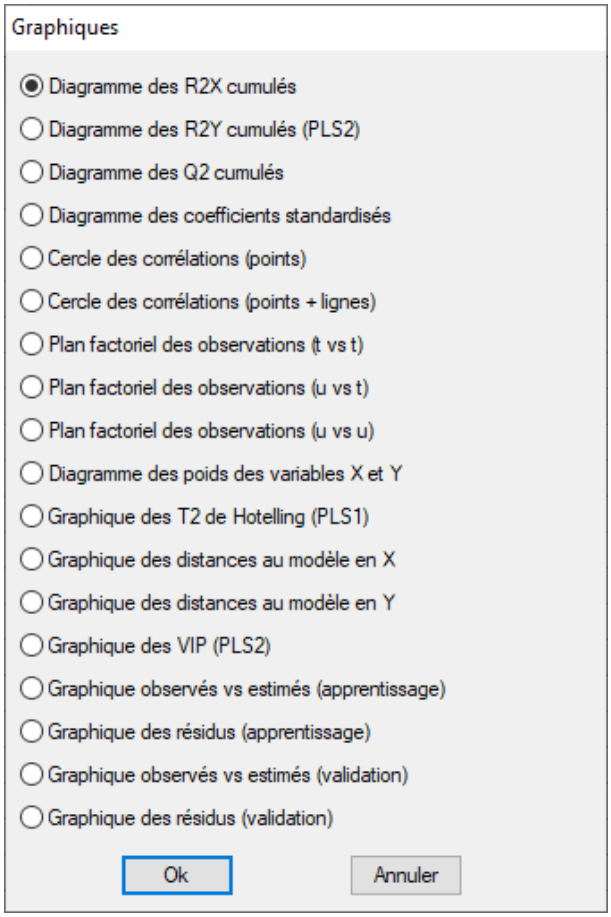


La barre d'outils 'Rapports et Graphiques' permet par l'icône 'Données'  de rappeler la boîte de dialogue d'entrée des données.

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

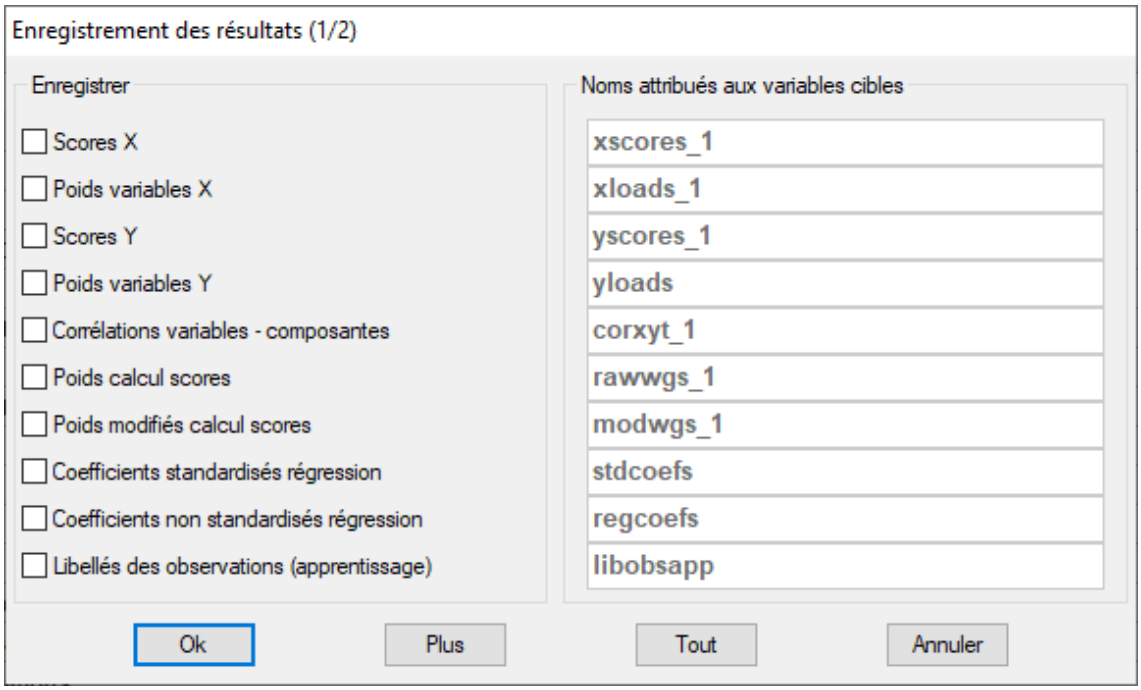


et l'icône 'Graphiques'  affiche la boîte de dialogue des options pour les graphiques :



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

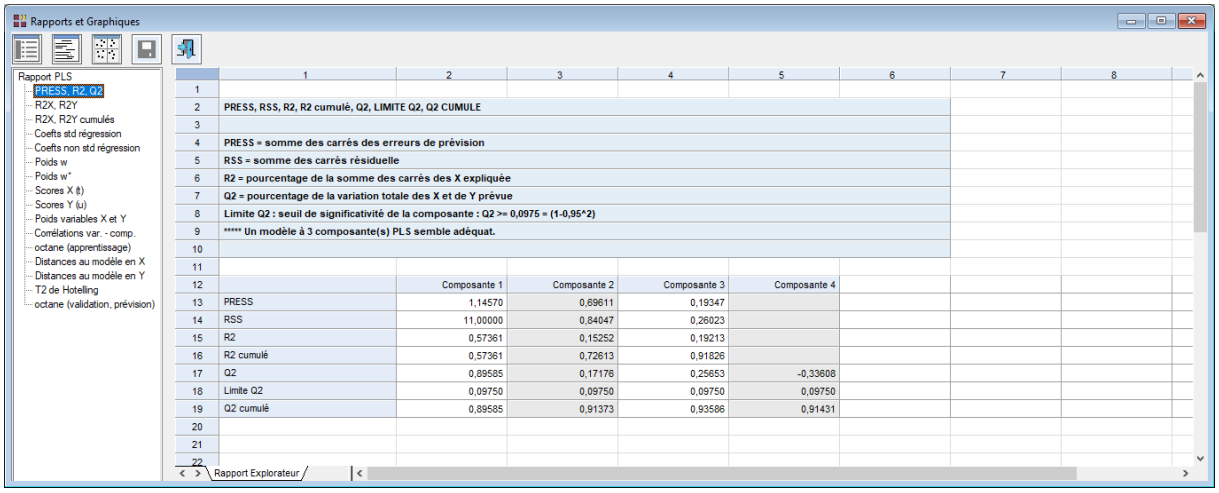
L'icône 'Quitter'  permet de quitter l'analyse.



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

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 des exemples du rapport pour notre analyse.



Ce premier tableau est important. Il n’est affiché que si la validation croisée a été mise en œuvre. Il donne les informations suivantes pour chaque composante :

- PRESS (PRediction Error Sum of Squares) : somme des carrés des erreurs de prévision
- RSS (Residual Sum of Squares) : somme des carrés résiduelle
- R2 : pourcentage de la somme des carrés des X expliquée

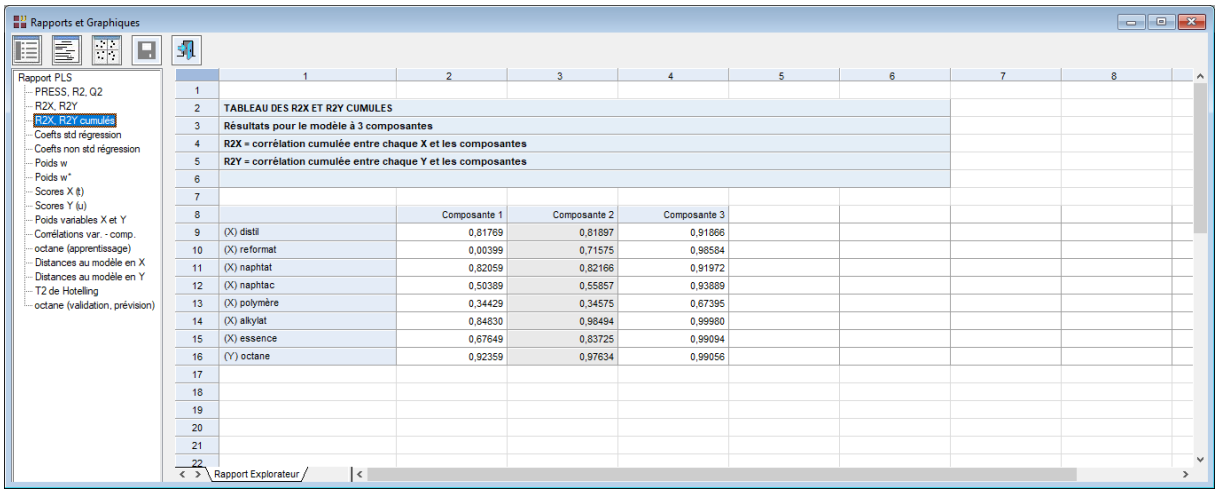
- Q2 : pourcentage de la variation totale des X et de Y prévue
- Limite Q2 : seuil de significativité de la composante égal à $(1-0,95^2) = 0,0975$

L'en-tête de ce tableau indique le nombre optimal de composantes déterminé par la validation croisée, ici 3.

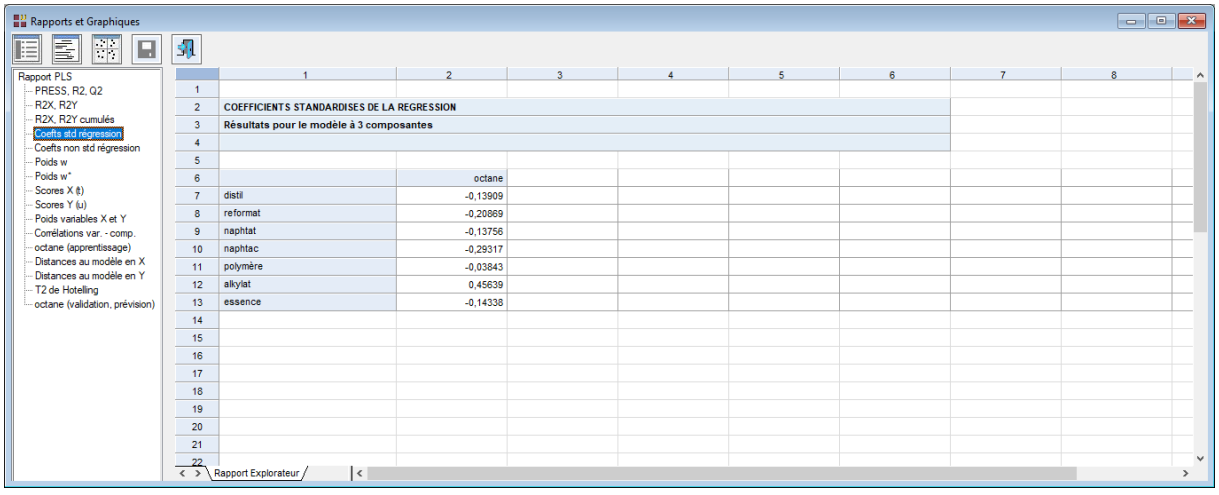
Note : si le nombre de composantes à extraire a été précisé dans la boîte de dialogue initiale et que la validation croisée a été demandée, alors il est conseillé de mettre en œuvre la méthode en précisant le nombre de composantes déterminé par la validation croisée.

Le deuxième tableau affiche les corrélations entre chaque X et le Y avec les composantes extraites. Il décrit le pouvoir explicatif de chaque composante.

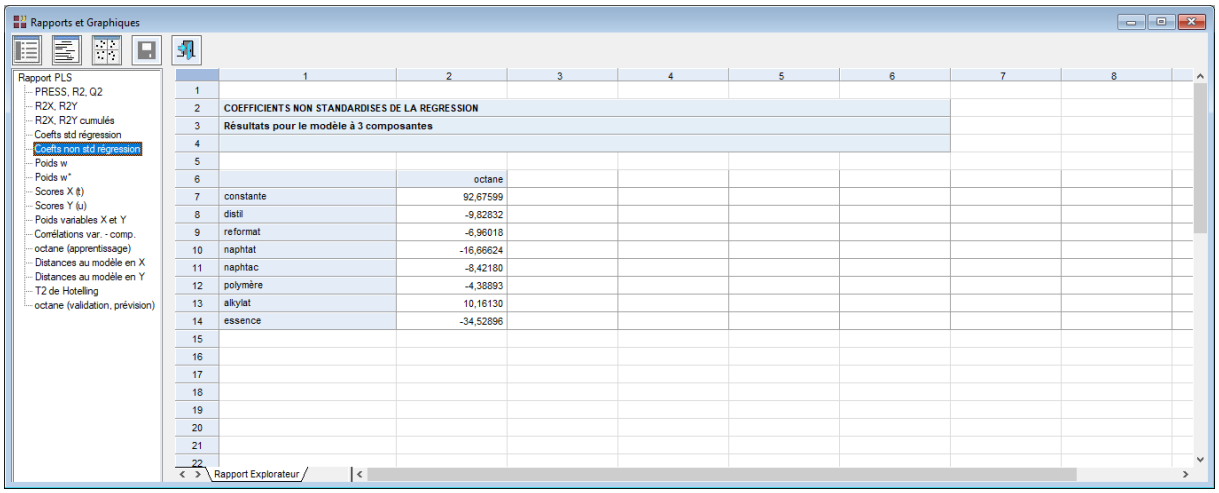
Le troisième tableau affiche ces informations cumulées.



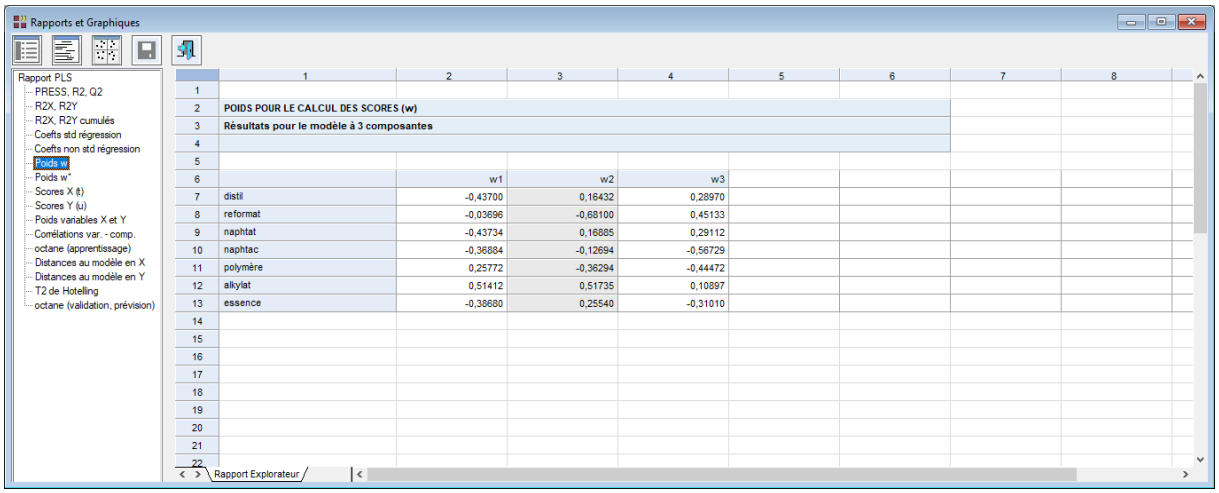
Le quatrième tableau affiche les coefficients standardisés de la régression pour le modèle à 3 composantes.



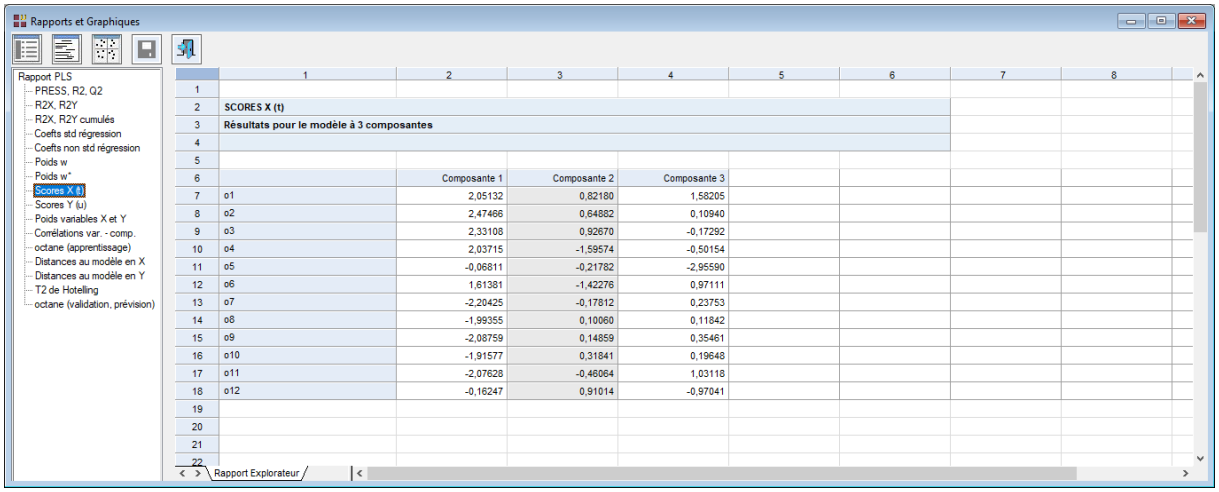
Le cinquième tableau affiche ces informations non standardisées c'est-à-dire dans les unités des variables X et de la variable Y.



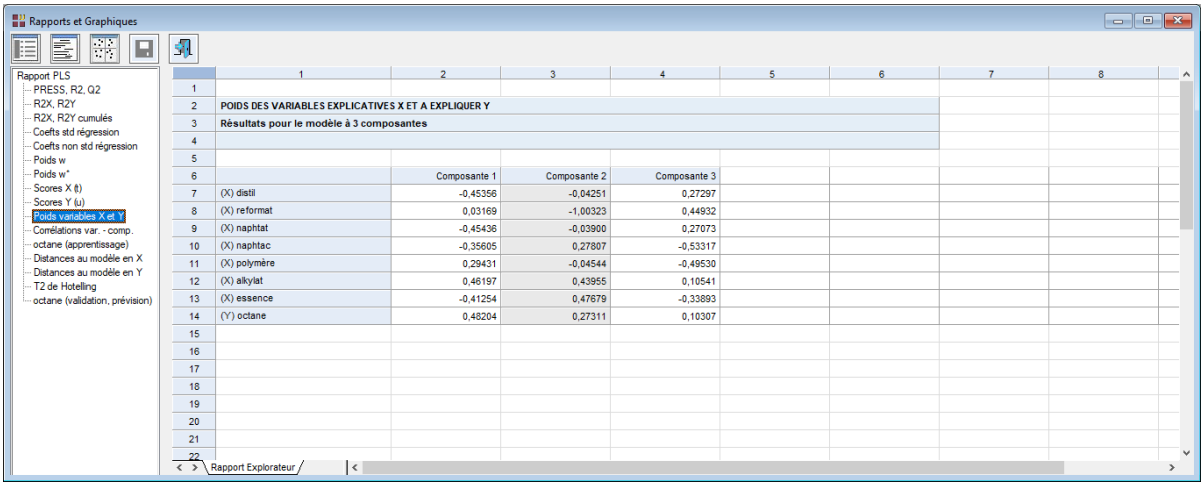
Les deux tableaux suivants affichent les poids w et w* (w modifiés pour tenir compte des nombres de degrés de liberté) qui sont utilisés pour le calcul des scores des observations.



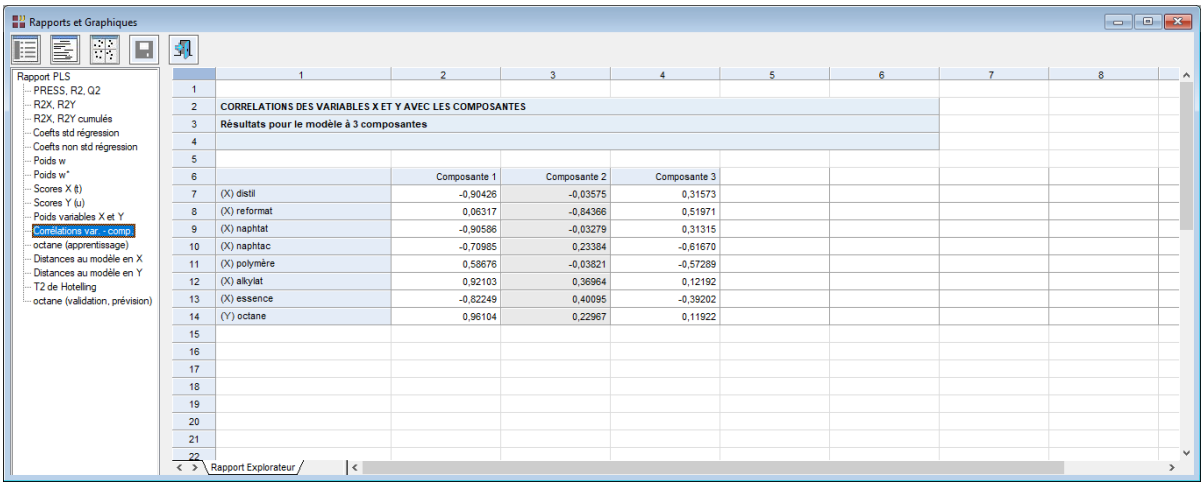
Le tableau des scores X affiche les coordonnées des observations sur les composantes t. Le tableau des scores Y affiche ces coordonnées sur les composantes u. En pratique seule la composante u_1 est interprétable et donc utilisée.



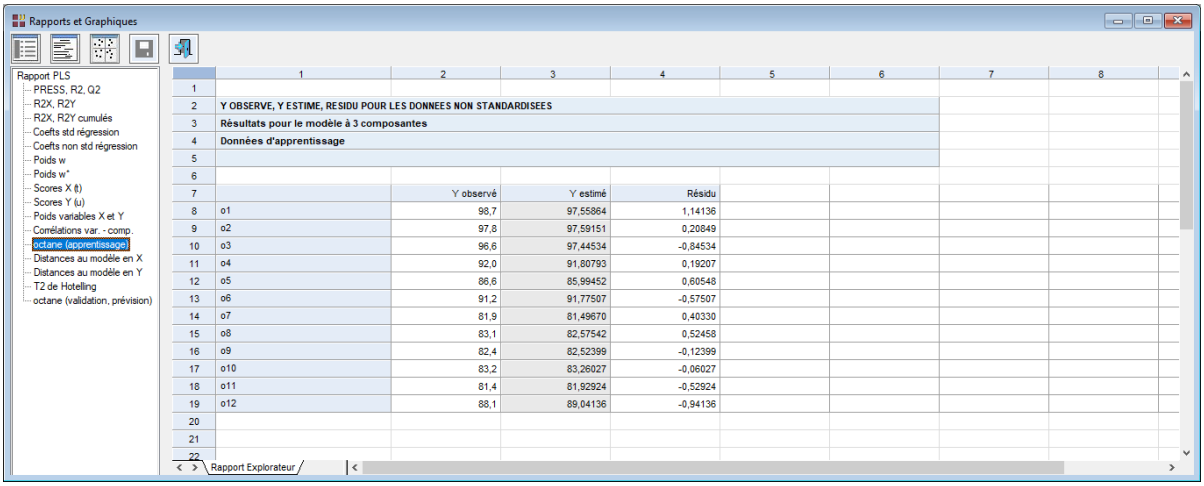
Le tableau suivant affiche les poids (loadings) des variables X et Y.



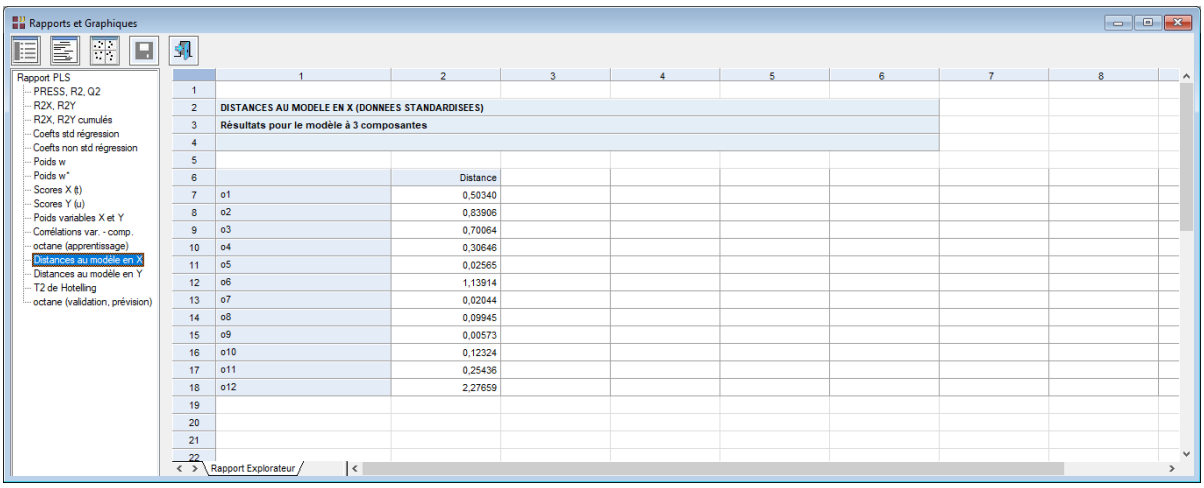
Le tableau suivant affiche les corrélations entre les variables X et Y et les composantes principales.



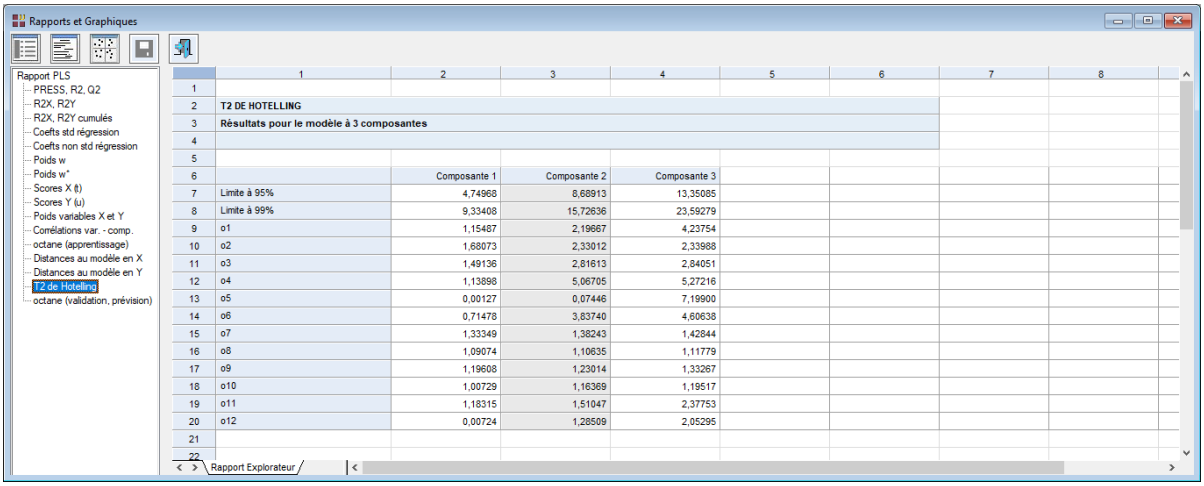
Les valeurs observées, estimées par le modèle à 3 composantes ainsi que les résidus sont ensuite affichés.



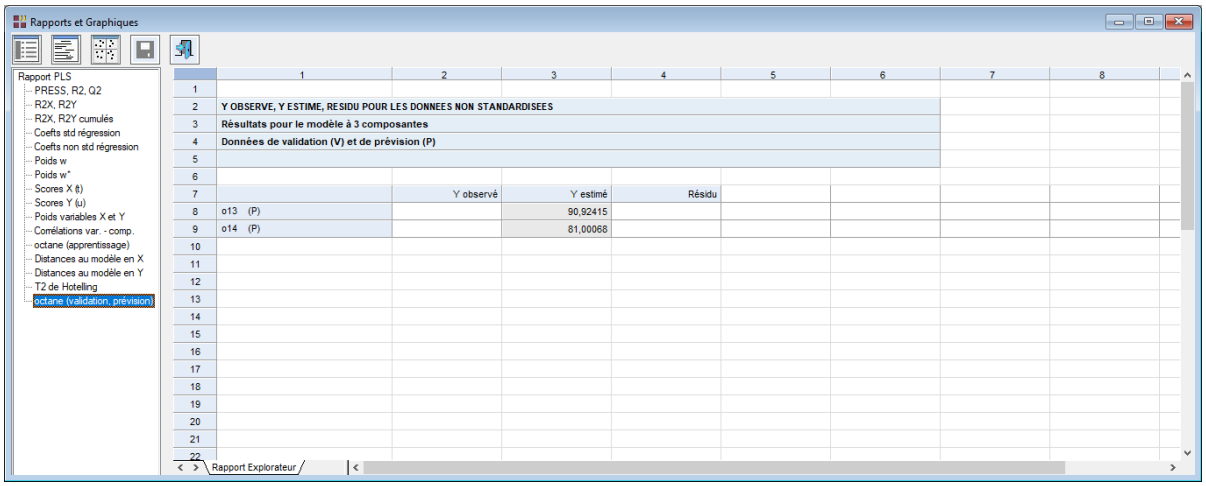
Les deux tableaux suivants affichent les distances des observations au modèle en X et au modèle en Y. Plus la distance est grande, moins l'observation caractérisée par ses valeurs en X et en Y est bien reconstituée par le modèle.



Le tableau suivant affiche les T2 de Hotelling. Il permet de détecter d'éventuelles observations atypiques.



Le dernier tableau affiche les valeurs observées, estimées et les résidus pour les données de validation et de prévision.

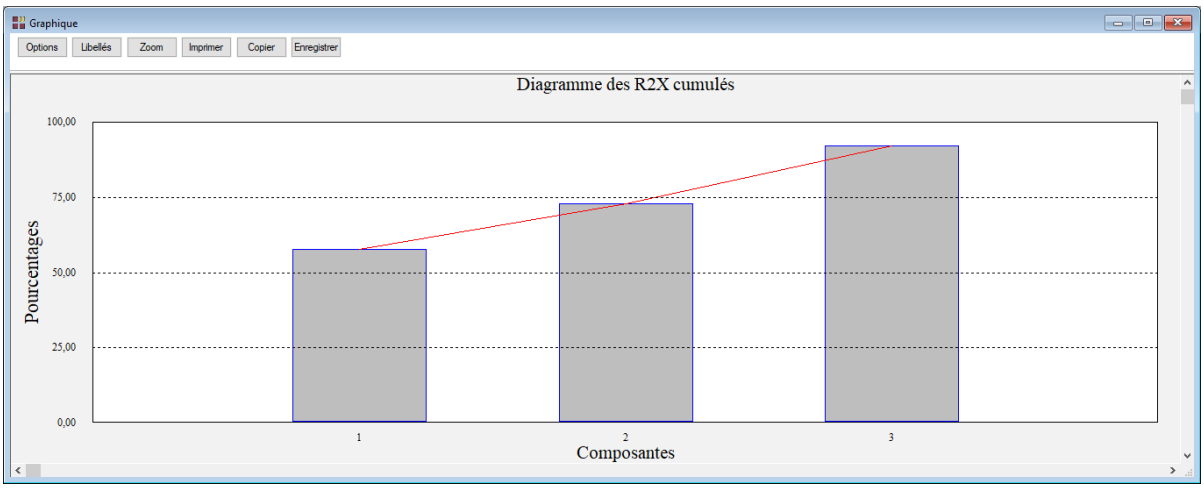


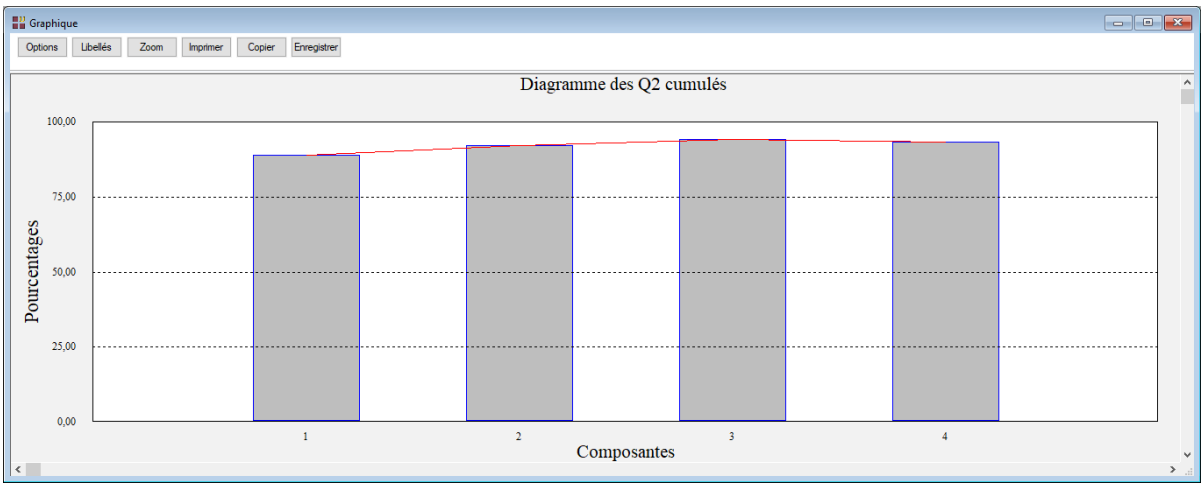
L'option Graphiques

- Diagrammes des R2X cumulés, des R2Y cumulés et des Q2 cumulés

Le diagramme des R2Y cumulés n'est proposé que pour si une régression PLS2 a été mise en œuvre.

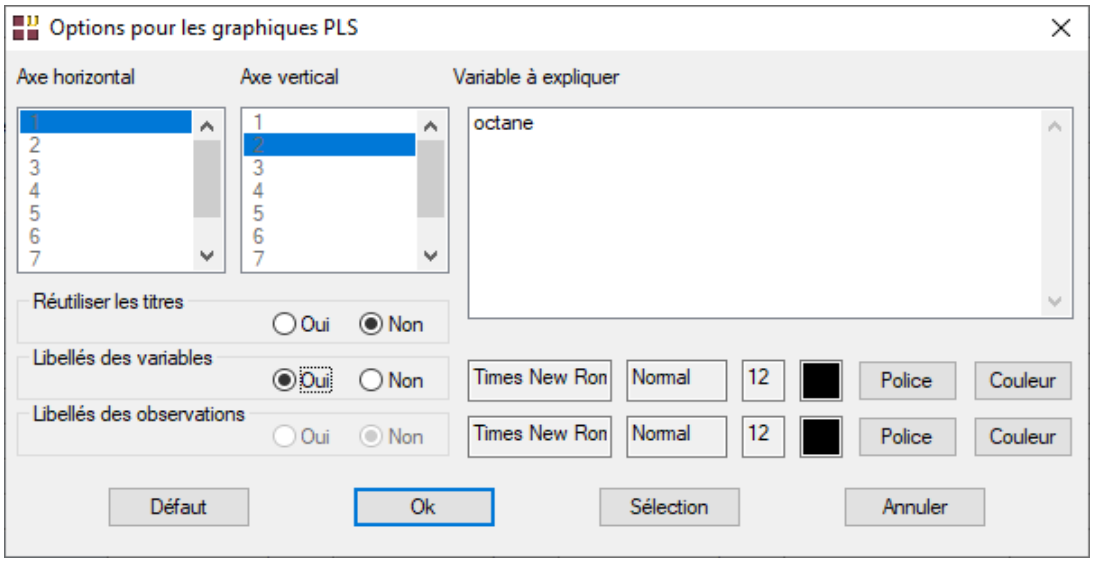
Le diagramme des Q2 cumulés n'est proposé que si la validation croisée a été mise en œuvre.

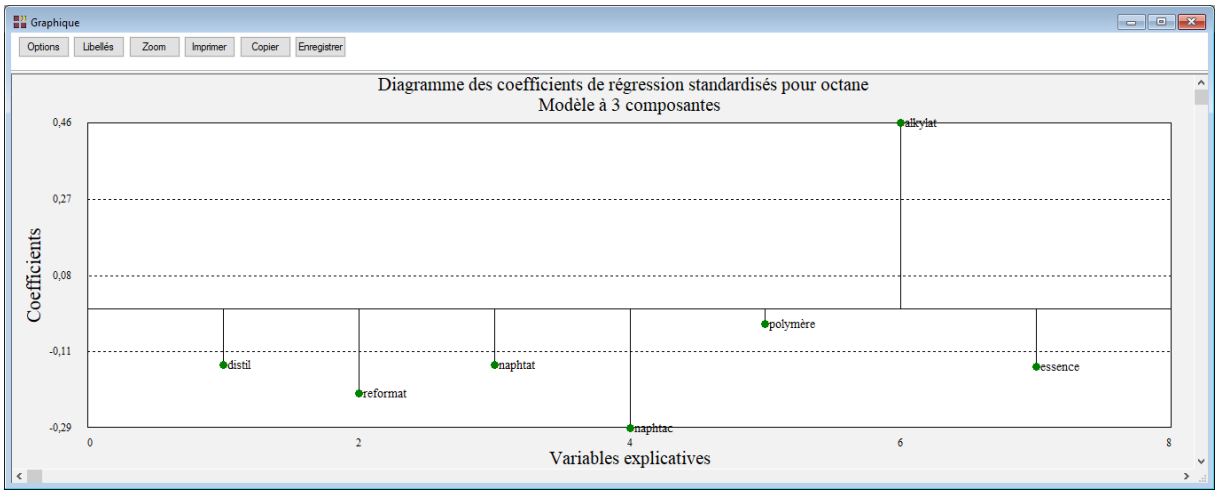




- Diagramme des coefficients standardisés

Une boîte de dialogue permet de choisir la variable Y (ici une unique variable octane) et de préciser si les libellés des variables X doivent être affichés.



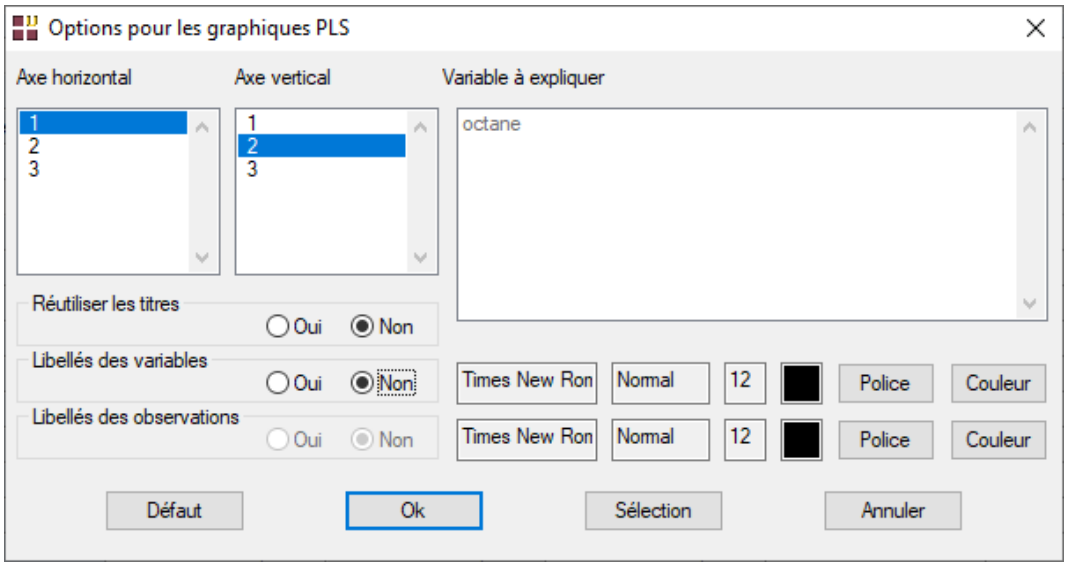


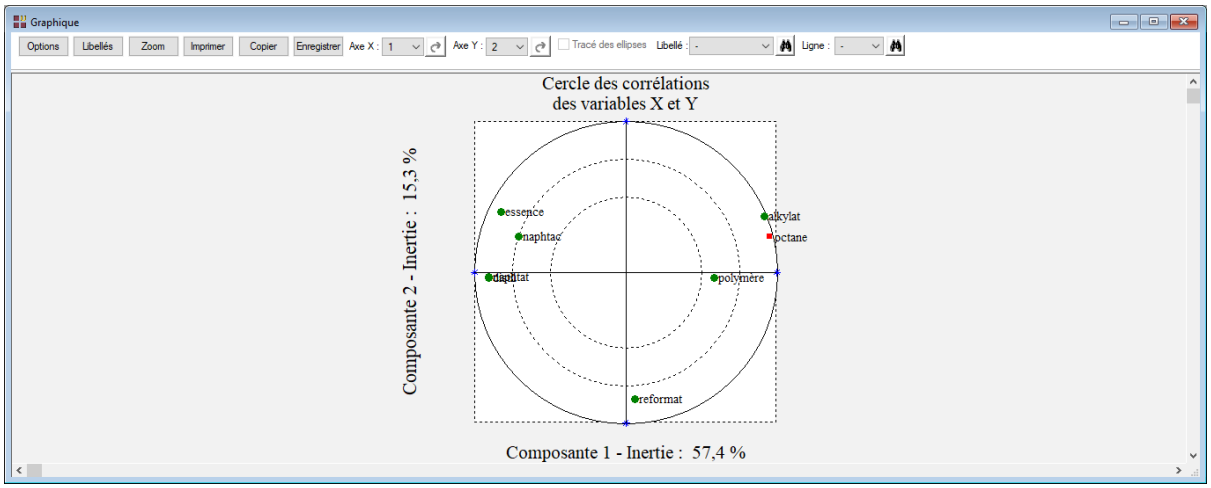
- Cercle des corrélations (points ou points + lignes)

Une boîte de dialogue permet d'indiquer les composantes PLS (axes) à représenter et de préciser si les libellés des variables X et Y doivent être affichés.

La barre d'outils affichée au-dessus du graphique permet de modifier le choix des axes et de localiser des variables par libellés ou lignes (numéros).

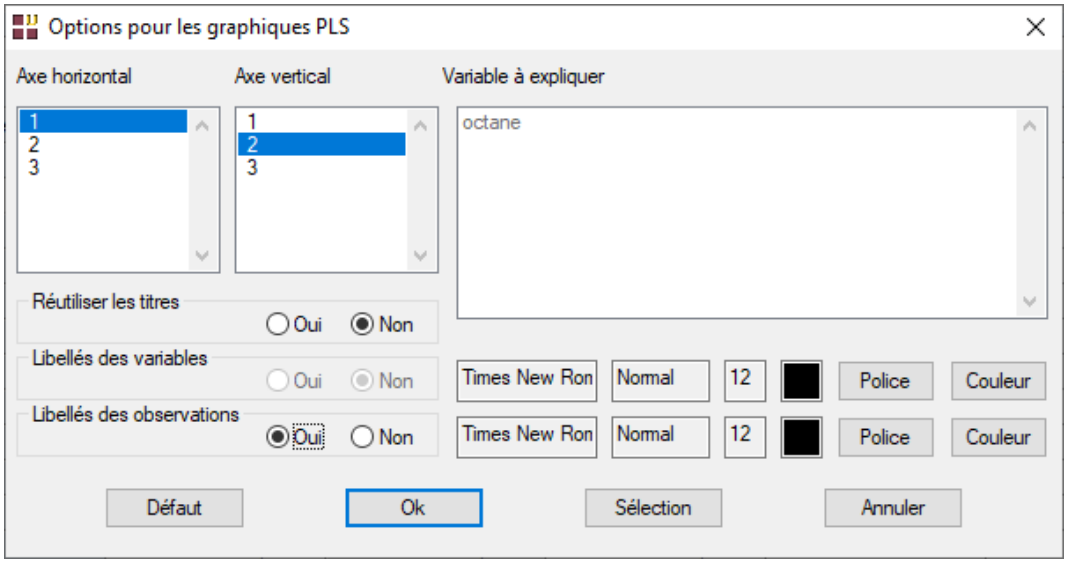
Cela est utile si le graphique a été affiché sans les libellés des variables dans le cas où ces variables sont nombreuses.

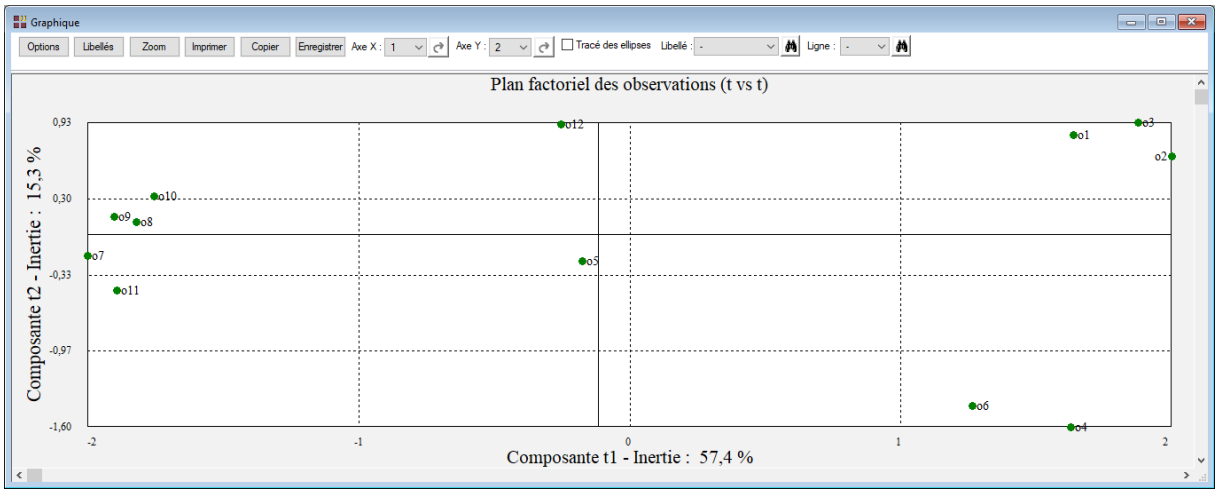




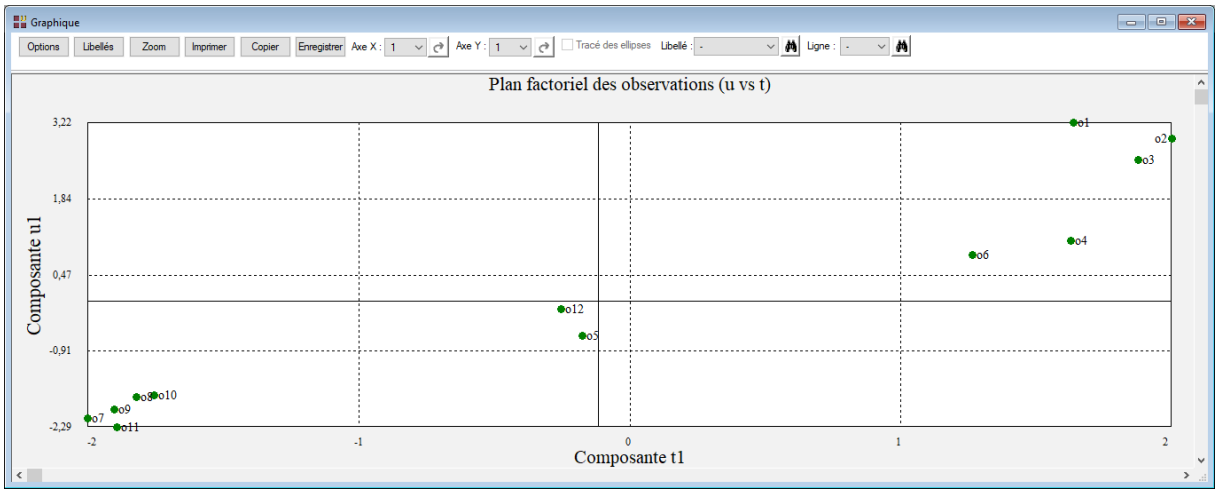
- Plan factoriel des observations (t vs t) - espace des X

Une boîte de dialogue permet d'indiquer les composantes PLS (axes) à représenter et de préciser si les libellés des variables X et Y doivent être affichés.



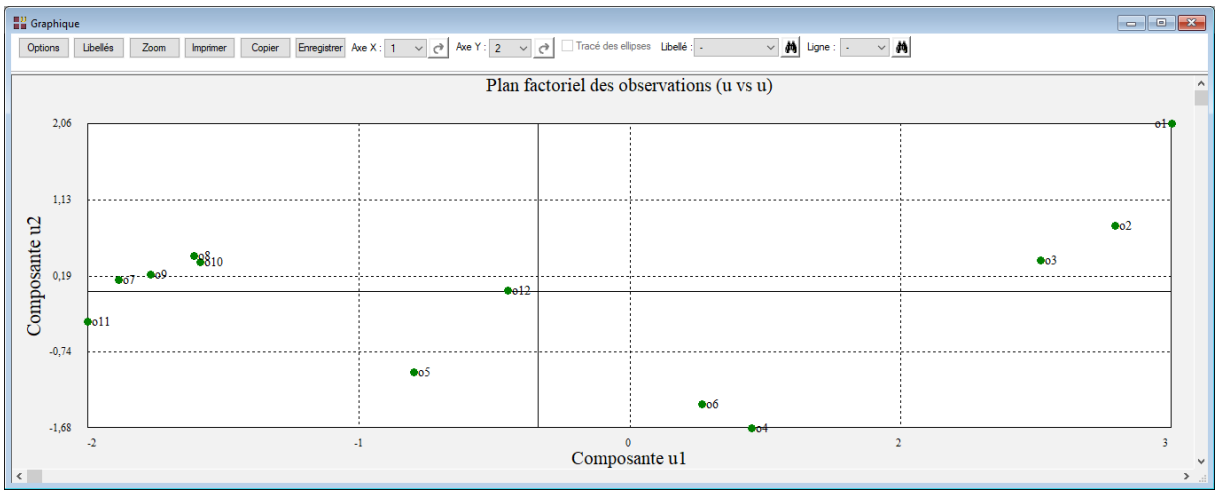


- Plan factoriel des observations (u vs t) - liaison entre l'espace des X et celui des Y

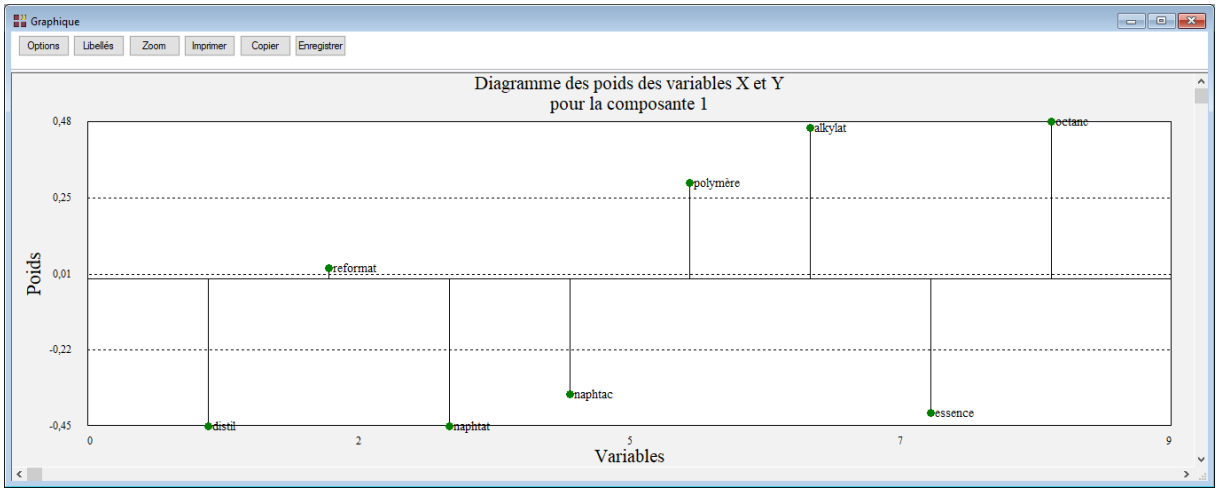


Ce graphique montre la forte corrélation R2Y entre Y et la première composante.

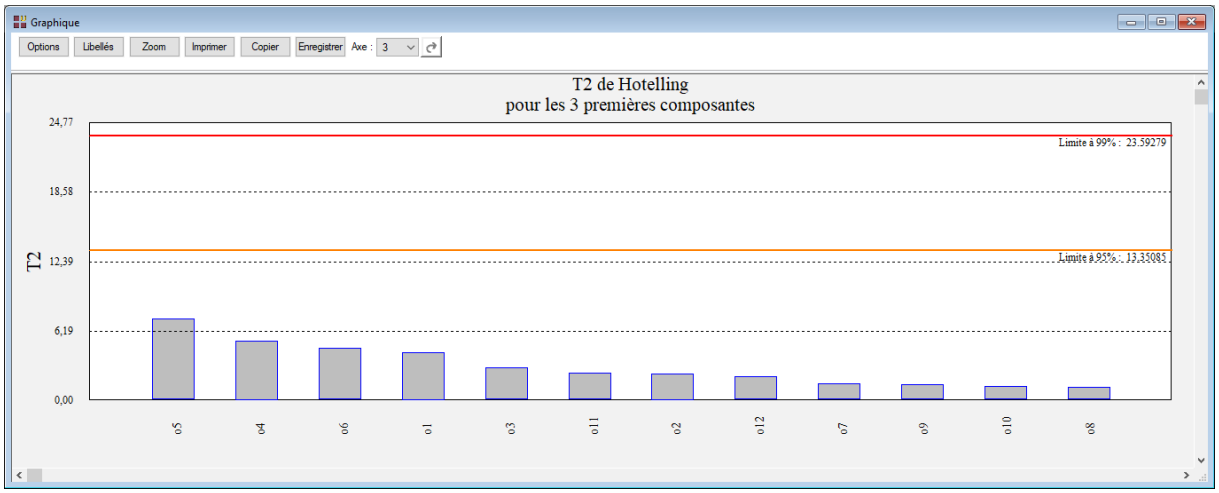
- Plan factoriel des observations (u vs u) - espace des Y



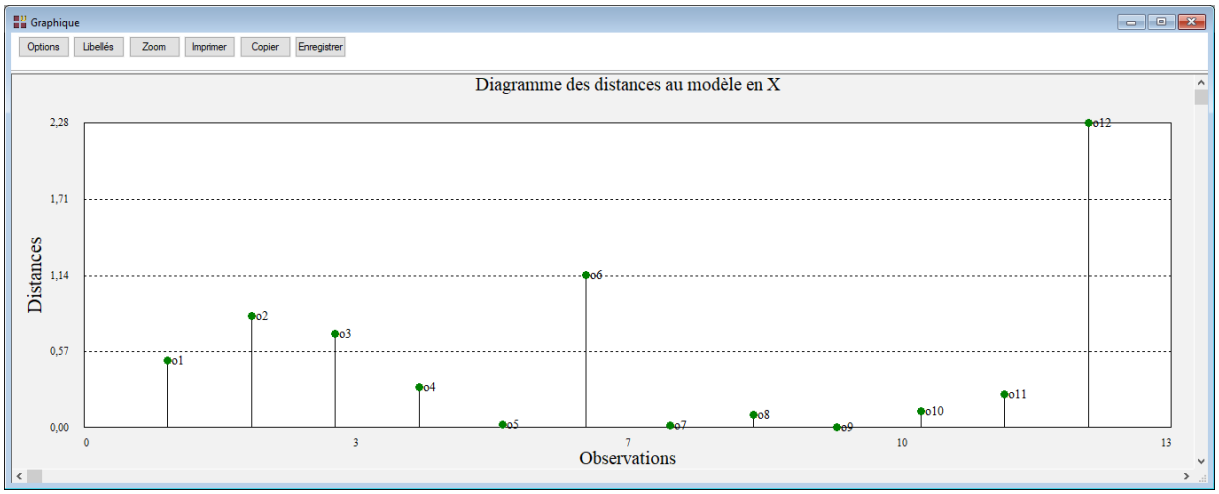
- Diagramme des poids des variables X et Y

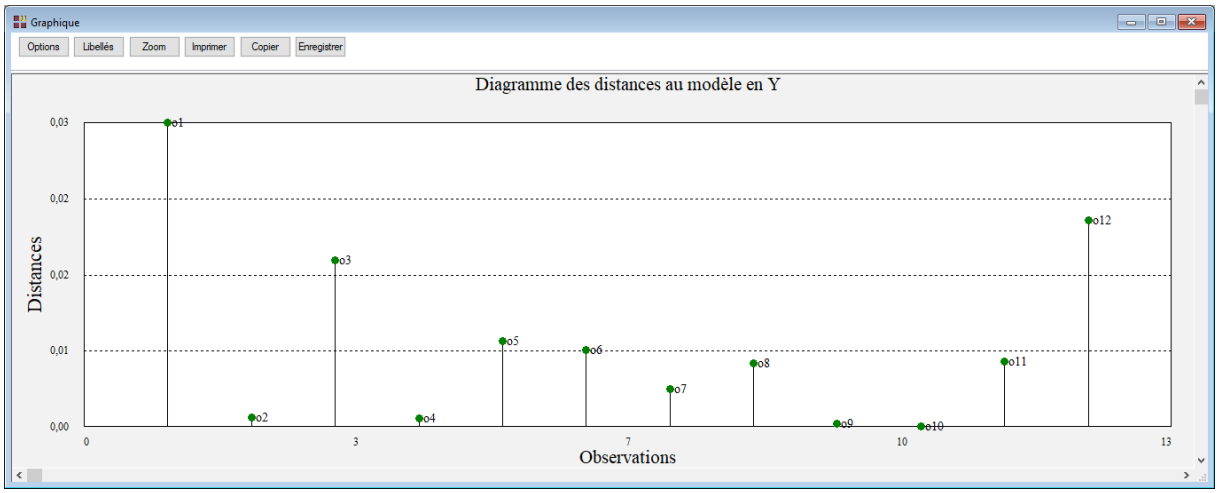


- Graphique des T2 de Hotelling

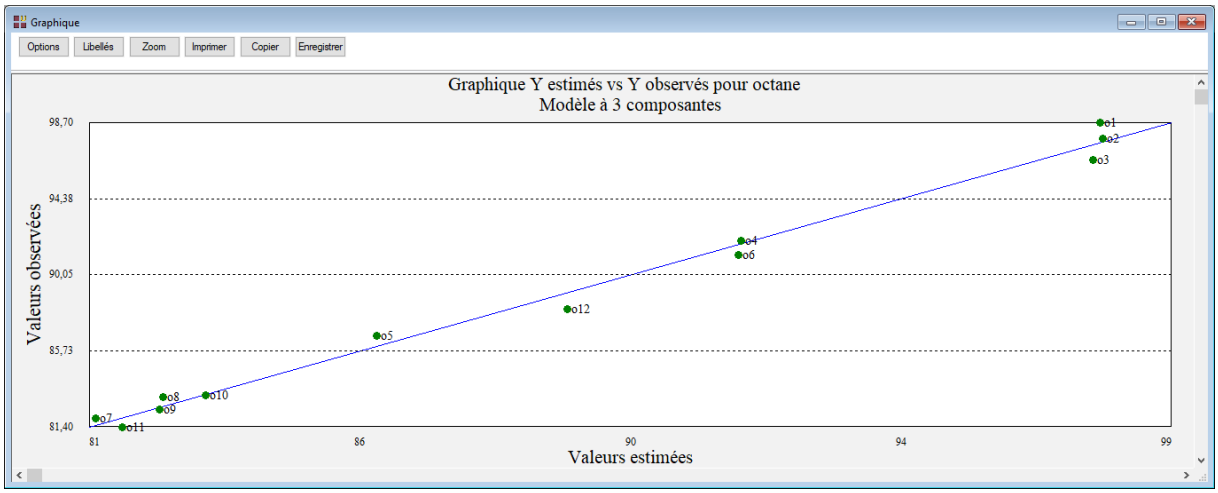


- Graphiques des distances au modèle en X et Y

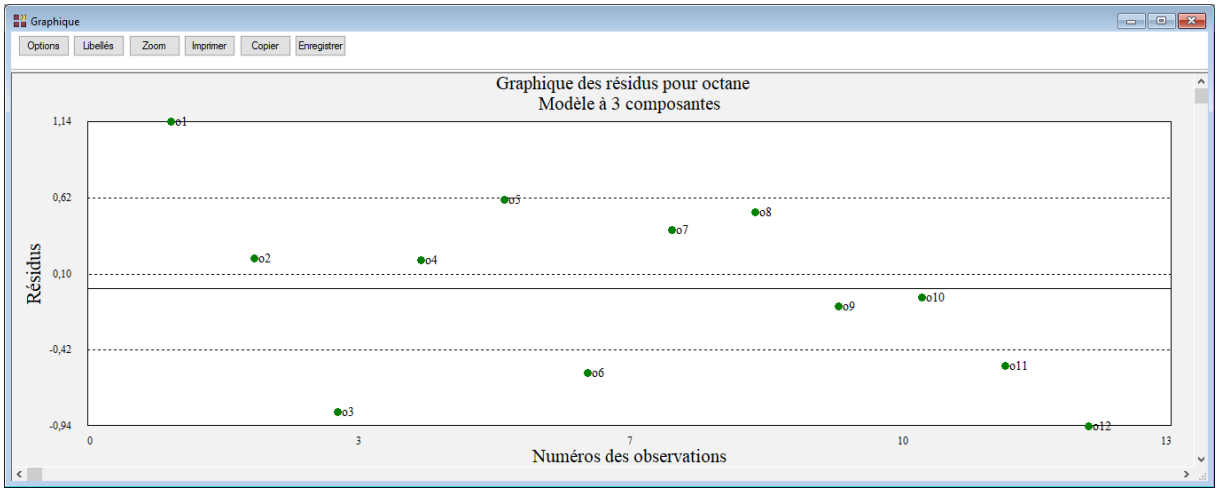




- Graphique observés vs estimés (apprentissage)



- Graphique des résidus (apprentissage)



- Graphique observés vs estimés (validation)

Ce graphique n'est disponible que s'il y a des données de validation, ce qui n'est pas le cas dans cet exemple.

- Graphique des résidus (validation)

Ce graphique n'est disponible que s'il y a des données de validation, ce qui n'est pas le cas dans cet exemple.

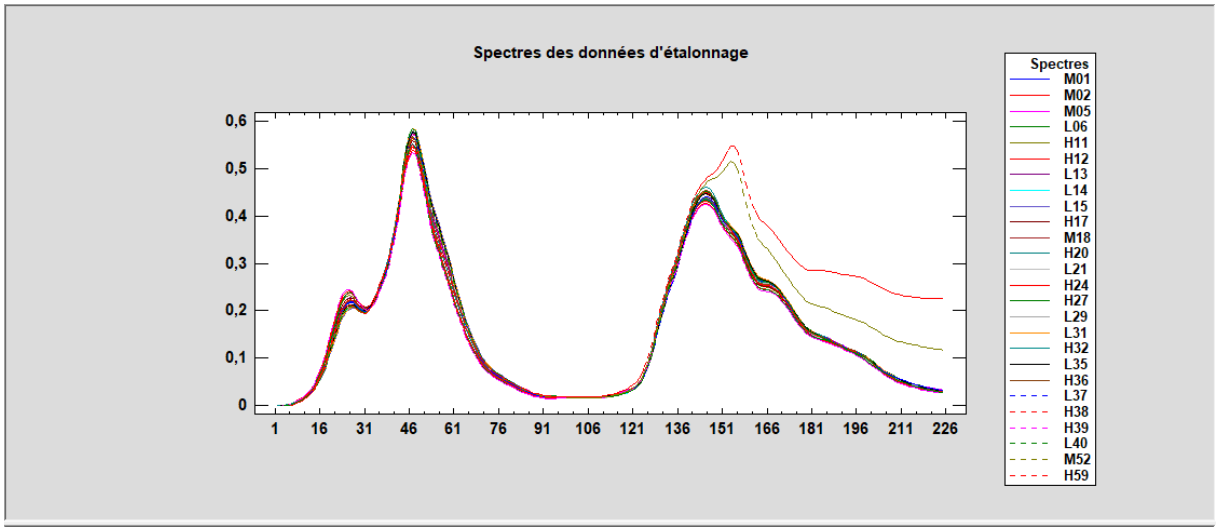
Exemple 2 : Fichier Octane2 (PLS1)

Ce deuxième exemple illustre l'étalonnage multidimensionnel sur un exemple de données de spectroscopie.

Il montre l'intérêt de la régression PLS qui permet de prendre en compte un grand nombre de variables explicatives sur un petit nombre d'échantillons.

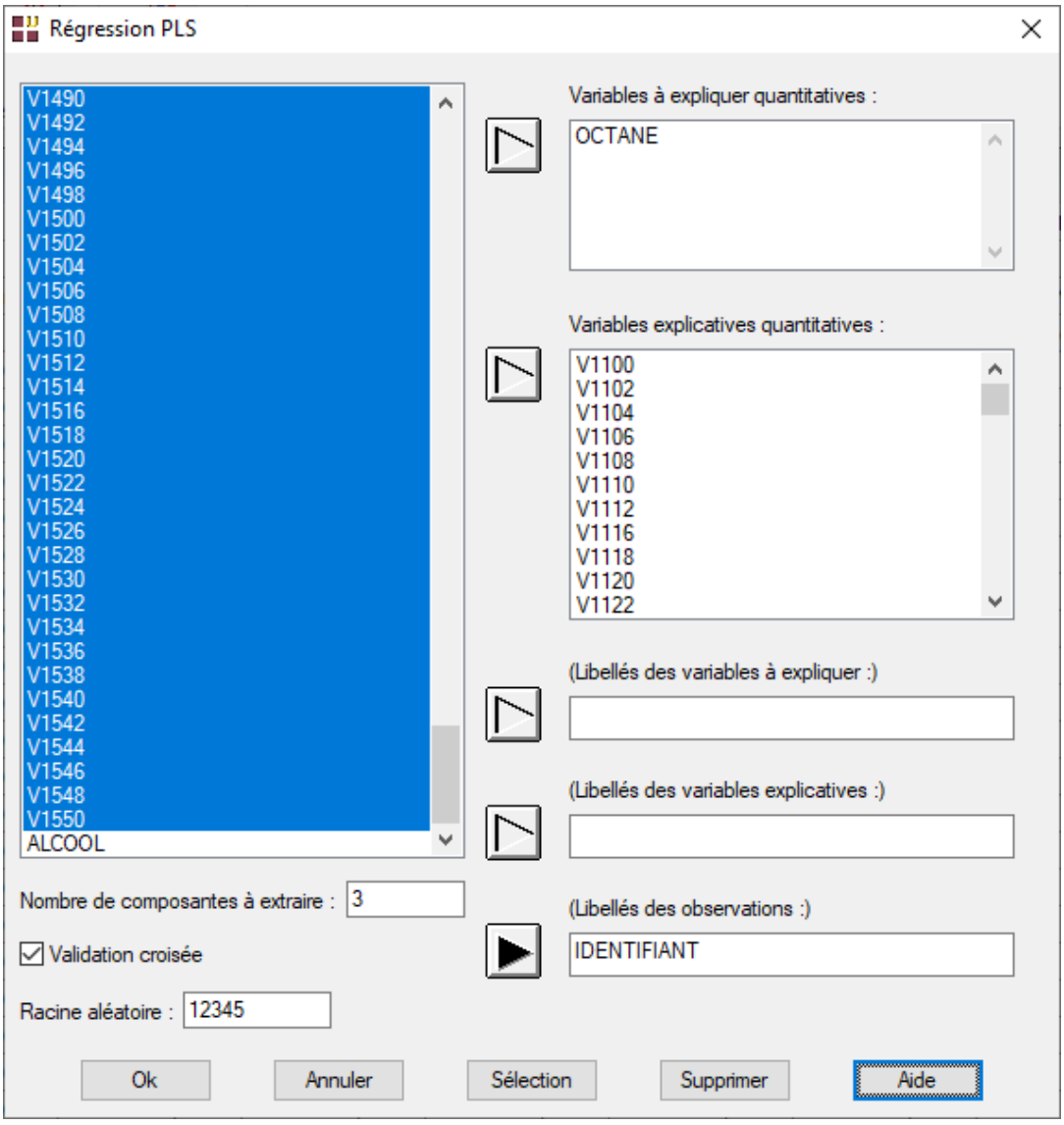
La variable Y représente l'indice d'octane et les 225 variables des valeurs d'absorbance à différentes longueurs d'onde.

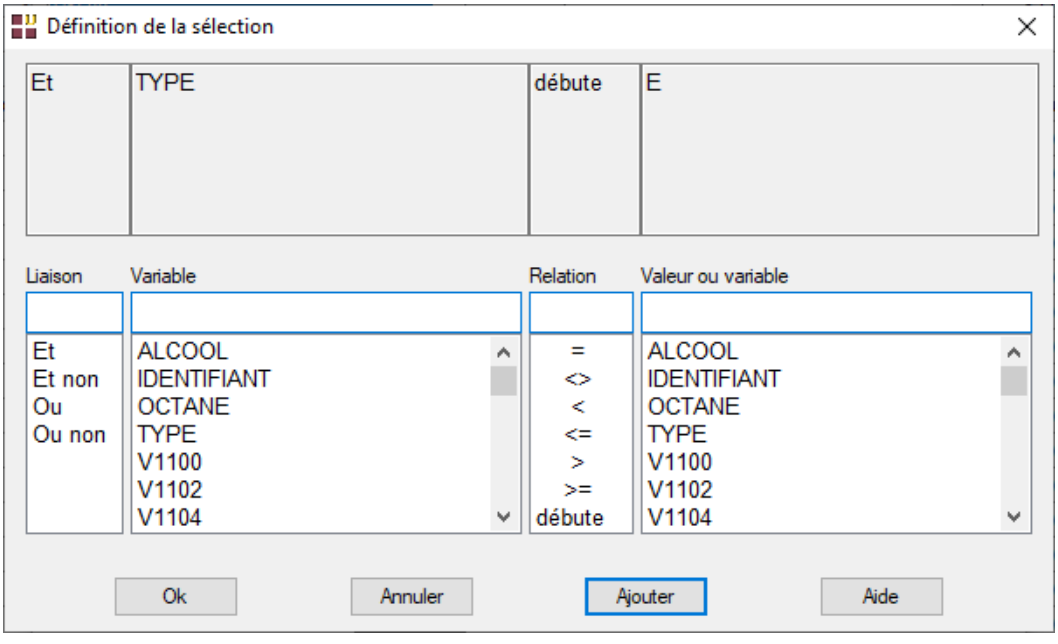
Le modèle est construit à partir d'un ensemble de 39 échantillons (26 d'étalonnage et 13 de validation) d'essence recueillis sur une période suffisamment longue pour être représentatif de la dispersion de l'ensemble de la production.



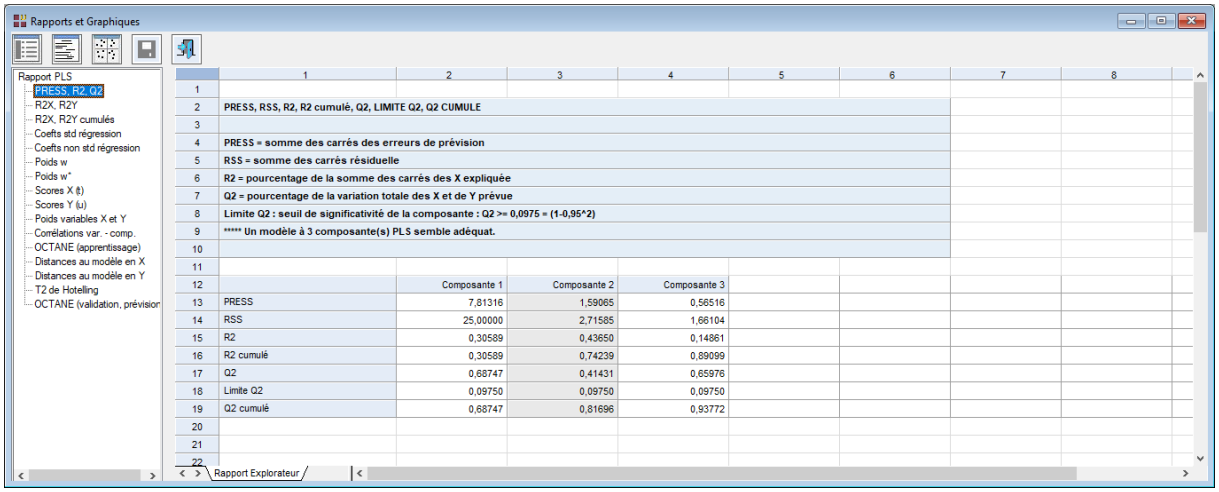
Renseignons la boîte de dialogue de la régression PLS comme montré ci-après en sélectionnant OCTANE comme variable à expliquer, les colonnes V1100 à V1550 comme variables explicatives et la colonne IDENTIFIANT comme libellés des observations.

Demandons 3 composantes et la validation croisée et sélectionnons les échantillons d'étalonnage (TYPE débute par E) en cliquant sur le bouton 'Sélection' :


Régression PLS
V1490
V1492
V1494
V1496
V1498
V1500
V1502
V1504
V1506
V1508
V1510
V1512
V1514
V1516
V1518
V1520
V1522
V1524
V1526
V1528
V1530
V1532
V1534
V1536
V1538
V1540
V1542
V1544
V1546
V1548
V1550
ALCOOL
Variables à expliquer quantitatives :
OCTANE
Variables explicatives quantitatives :
V1100
V1102
V1104
V1106
V1108
V1110
V1112
V1116
V1118
V1120
V1122
(Libellés des variables à expliquer :)
(Libellés des variables explicatives :)
(Libellés des observations :)
IDENTIFIANT
Nombre de composantes à extraire : 3
Validation croisée
Racine aléatoire : 12345
Ok
Annuler
Sélection
Supprimer
Aide

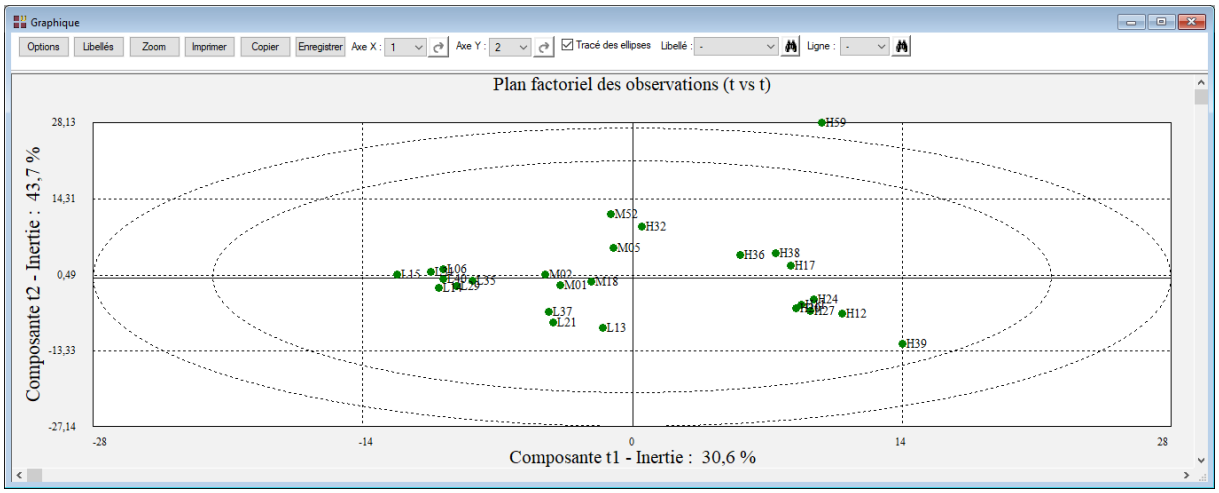

Définition de la sélection
Et
TYPE
débute
E
Liaison
Variable
Relation
Valeur ou variable
Et
Et non
Ou
Ou non
ALCOOL
IDENTIFIANT
OCTANE
TYPE
V1100
V1102
V1104
=
<>
<
<=
>
>=
débute
ALCOOL
IDENTIFIANT
OCTANE
TYPE
V1100
V1102
V1104
Ok
Annuler
Ajouter
Aide

Après quelques instants, le rapport s'affiche.



	PRESS, RSS, R2, R2 cumulé, Q2, LIMITE Q2, Q2 CUMULE			
	PRESS = somme des carrés des erreurs de prévision			
	RSS = somme des carrés résiduelle			
	R2 = pourcentage de la somme des carrés des X expliquée			
	Q2 = pourcentage de la variation totale des X et de Y prévue			
	Limite Q2 : seuil de significativité de la composante : Q2 >= 0,0975 = (1-0,95^2)			
	***** Un modèle à 3 composante(s) PLS semble adéquat.			
		Composante 1	Composante 2	Composante 3
	PRESS	7,81316	1,59065	0,56516
	RSS	25,00000	2,71585	1,66104
	R2	0,30589	0,43650	0,14861
	R2 cumulé	0,30589	0,74239	0,89099
	Q2	0,68747	0,41431	0,65976
	Limite Q2	0,09750	0,09750	0,09750
	Q2 cumulé	0,68747	0,81696	0,93772

Affichons le plan factoriel des observations (t1 vs t2) et demandons le tracé des ellipses de Hotelling à 95% et 99%.



Il y a trois niveaux d'octane.

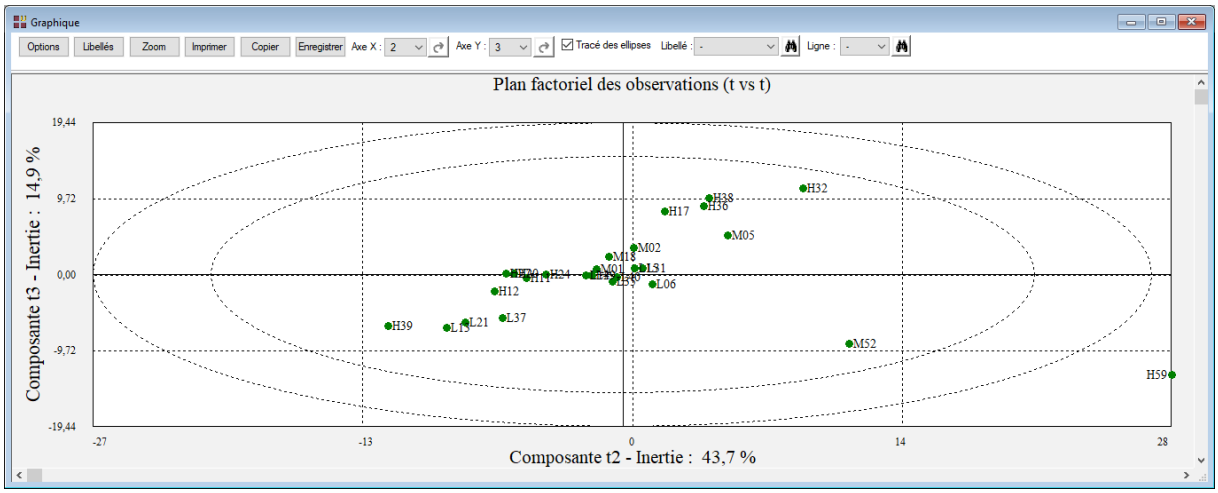
Les identifiants débutant par 'L' indique un niveau bas, ceux débutant par 'M' un niveau moyen et ceux débutant par 'H' un niveau élevé.

Dans ce graphique, les trois niveaux d'octane sont bien séparés et l'échantillon 'H59' avec addition d'alcool y apparaît comme un point atypique.

L'échantillon 'M52' (niveau moyen) également avec addition d'alcool est proche du groupe 'H' (niveau élevé).

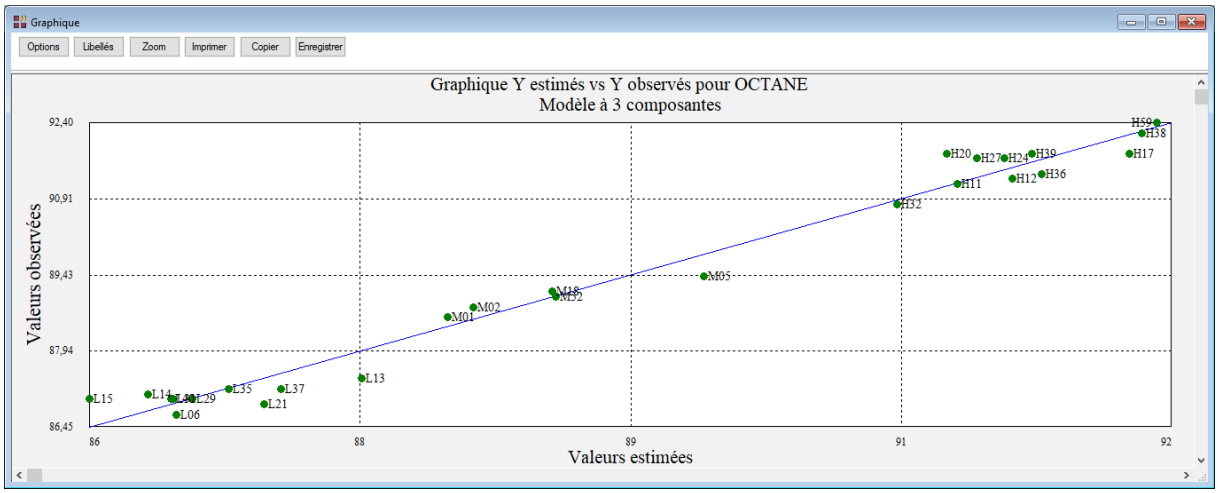
L'addition d'alcool a pour effet d'augmenter l'indice d'octane.

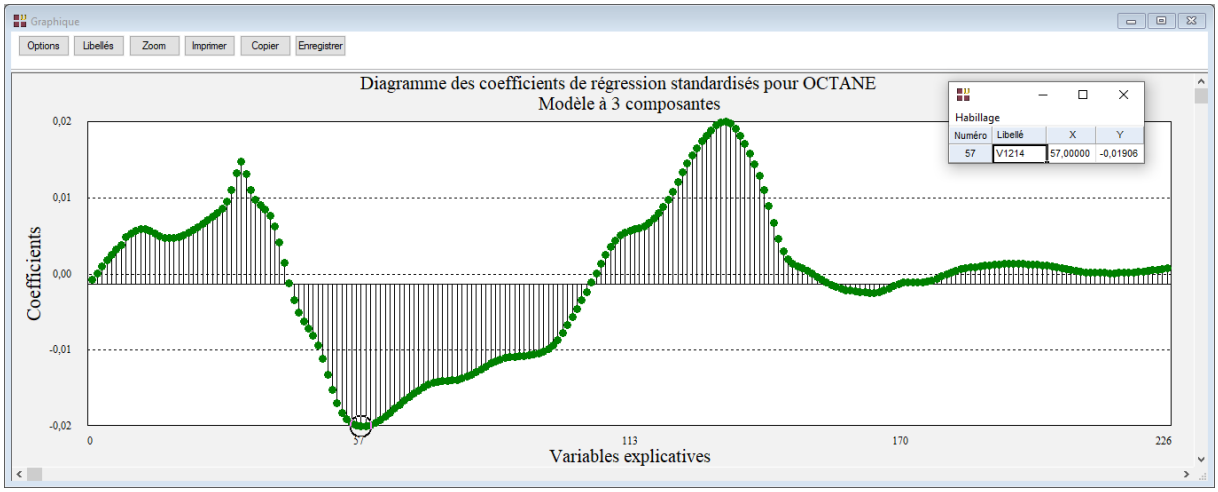
Le plan t2 vs t3 montre bien que ces deux échantillons avec addition d'alcool sont différents des autres échantillons.

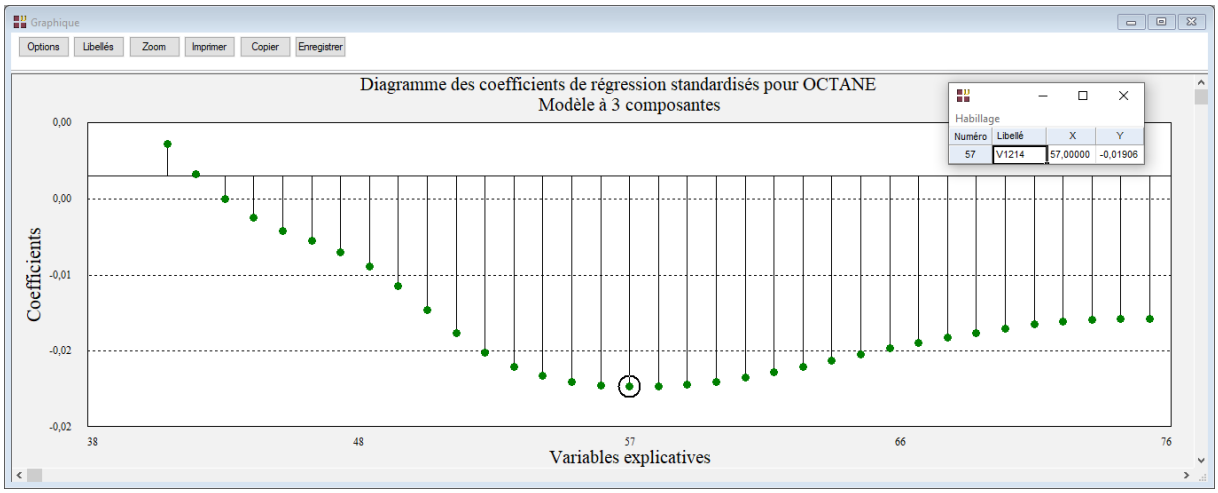


Le graphique observés vs estimés (apprentissage) confirme la qualité du modèle.

Le diagramme des coefficients de régression standardisés permet de visualiser les longueurs d'onde les plus utiles pour prévoir l'indice d'octane. Il est possible en cliquant sur un point dans ce graphique d'afficher la longueur d'onde correspondante. Le bouton Zoom dans la barre d'outils permet de zoomer de diverses façons dans le graphique.

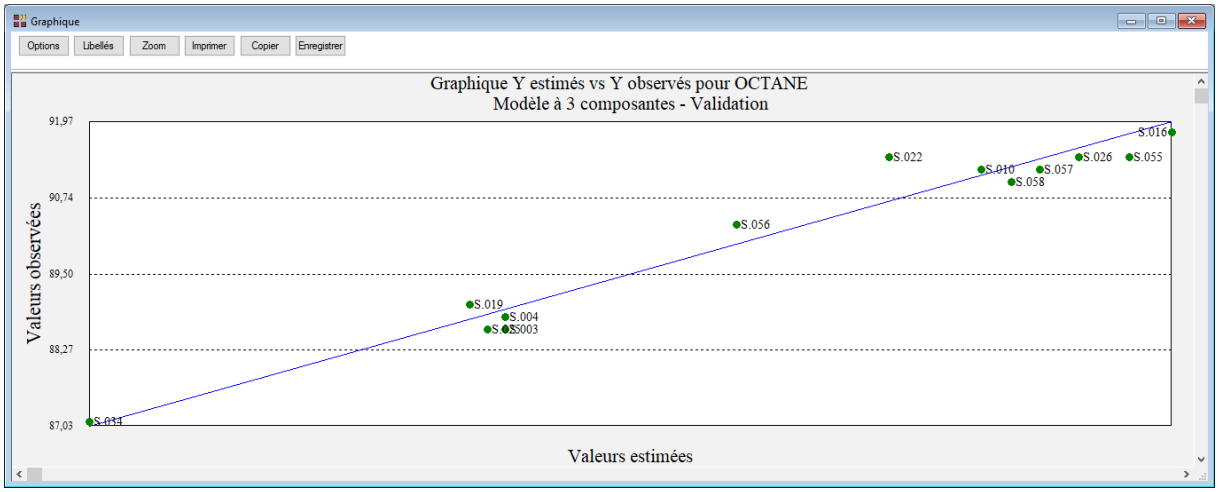






Les deux longueurs d'onde 1214 et 1366 sont donc les deux variables principales pour la prévision de l'indice d'octane.

Le graphique observés vs estimés (validation et prévision) affiche les données des 13 spectres de validation.



Exemple 3 : Fichier Thé (PLS2)

Ce troisième exemple illustre l'usage de la régression PLS en analyse conjointe. Il s'agit de relier les préférences de consommateurs aux caractéristiques des produits évalués.

Des scénarios sont construits représentant des tasses de thé hypothétiques caractérisées par quatre facteurs :

1. Température (1 = chaud, 2 = tiède, 3 = froid)
2. Sucre (1 = sans sucre, 2 = un sucre, 3 = deux sucres)
3. Force (1 = fort, 2 = moyen, 3 = léger)
4. Citron (1 = avec du citron, 2 = sans citron)

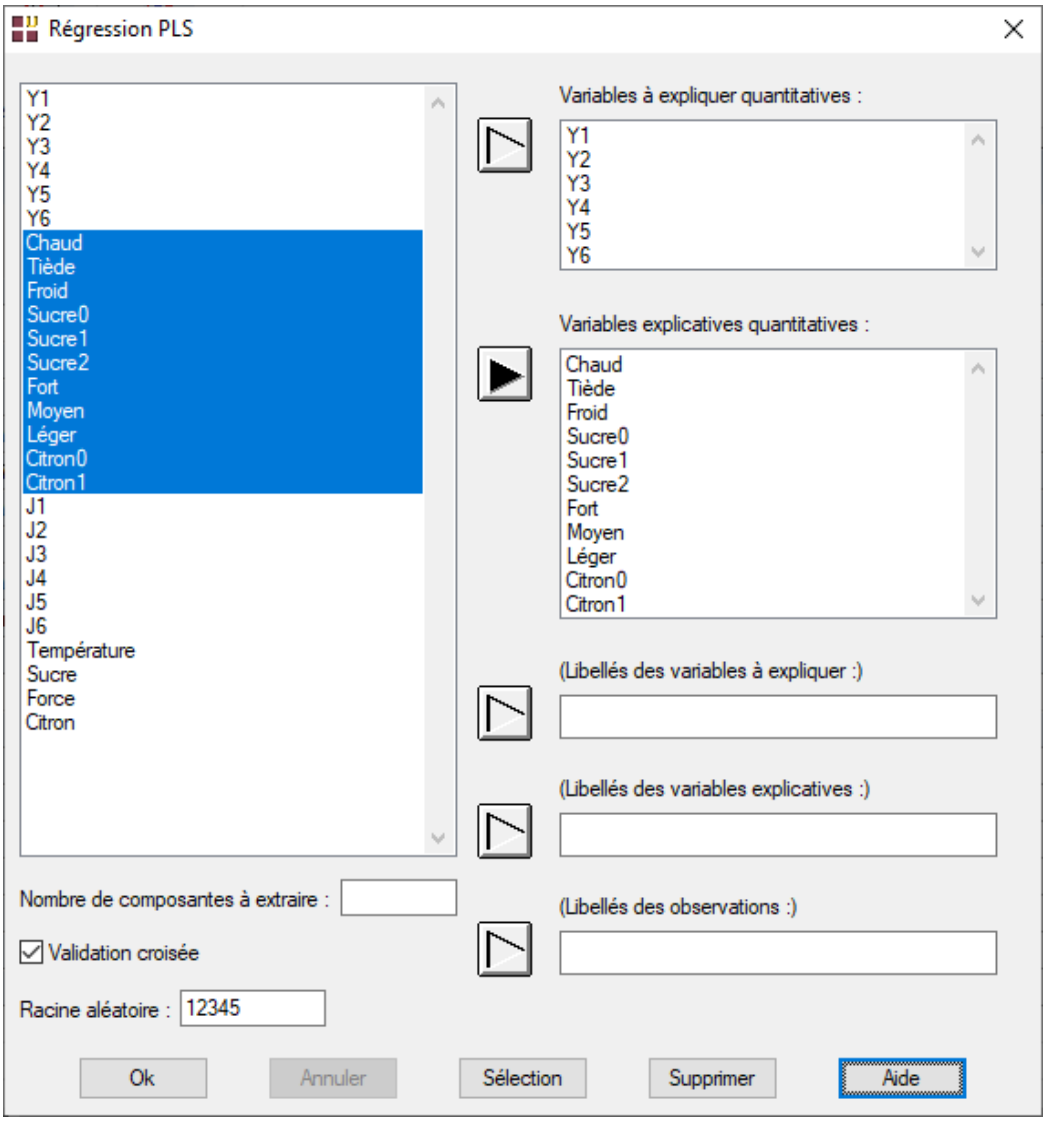
Parmi les 3x3x3x2 = 54 scénarios possibles, on sélectionne 18 combinaisons formant un plan orthogonal (toutes les paires de variables indicatrices de modalités n'appartenant pas aux mêmes facteurs sont non corrélées). On demande ensuite à chacun des 6 juges interrogés nommés J1 à J6 de classer par ordre de préférence les 18 scénarios.

Les variables X sont constituées des variables indicatrices des modalités des facteurs. Elles sont nommées : Chaud, Tiède, Froid, Sucre0, Sucre1, Sucre2, Fort, Moyen, Léger, Citron0 et Citron1.

Les variables Y1 à Y6 sont obtenues en inversant les rangs fournis par les juges de manière à obtenir des corrélations positives entre les classements et les variables indicatrices des caractéristiques préférées.

La régression PLS de Y sur X permet une modélisation des classements à l'aide des caractéristiques des scénarios.

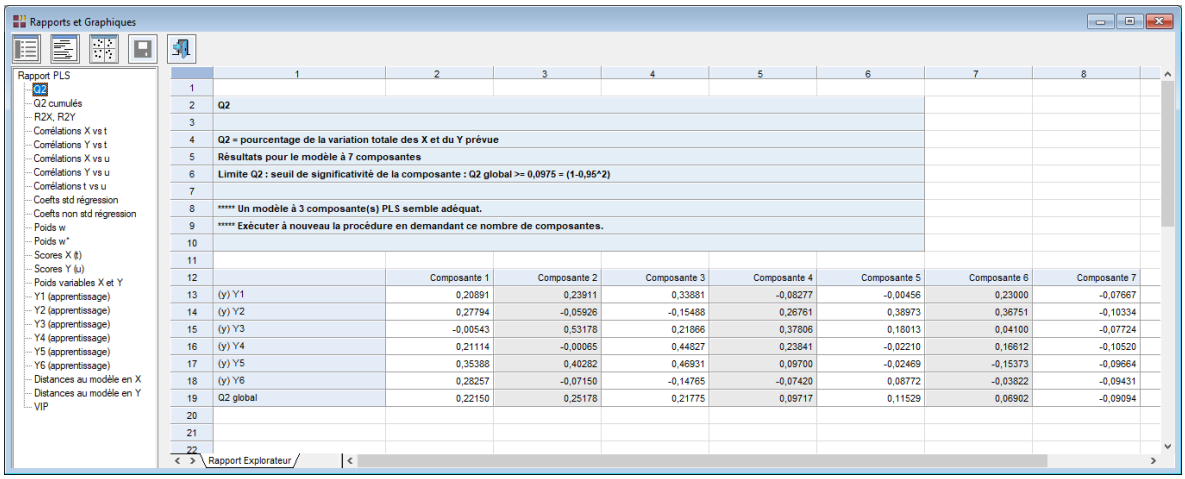
Renseignons la boîte de dialogue de la régression PLS sans préciser le nombre de composantes PLS qui sera ainsi déterminé par la validation croisée.



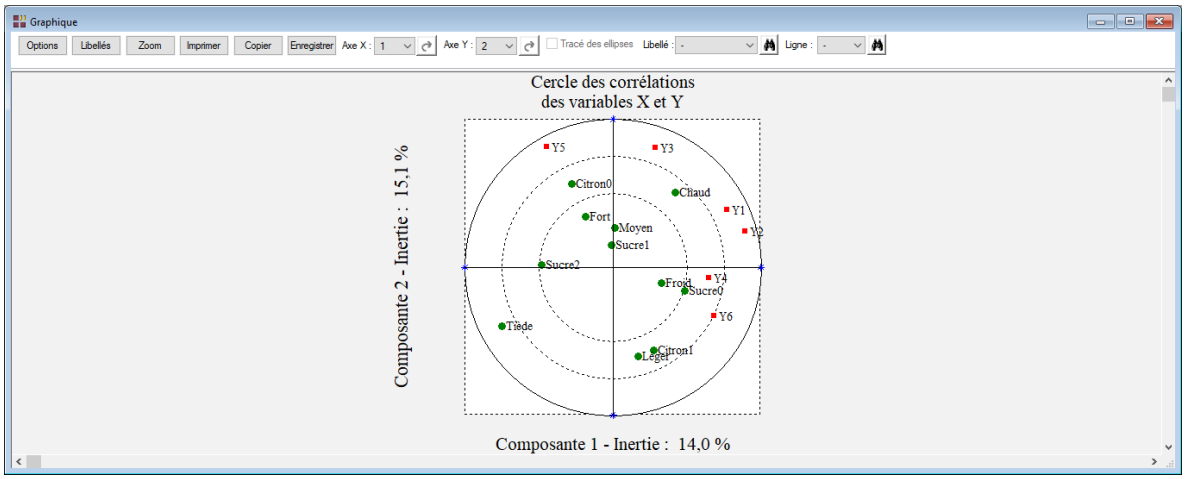
Après quelques instants, le rapport nous indique qu'un modèle à 4 composantes semble adéquat. Exécutons donc à nouveau la régression PLS en précisant que 4 composantes doivent être extraites.

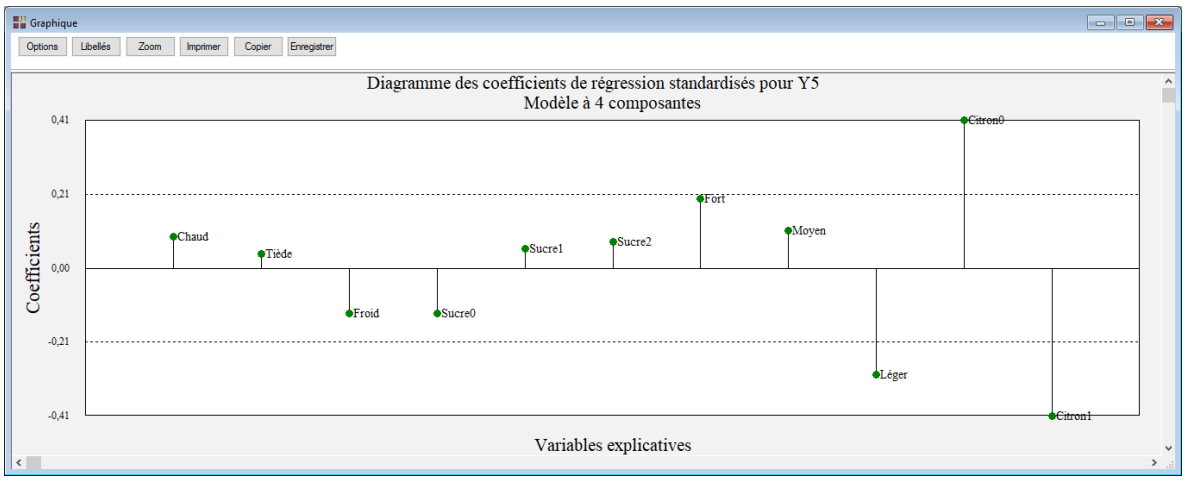
Affichons le cercle des corrélations pour visualiser les poids des variables X et Y et préciser les caractéristiques importantes pour chacun des juges.

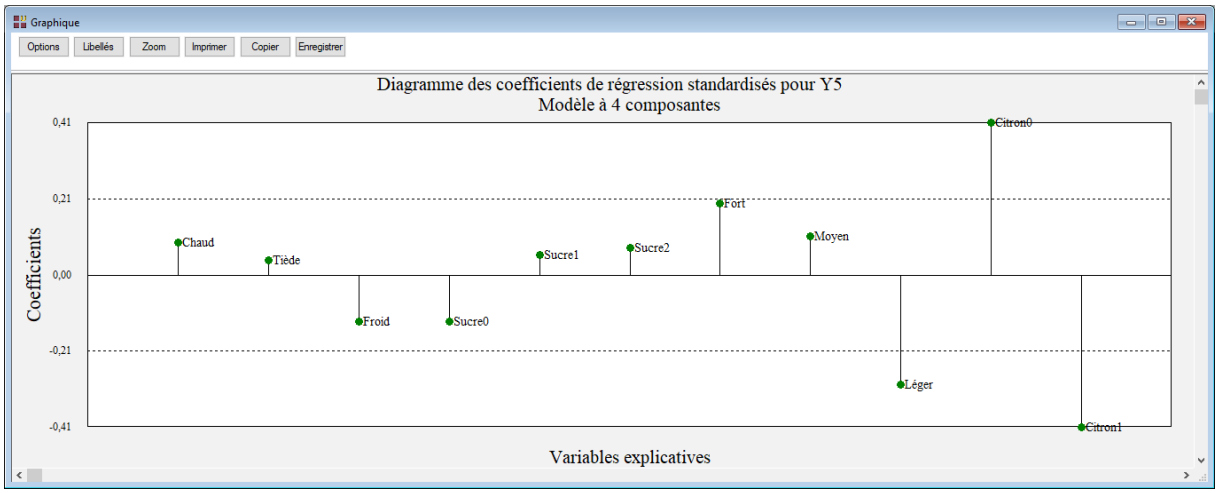
Le tableau des coefficients standardisés de régression et les graphiques associés indiquent également les caractéristiques importantes (en positif ou en négatif) pour chaque juge. Par exemple, Chaud et Tiède pour le juge 1 et Citron0 et Citron1 pour le juge 5.



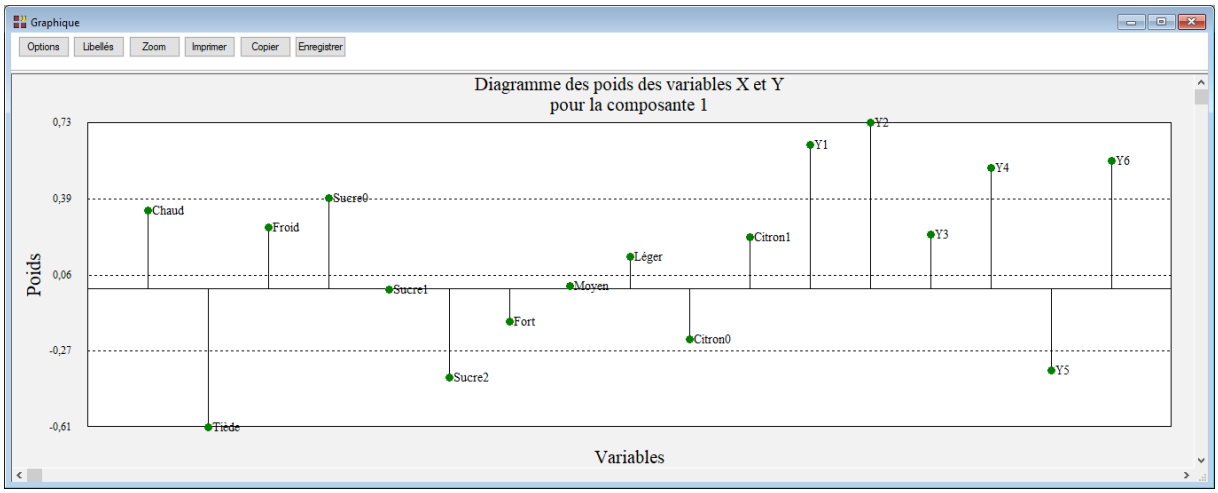
		Composante 1	Composante 2	Composante 3	Composante 4	Composante 5	Composante 6	Composante 7
1								
2	**Q2**							
3								
4	**Q2 = pourcentage de la variation totale des X et du Y prévue**							
5	**Résultats pour le modèle à 7 composantes**							
6	**Limite Q2 : seuil de significativité de la composante : Q2 global >= 0,0975 = (1-0,95^2)**							
7								
8	******* Un modèle à 3 composante(s) PLS semble adéquat.**							
9	******* Exécuter à nouveau la procédure en demandant ce nombre de composantes.**							
10								
11								
12		Composante 1	Composante 2	Composante 3	Composante 4	Composante 5	Composante 6	Composante 7
13	(y) Y1	0,20891	0,23911	0,33881	-0,08277	-0,00456	0,23000	-0,07667
14	(y) Y2	0,27794	-0,05926	-0,15488	0,26761	0,38973	0,36751	-0,10334
15	(y) Y3	-0,00543	0,53178	0,21866	0,37806	0,18013	0,04100	-0,07724
16	(y) Y4	0,21114	-0,00065	0,44827	0,23841	-0,02210	0,16612	-0,10520
17	(y) Y5	0,35388	0,40282	0,46931	0,09700	-0,02469	-0,15373	-0,09664
18	(y) Y6	0,28257	-0,07150	-0,14765	-0,07420	0,08772	-0,03822	-0,09431
19	Q2 global	0,22150	0,25178	0,21775	0,09717	0,11529	0,06902	-0,09094
20								
21								
22								



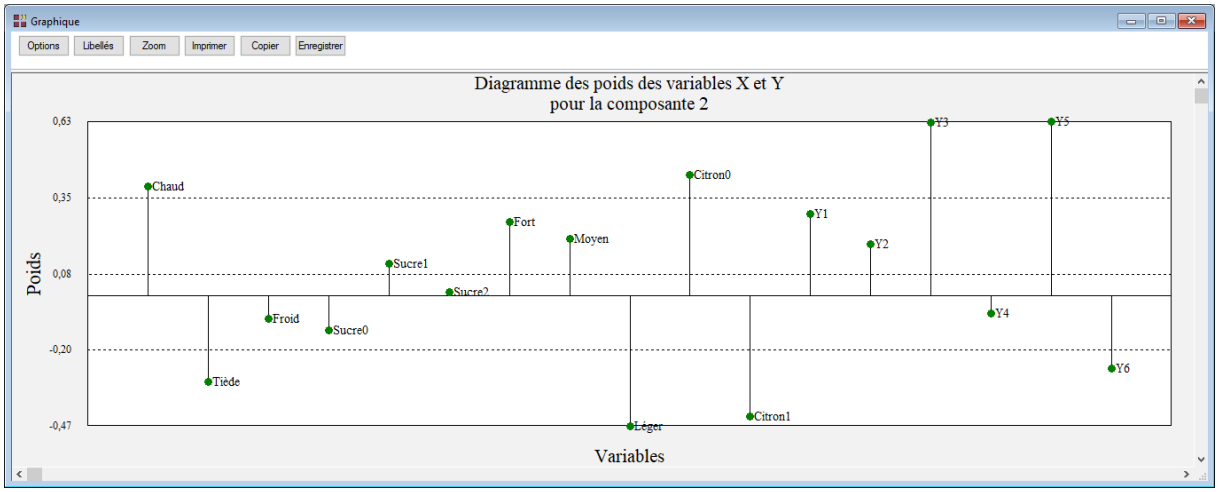




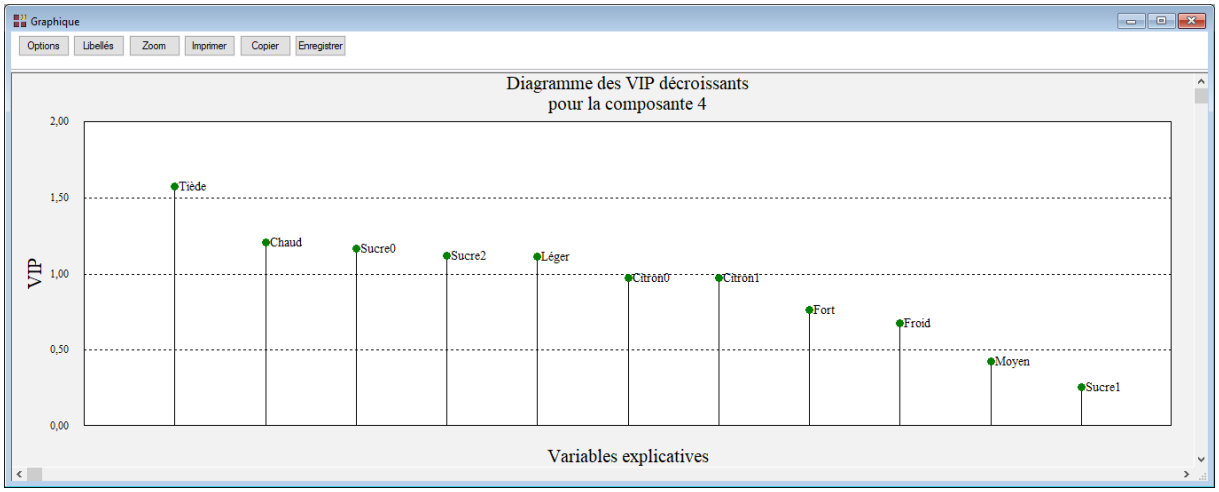
Le diagramme des poids des variables X et Y pour la composante 1 rassemble les juges 1, 2, 4 et 6. Ils apprécient le thé chaud ou froid, sans sucre et rejettent le thé tiède ou très sucré.



La deuxième dimension regroupe les juges 3 et 5. Ils apprécient le thé chaud, fort ou moyen, sans citron. Ils rejettent le thé tiède, léger avec du citron.



Le graphique des VIP pour la dernière composante classe les modalités par ordre d'importance dans la construction de l'ensemble des préférences. Les modalités essentielles ont un VIP supérieur ou voisin de 1 : Tiède, Chaud, Sucre0, Sucre2, Léger, Citron0 et Citron1. Les autres modalités exercent une influence moindre sur les préférences.



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*
cortu	Corrélations entre t et u (PLS2)
corxt	Corrélations entre X et t (PLS2)
corxu	Corrélations entre X et u (PLS2)
corxyt	Corrélations entre X et Y (PLS1)
coryt	Corrélations entre Y et t (PLS2)
coryu	Corrélations entre Y et u (PLS2)
distx	Distances au modèle en X
disty	Distances au modèle en Y
modwgs	Poids modifiés pour le calcul des scores
rawwgs	Poids pour le calcul des scores
regcoefs	Coefficients non standardisés de régression
stdcoefs	Coefficients standardisés de régression
T2	T2 de Hotelling (PLS1)
VIP	VIP - Variable Importance in the Projection - (PLS2)
xloads	Poids des variables X
xscores	Scores sur les composantes (espace des X)
yloads	Poids des variables Y
yscores	Scores sur les composantes (espaces des Y)

libobsapp	Libellés des observations (apprentissage)
prevapp	Valeurs estimées des Y (apprentissage)
residapp	Résidus (apprentissage)
libobsvp	Libellés des observations (validation et prévision)
prevvp	Valeurs estimées des Y (validation et prévision)

Références

Documentation du package R 'plsdepot' (2016)

http://www.gastonsanchez.com/plsdepot

Exemple 1 : Octane1 – données de Cornell (1990)

La régression PLS – Théorie et Pratique – Michel Tenenhaus – Editions Technip

Exemple 2 : Octane2

La régression PLS – Théorie et Pratique – Michel Tenenhaus – Editions Technip
UOP Guided Wave, Inc. – Esbensen, Schönkopf, Midtgaard (1994)

Exemple 3 : Thé

La régression PLS – Théorie et Pratique – Michel Tenenhaus – Editions Technip