



UNIWIN VERSION 9.7.0

CLASSIFICATION PAR CARTES AUTO-ORGANISATRICES

Révision : 02/09/2023

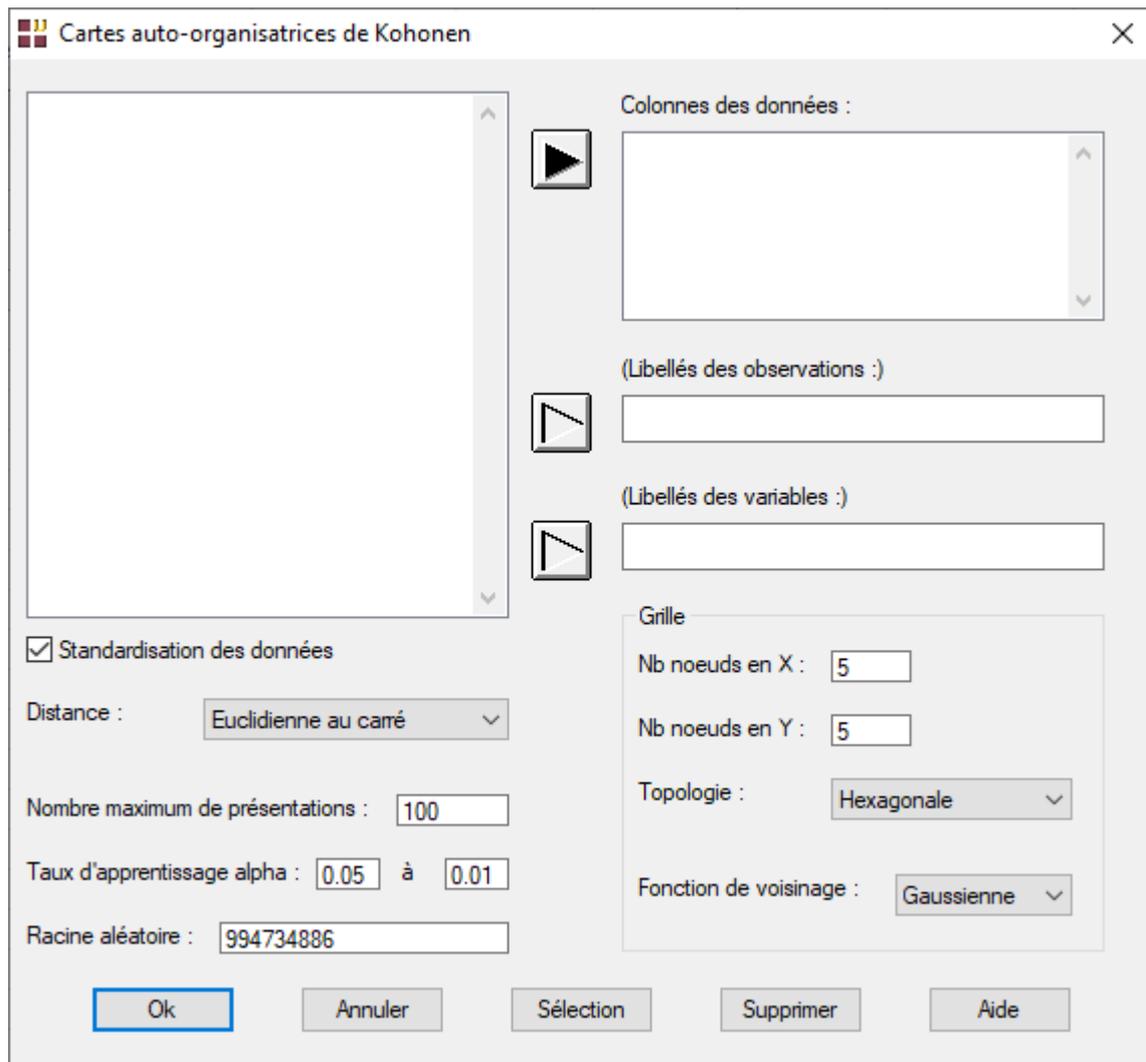
Définition..... 1
Entrée des données 1
Données manquantes 3
Exemple 1 : Fichier Wines 3
L'option Rapports 6
L'option Graphiques 7
Exemple 2 : Fichier Waveform..... 13
Les variables internes créées par la procédure 19
Références 19

Définition

Les cartes auto-adaptatives, cartes auto-organisatrices ou cartes topologiques forment une classe de réseau de neurones artificiels fondée sur des méthodes d'apprentissage non-supervisées. Elles sont souvent désignées par le terme anglais « self organizing maps » (SOM), ou encore cartes de Kohonen du nom du statisticien ayant développé le concept en 1984. La littérature utilise aussi les dénominations : « réseau de Kohonen », « réseau auto-adaptatif » ou « réseau auto-organisé ». Elles sont utilisées pour cartographier un espace réel, c'est-à-dire pour étudier la répartition de données dans un espace à grande dimension. En pratique, cette cartographie peut servir à réaliser des tâches de discrétisation, quantification vectorielle ou classification (source Wikipédia). Cette procédure est basée sur le package R 'kohonen'.

Entrée des données

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



Cette boîte de dialogue permet de préciser les colonnes des données à analyser, les libellés optionnels des observations et des variables.

Elle permet également de préciser :

- les paramètres de la grille :
 - nombre de nœuds en X
 - nombre de nœuds en Y
 - topologie : hexagonale ou rectangulaire
 - fonction de voisinage : gaussienne ou circulaire
- si les données sont standardisées ou non
- la distance utilisée : euclidienne au carré, euclidienne, valeur absolue
- le nombre maximum de présentations
- la plage pour le taux d'apprentissage alpha
- la racine aléatoire

Données manquantes

Les données manquantes ne sont pas autorisées dans les colonnes des données.

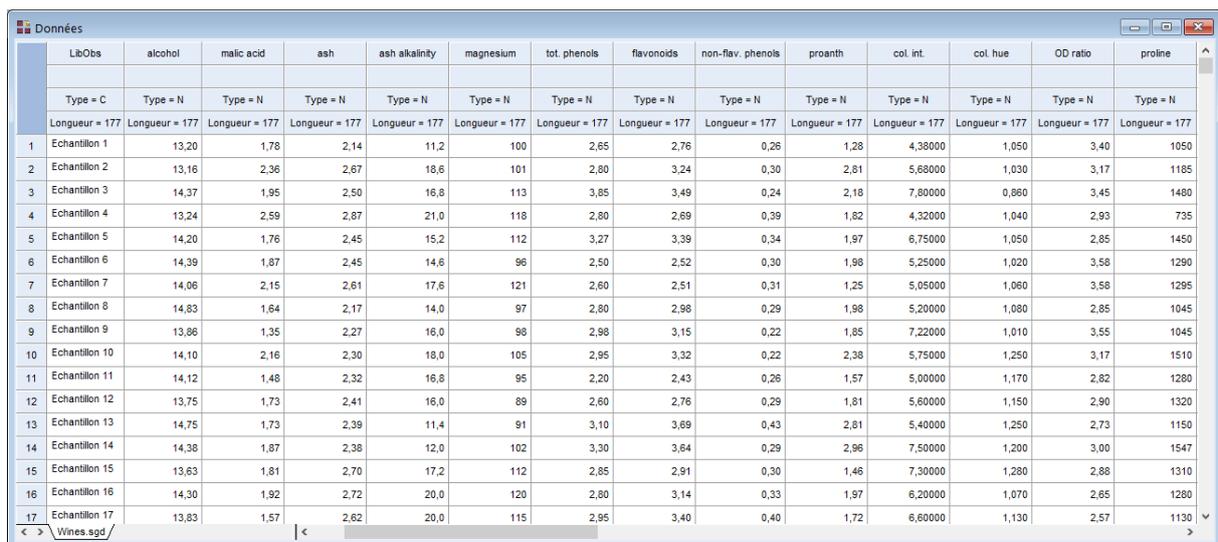
Exemple 1 : Fichier Wines

Les données sont le résultat de l'analyse chimique de 177 échantillons de vins issus de trois cultivars différents et provenant d'une même région en Italie (source : <https://archive.ics.uci.edu/ml/datasets/wine>). Elles sont constituées des treize caractéristiques chimiques et spectroscopiques suivantes :

alcohol
malic acid
ash
ash alkalinity
magnesium
tot. phenols
flavonoids
non-flav. phenols
proanth
col. int.
col. hue
OD ratio
proline

Alcool
Acide malique
Cendres
Alcalinité des cendres
Magnésium
Total des phénols
Flavonoïdes
Phénols non-flavonoïdes
Proanthocyanidines
Intensité de couleur
Teinte de couleur
Rapport OD
Proline

Les données du tableur sont contenues dans le fichier Wines.sgd. Une partie du fichier des données est montrée ci-dessous.



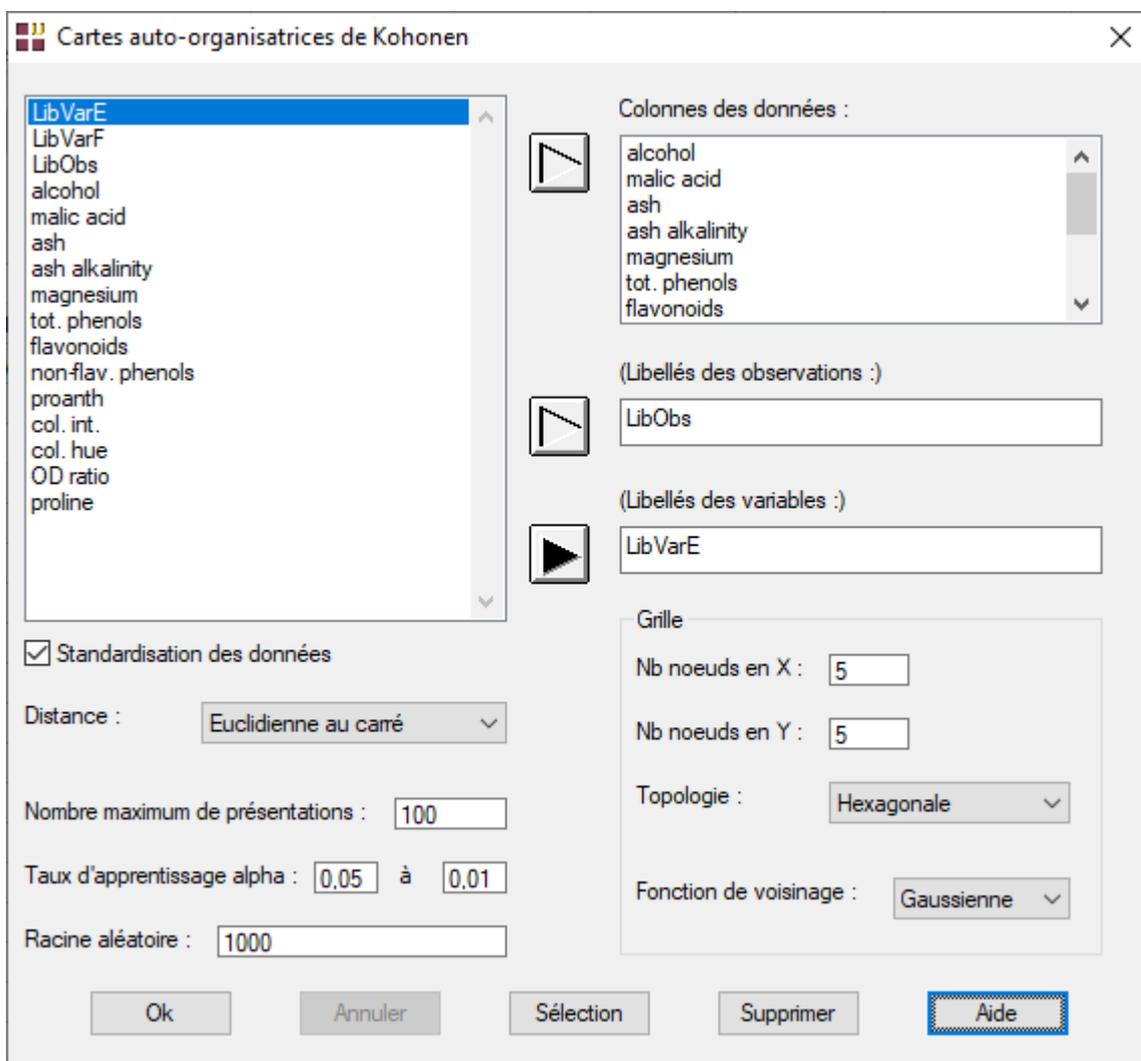
LibObs	alcohol	malic acid	ash	ash alkalinity	magnesium	tot. phenols	flavonoids	non-flav. phenols	proanth	col. int.	col. hue	OD ratio	proline
Type = C	Type = N	Type = N	Type = N	Type = N	Type = N	Type = N	Type = N	Type = N	Type = N	Type = N	Type = N	Type = N	Type = N
1	13,20	1,78	2,14	11,2	100	2,65	2,76	0,26	1,28	4,38000	1,050	3,40	1050
2	13,16	2,36	2,67	18,6	101	2,80	3,24	0,30	2,81	5,68000	1,030	3,17	1185
3	14,37	1,95	2,50	16,8	113	3,85	3,49	0,24	2,18	7,80000	0,860	3,45	1480
4	13,24	2,59	2,87	21,0	118	2,80	2,69	0,39	1,82	4,32000	1,040	2,93	735
5	14,20	1,76	2,45	15,2	112	3,27	3,39	0,34	1,97	6,75000	1,050	2,85	1450
6	14,39	1,87	2,45	14,6	96	2,50	2,52	0,30	1,98	5,25000	1,020	3,58	1290
7	14,06	2,15	2,61	17,6	121	2,60	2,51	0,31	1,25	5,05000	1,060	3,58	1295
8	14,83	1,64	2,17	14,0	97	2,80	2,98	0,29	1,98	5,20000	1,080	2,85	1045
9	13,86	1,35	2,27	16,0	98	2,98	3,15	0,22	1,85	7,22000	1,010	3,55	1045
10	14,10	2,16	2,30	18,0	105	2,95	3,32	0,22	2,38	5,75000	1,250	3,17	1510
11	14,12	1,48	2,32	16,8	95	2,20	2,43	0,26	1,57	5,00000	1,170	2,82	1280
12	13,75	1,73	2,41	16,0	89	2,60	2,76	0,29	1,81	5,60000	1,150	2,90	1320
13	14,75	1,73	2,39	11,4	91	3,10	3,69	0,43	2,81	5,40000	1,250	2,73	1150
14	14,38	1,87	2,38	12,0	102	3,30	3,64	0,29	2,96	7,50000	1,200	3,00	1547
15	13,63	1,81	2,70	17,2	112	2,85	2,91	0,30	1,46	7,30000	1,280	2,88	1310
16	14,30	1,92	2,72	20,0	120	2,80	3,14	0,33	1,97	6,20000	1,070	2,65	1280
17	13,83	1,57	2,62	20,0	115	2,95	3,40	0,40	1,72	6,60000	1,130	2,57	1130

L'objectif est de réduire la dimensionalité de ces données à treize dimensions à un espace à deux dimensions. Les échantillons sont d'abord regroupés en nœuds d'échantillons similaires. Ces nœuds sont ensuite affichés sur la carte à deux dimensions dans laquelle les nœuds similaires sont les uns à côtés des autres.

Cliquons sur l'icône 'SOM' dans le ruban 'Décrire' pour renseigner la boîte de dialogue d'entrée des données comme montré ci-après.

Sélectionnons les treize caractéristiques (de alcohol à proline), standardisons ces données et précisons les libellés pour les observations (LibObs) et pour les variables (LibVarE).

Choisissons une grille 5 x 5 de topologie hexagonale avec une fonction de voisinage gaussienne.



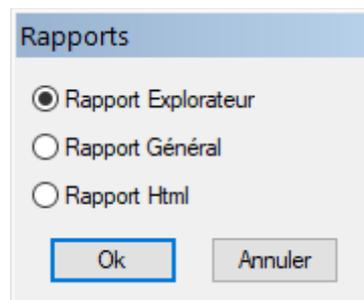
Conservons les valeurs par défaut pour la distance (euclidienne au carré), le nombre de présentations du jeu de données à l'algorithme (100) et la plage pour le taux d'apprentissage (décroissance linéaire de 0,05 à 0,01).

La racine aléatoire est utilisée pour initialiser aléatoirement les poids des nœuds. Nous la fixons pour cet exemple à 1000.

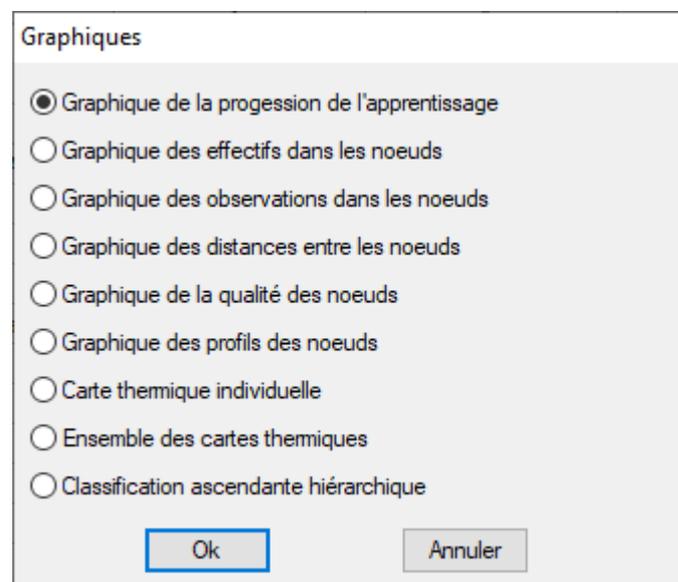
Cliquons sur Ok. UNIWIN débute les calculs. Après quelques instants, la fenêtre Rapports et Graphiques s'affiche.

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.



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.

Enregistrement des résultats (1/1)

Enregistrer

Progression de l'apprentissage

Profils des noeuds

Noeuds des observations

Effectifs dans les noeuds

Distances aux profils

Noms attribués aux variables cibles

progression

profils_1

noeudsobs

effectifs

distances

Ok Plus Tout Annuler

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 ci-après trois exemples du rapport pour l'analyse SOM : Explorateur, Général, HTML.

Ce rapport affiche :

- Effectifs dans les nœuds
- Affectations des nœuds dans les classes
- Affectations des observations dans les classes
- Profils des nœuds

Rapports et Graphiques

Rapport SOM

- Effectifs dans les noeuds
- Affectations des noeuds
- Affectations des observations**
- Profils des noeuds

	1	2	3	4	5	6	7	8
1								
2	AFFECTATIONS DES OBSERVATIONS DANS LES CLASSES (CAH)							
3	Classe = classe d'appartenance de l'observation							
4	Noeud = noeud d'appartenance de l'observation							
5	Distance = distance de l'observation au profil de son noeud							
6								
7								
8		Classe	Noeuds	Distance				
9	Echantillon 1	1	6	3,84786				
10	Echantillon 2	1	21	3,74485				
11	Echantillon 3	1	16	3,62386				
12	Echantillon 4	1	22	1,39484				
13	Echantillon 5	1	16	0,90936				
14	Echantillon 6	1	11	1,65330				
15	Echantillon 7	1	17	3,38353				
16	Echantillon 8	1	11	3,36519				
17	Echantillon 9	1	11	1,95375				
18	Echantillon 10	1	16	2,28168				
19	Echantillon 11	1	11	2,38098				
20	Echantillon 12	1	11	1,45464				
21	Echantillon 13	1	16	6,54256				
22	Echantillon 14	1	16	5,38731				

Rapport Explorateur

Rapports et Graphiques

1		2	3	4	5	6	7	8	9	10	11	12
2	(C) UNIWIN version 9.5.0											
3												
4	DATE : 29/12/2022											
5	ORDINATEUR : LAPTOP-8CVD8RT1											
6	UTILISATEUR : Christian											
7	FICHIER(S) DE DONNEES OUVERT(S) : WINES.SGD											
8												
9	RESULTATS DE L'ANALYSE CARTES AUTO-ORGANISATRICES											
10												
11	Sélection :											
12	Aucune											
13												
14	Données standardisées											
15	Grille : 5 noeuds horizontaux par 5 noeuds verticaux											
16	Topologie : hexagonale											
17	Fonction de voisinage : gaussienne											
18	Distance entre les noeuds : euclidienne au carré											
19	Nombre maximum de présentations : 100											
20	Taux d'apprentissage alpha : 0,05 à 0,01											
21	Nombre d'observations : 177											
22	Nombre de variables : 13											

Rapport Général /

Rapports et Graphiques

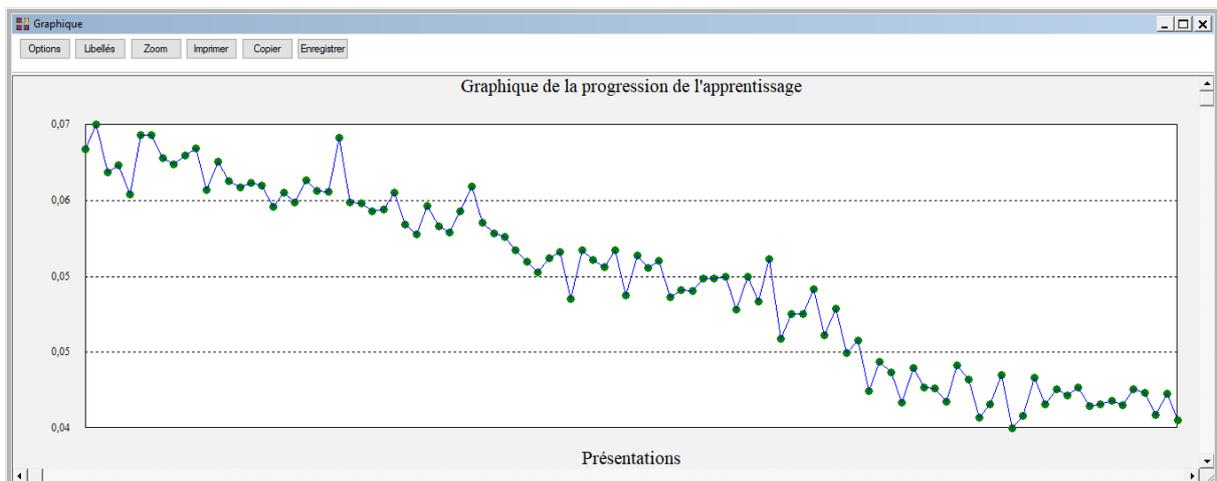
AFFECTATIONS DES OBSERVATIONS DANS LES CLASSES (CAH)

Noeud = noeud d'appartenance de l'observation
Distance = distance de l'observation au profil de son noeud

	Classe	Noeuds	Distance
Echantillon 1	1	6	3,04786
Echantillon 2	1	21	3,74485
Echantillon 3	1	16	3,62386
Echantillon 4	1	22	1,39484
Echantillon 5	1	16	0,90936
Echantillon 6	1	11	1,65330
Echantillon 7	1	17	3,38353
Echantillon 8	1	11	3,36519
Echantillon 9	1	11	1,95375
Echantillon 10	1	16	2,28168
Echantillon 11	1	11	2,38098
Echantillon 12	1	11	1,45464

L'option Graphiques

Graphique de la progression de l'apprentissage



Ce graphique montre la progression de l'apprentissage par rapport aux nombres de présentations.

Si la courbe continue de diminuer avec les présentations et ne se stabilise pas, alors il faut relancer la procédure en précisant un nombre plus important de présentations.

Dans notre exemple, environ 75 présentations auraient été suffisantes.

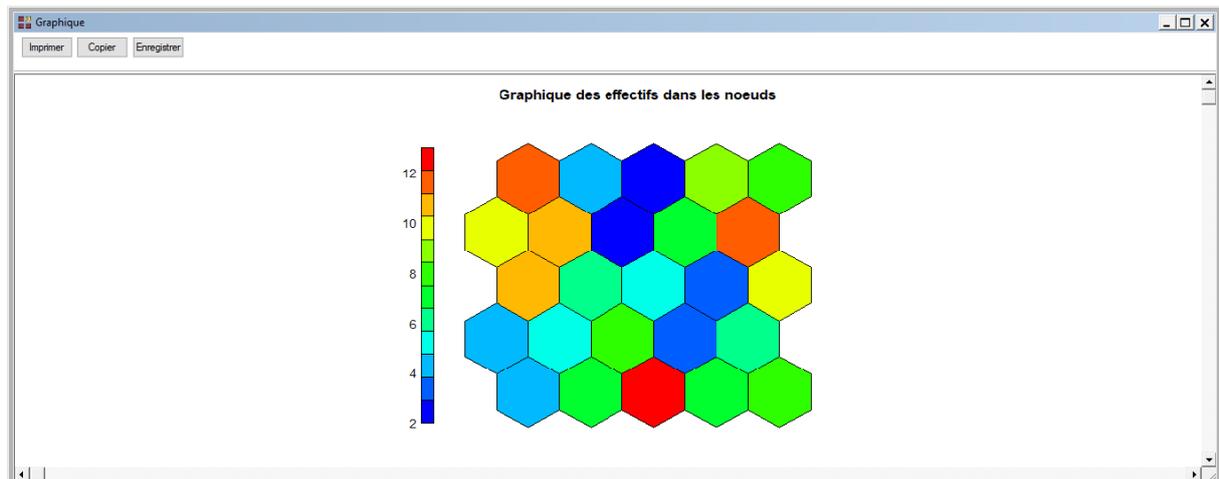
Graphique des effectifs dans les nœuds

Ce graphique affiche les effectifs dans chacun des nœuds.

Ces effectifs permettent d'identifier les zones de fortes densités.

Une répartition homogène est généralement désirée.

La taille de la grille doit être réduite s'il y a de nombreuses cellules vides et augmentée si des zones de trop fortes densités apparaissent.

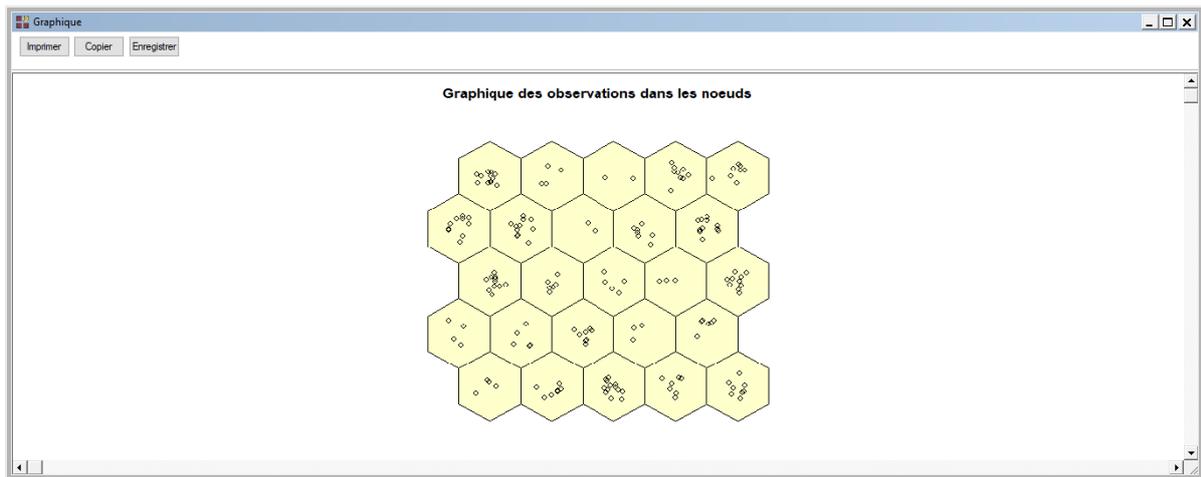


Dans tous les graphiques de cette grille 5x5, la numérotation des nœuds est faite de la façon suivante :

21	22	23	24	25
16	17	18	19	20
11	12	13	14	15
6	7	8	9	10
1	2	3	4	5

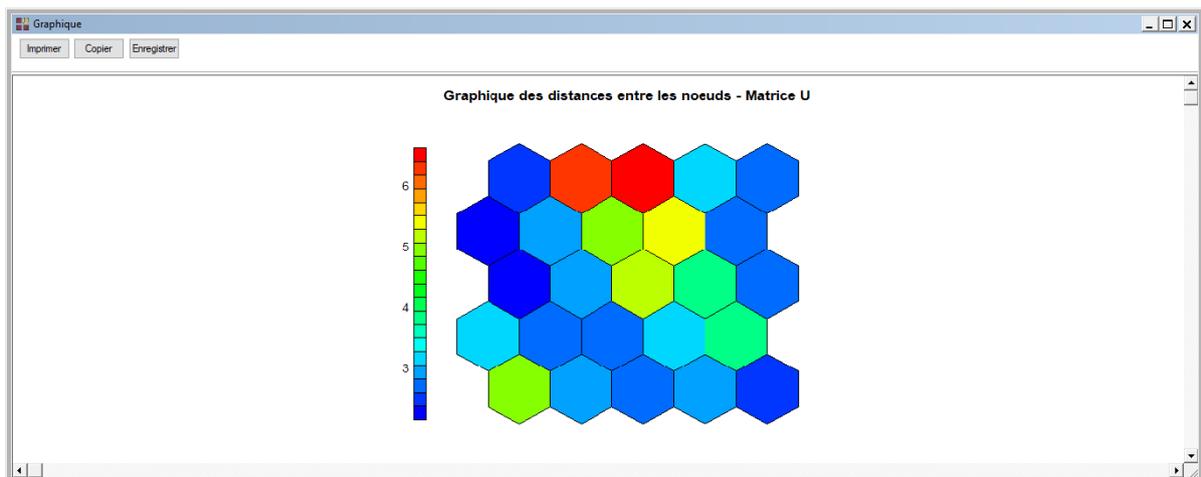
Graphique des observations dans les nœuds

Ce graphique affiche les observations contenues dans chacun des nœuds.

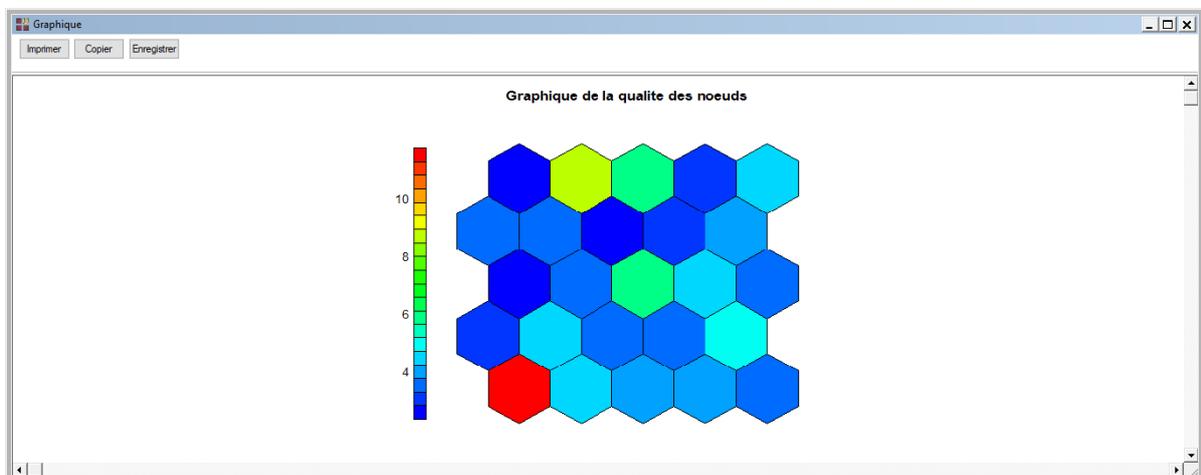


Graphique des distances entre les nœuds

Ce graphique, appelé 'Matrice U', indique la somme des distances aux voisins immédiats pour chaque nœud. Les faibles distances indiquent des nœuds similaires, les grandes distances des nœuds dissimilaires. Ce graphique peut permettre d'identifier des classes dans la grille.



Graphique de la qualité des nœuds

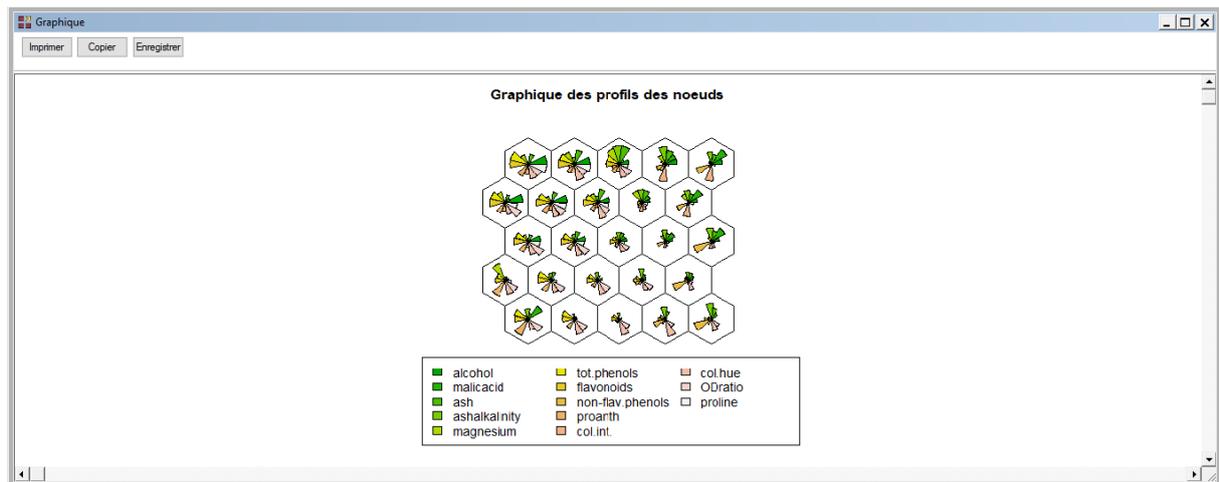


Ce graphique affiche pour chaque nœud la distance moyenne des observations affectées à ce nœud.

Plus les distances sont faibles, plus les observations sont bien représentées par les profils des nœuds.

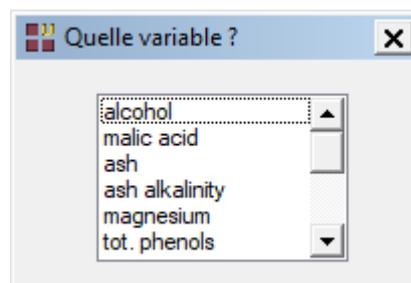
Graphique des profils des nœuds (segments)

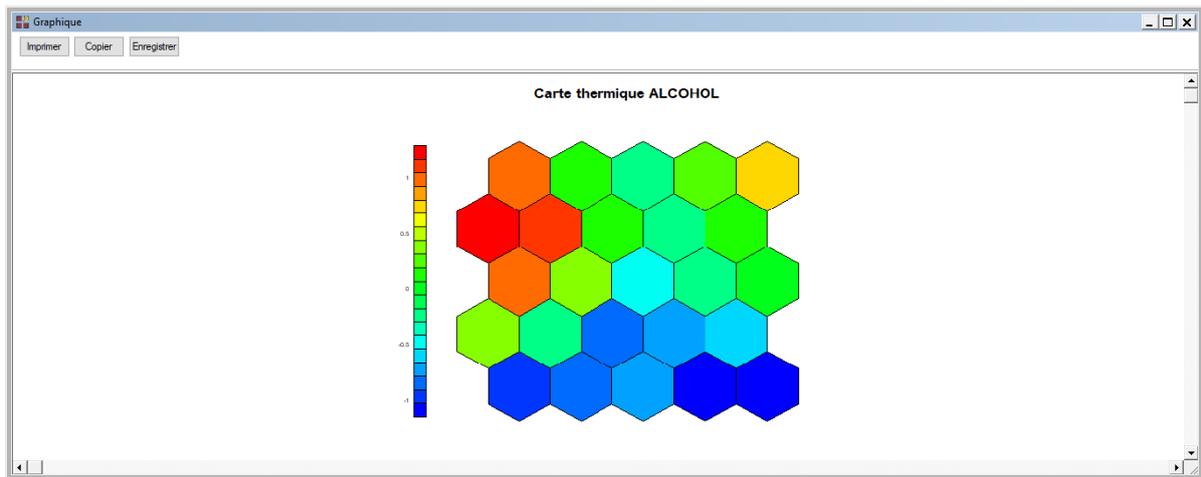
Ce graphique (ici sous la forme d'éventails) permet d'établir le rôle des variables dans la définition des différentes zones qui composent la grille.



Carte thermique individuelle

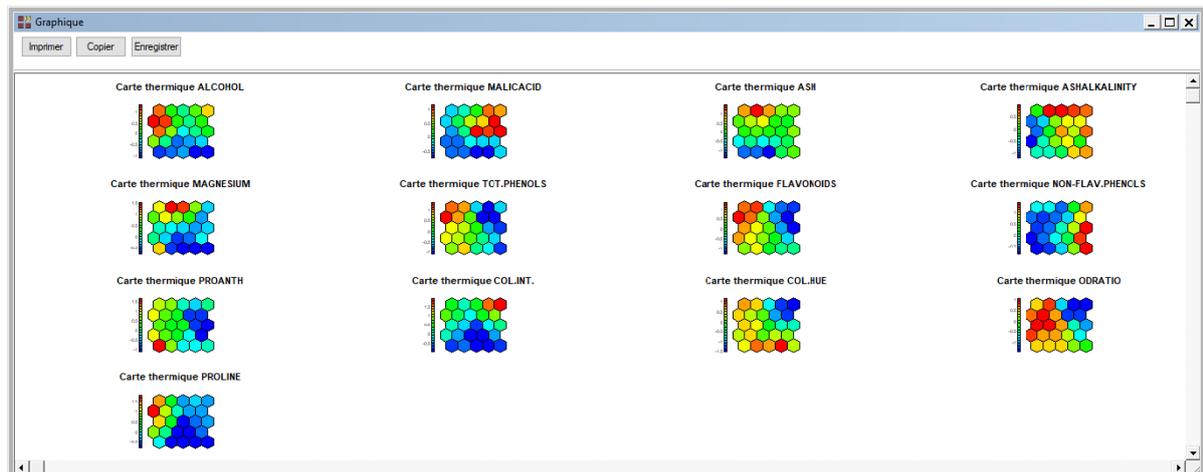
Ce graphique permet de visualiser les valeurs normalisées de la variable sélectionnée dans chaque nœud de la carte.





Ensemble des cartes thermiques

Ce graphique permet de visualiser les valeurs normalisées de l'ensemble des variables dans chaque nœud de la carte.

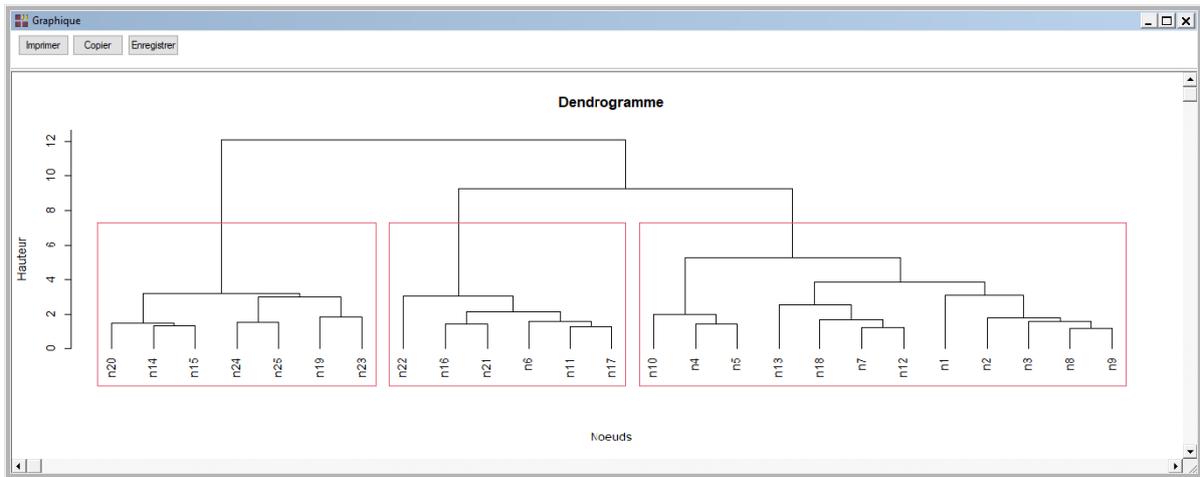
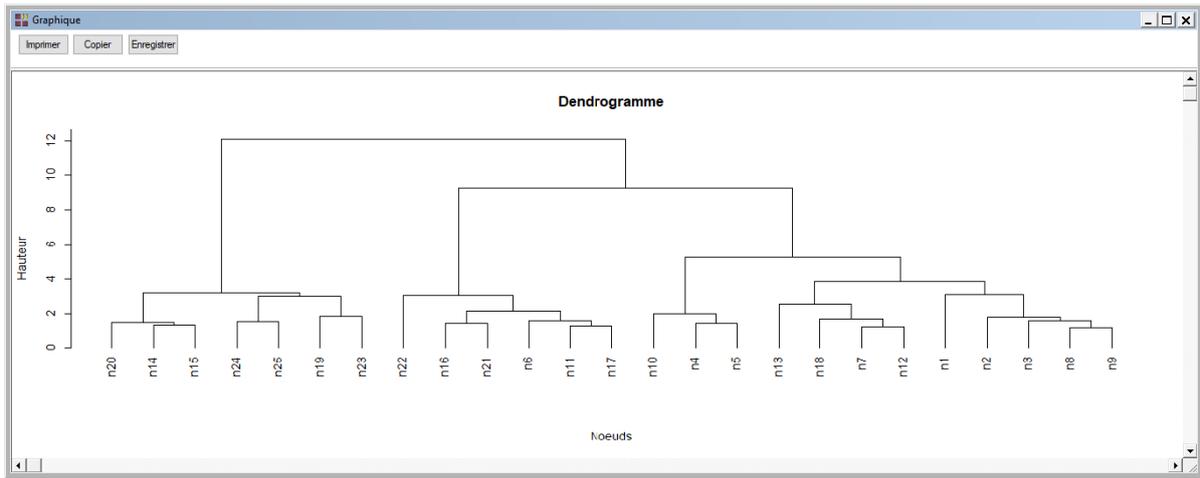


Dendrogramme (Classification ascendante hiérarchique)

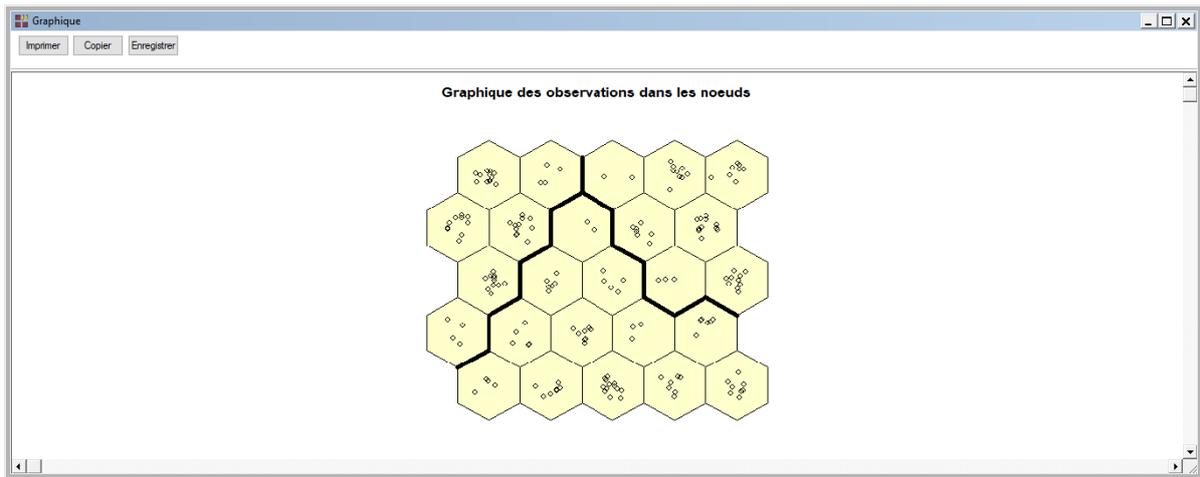
Ce graphique illustre le regroupement des nœuds issu d'une classification ascendante hiérarchique par la méthode de Ward en utilisant les distances entre les nœuds dans la grille (matrice U). Trois classes ont été demandées.

Ce graphique illustre le regroupement des nœuds issu d'une classification ascendante hiérarchique par la méthode de Ward en utilisant les distances entre les nœuds dans la grille (matrice U). Trois classes ont été demandées.

Note : Pour visualiser dans le rapport les affectations des nœuds ou des observations dans les classes, il faut redemander la génération du rapport.



Les frontières entre les classes formées sont alors affichées dans tous les graphiques de la grille.



Les nœuds adjacents sont regroupés dans la même classe. Cela est normal en considération des propriétés de proximités de la grille. Des hiatus peuvent apparaître parfois à cause des contraintes de l'algorithme de classification.

Exemple 2 : Fichier Waveform

Les données Waveform, décrites dans le livre de Breiman (1984), sont composées de 5000 observations et de 21 descripteurs nommés V1 à V21. Une partie de ce fichier est montrée ci-dessous :

Sélection	LibObs	LibVar	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
Type = N	Type = C	Type = C	Type = N										
Longueur = 5000	Longueur = 5000	Longueur = 21	Longueur = 5000										
1	1 Obs 1	V1	-1,23	-1,56	-1,75	-0,28	0,80	2,22	0,85	0,21	-0,20	0,89	1,1
2	1 Obs 2	V2	-0,69	2,43	0,61	2,08	2,30	3,25	5,52	4,55	2,97	2,22	2,6
3	0 Obs 3	V3	-0,12	-0,94	1,29	2,59	2,42	3,55	4,94	3,25	1,90	2,07	0,6
4	0 Obs 4	V4	0,86	0,29	2,19	-0,02	1,13	2,51	2,37	5,45	5,45	4,84	4,6
5	1 Obs 5	V5	1,16	0,37	0,40	-0,59	2,66	1,00	2,69	4,06	5,34	3,53	4,6
6	0 Obs 6	V6	0,00	0,77	1,32	0,29	-1,28	0,84	1,60	1,55	2,93	4,76	5,6
7	0 Obs 7	V7	0,87	1,07	-0,65	1,46	0,84	2,70	3,67	2,94	3,81	5,20	8,1
8	0 Obs 8	V8	-0,22	-0,91	-1,18	0,35	-1,92	-1,59	1,91	0,75	1,72	2,02	3,6
9	1 Obs 9	V9	-1,11	-1,14	-0,89	0,00	0,53	0,44	0,24	2,15	1,64	1,75	3,6
10	0 Obs 10	V10	-0,75	1,10	-1,90	1,43	0,47	0,40	0,86	3,51	2,62	4,50	6,6
11	0 Obs 11	V11	0,14	-1,16	1,42	2,28	3,10	3,15	3,49	4,54	1,40	3,41	3,4
12	0 Obs 12	V12	1,32	-0,40	-0,69	4,17	3,86	4,00	5,24	3,88	2,17	1,82	3,6
13	0 Obs 13	V13	-0,93	2,46	1,20	2,97	2,91	3,57	3,66	4,19	3,22	3,53	2,4
14	0 Obs 14	V14	-1,06	0,59	1,01	3,33	2,05	3,20	4,70	4,21	4,73	2,22	2,6
15	1 Obs 15	V15	1,86	0,37	-0,35	0,74	0,84	0,21	1,97	1,52	1,85	2,39	3,6
16	0 Obs 16	V16	-0,51	-0,46	0,35	-1,67	0,26	2,45	-0,09	2,03	0,79	1,42	1,1

Dans cet exemple, nous utiliserons la colonne Sélection pour définir une sélection aléatoire de 1500 observations parmi les 5000.

Cliquons sur l'icône 'SOM' dans le ruban 'Décrire' pour renseigner la boîte de dialogue d'entrée des données comme montré ci-après.

Nous sélectionnons les colonnes de données V1 à V21, définissons les libellés optionnels des observations et des variables.

Les colonnes données sont standardisées.

La distance choisie la distance euclidienne au carré.

Une grille 15 x 10 de topologie rectangulaire est définie avec une fonction de voisinage circulaire.

Le nombre de présentations est fixé à 100 et la racine aléatoire est définie à 100.

Cliquons sur le bouton 'Sélection' pour définir la sélection des 1500 observations parmi les 5000.

Cartes auto-organisatrices de Kohonen

Sélection
 LibObs
LibVar
 V1
 V2
 V3
 V4
 V5
 V6
 V7
 V8
 V9
 V10
 V11
 V12
 V13
 V14
 V15
 V16
 V17

Colonnes des données :
 V1
 V2
 V3
 V4
 V5
 V6
 V7

(Libellés des observations :)
 LibObs

(Libellés des variables :)
 LibVar

Grille
 Nb noeuds en X : 15
 Nb noeuds en Y : 10
 Topologie : Rectangulaire
 Fonction de voisinage : Circulaire

Standardisation des données

Distance : Euclidienne au carré

Nombre maximum de présentations : 100

Taux d'apprentissage alpha : 0,05 à 0,01

Racine aléatoire : 1000

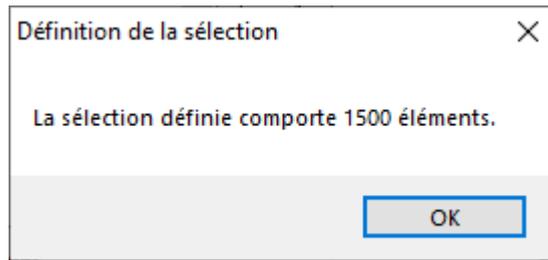
Ok Annuler Sélection Supprimer Aide

Définition de la sélection

Et Sélection = 1

Liaison	Variable	Relation	Valeur ou variable
Et	CLASSE	=	CLASSE
Et non	LibObs	<>	LibObs
Ou	LibVar	<	LibVar
Ou non	Sélection	<=	Sélection
	V1	>	V1
	V10	>=	V10
	V11	début	V11

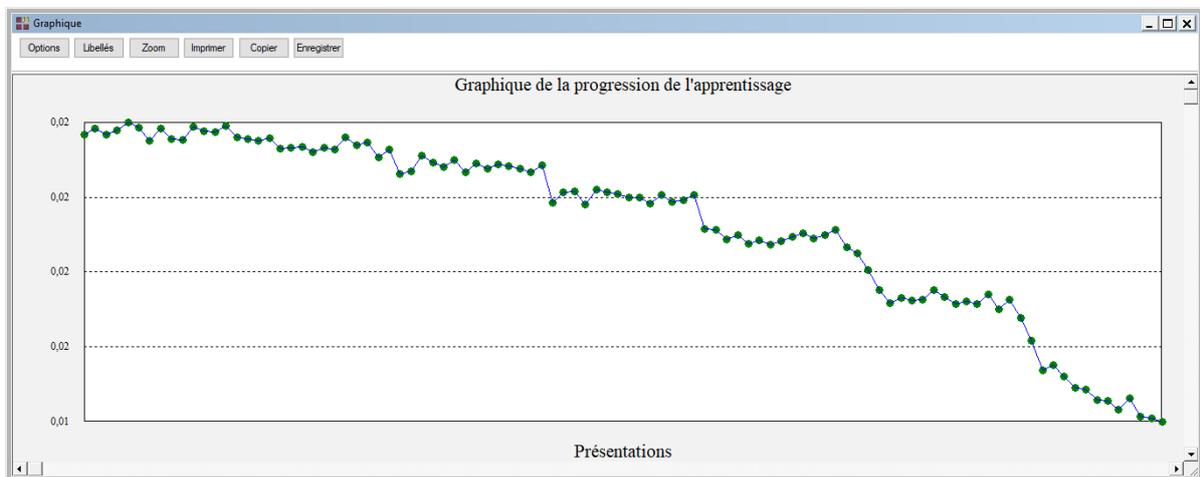
Ok Annuler Ajouter Aide



Après avoir renseigné cette boîte de dialogue, le logiciel débute les calculs. Après quelques instants, la fenêtre Rapports et Graphiques s'affiche.

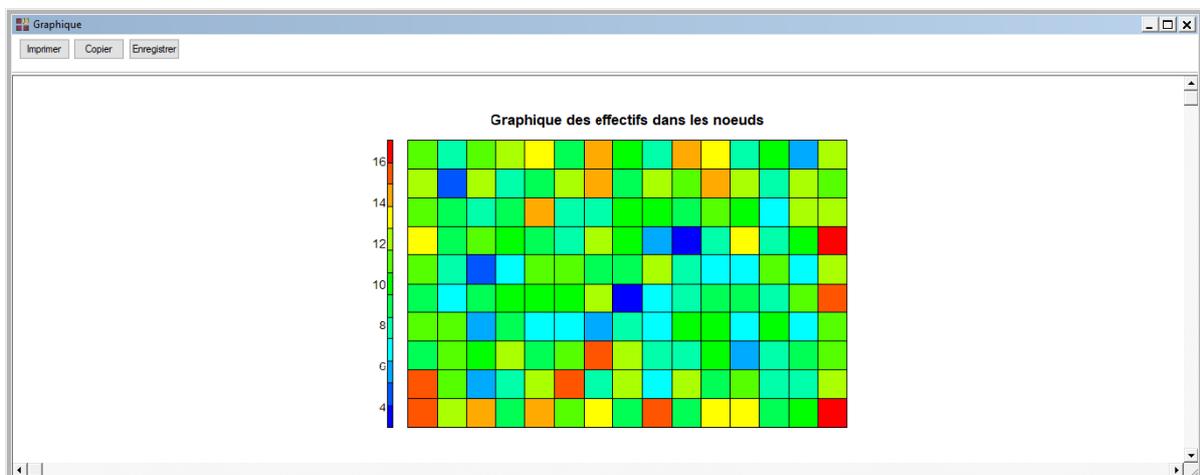
Comme dans l'exemple précédent, le rapport affiche les effectifs dans les nœuds, les affectations des nœuds dans les classes, les affectations des observations dans les classes et les profils des nœuds.

Graphique de la progression de l'apprentissage

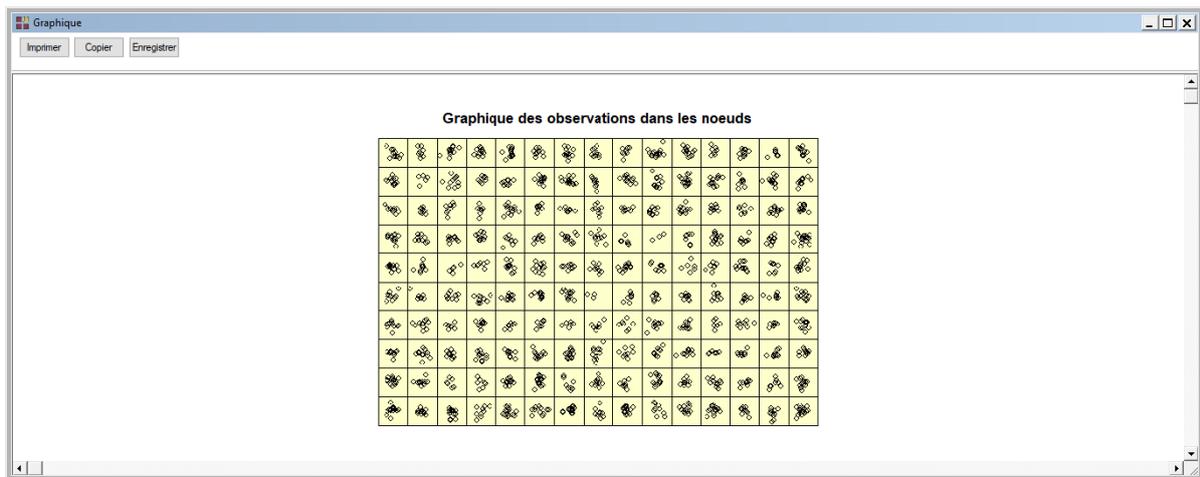


Il pourrait être utile de relancer cette procédure en augmentant le nombre de présentations pour mieux visualiser la stabilisation de la décroissance de cette courbe.

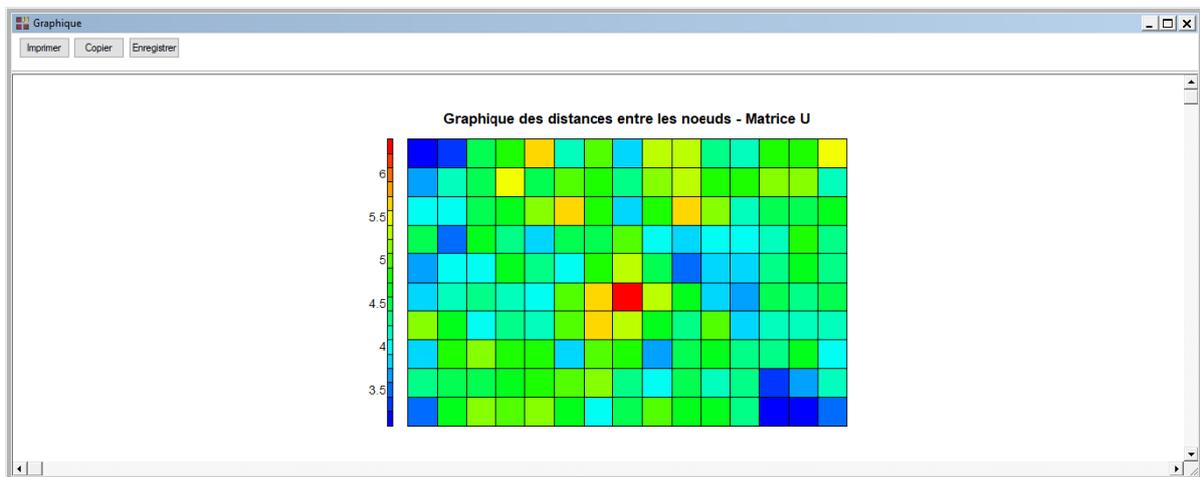
Graphique des effectifs dans les nœuds



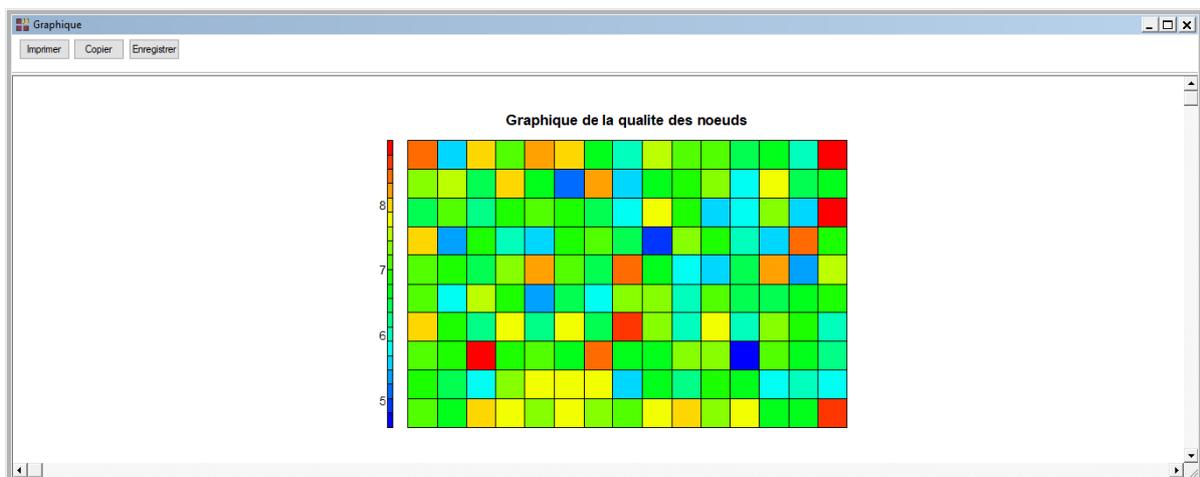
Graphique des observations dans les nœuds



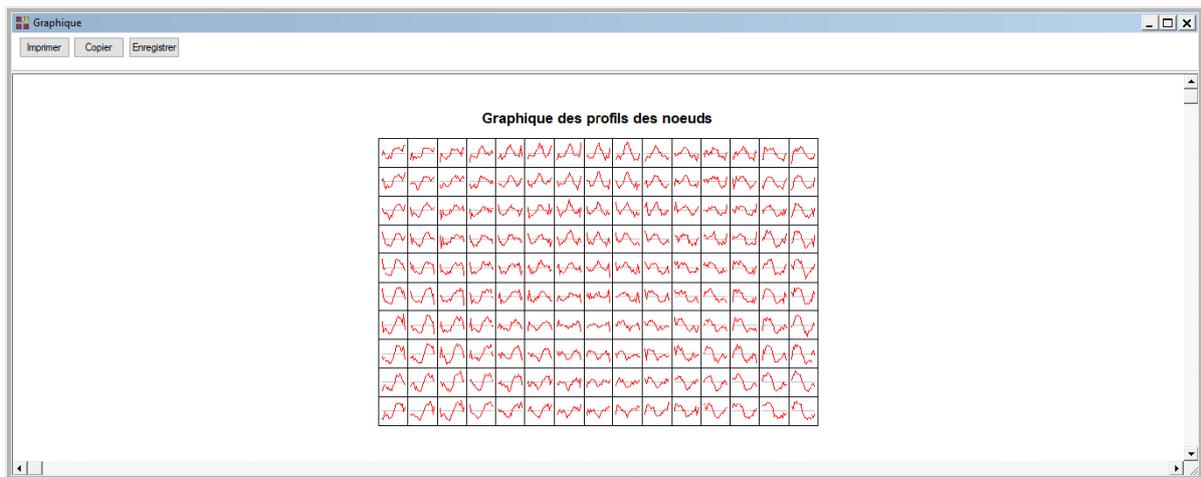
Graphique des distances entre les nœuds (matrice U)



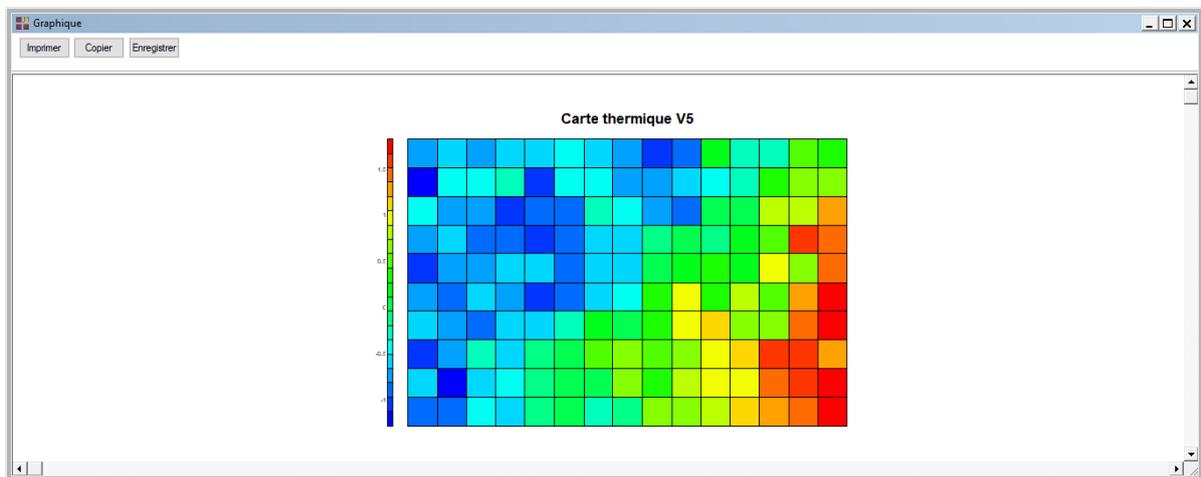
Graphique de la qualité des nœuds



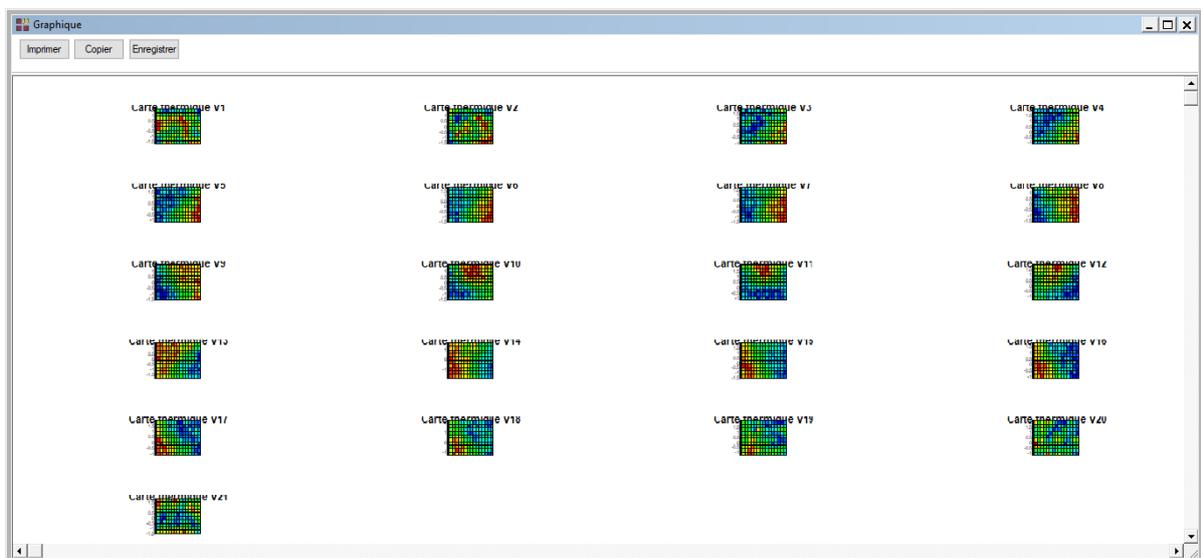
Graphique des profils des nœuds (lignes)



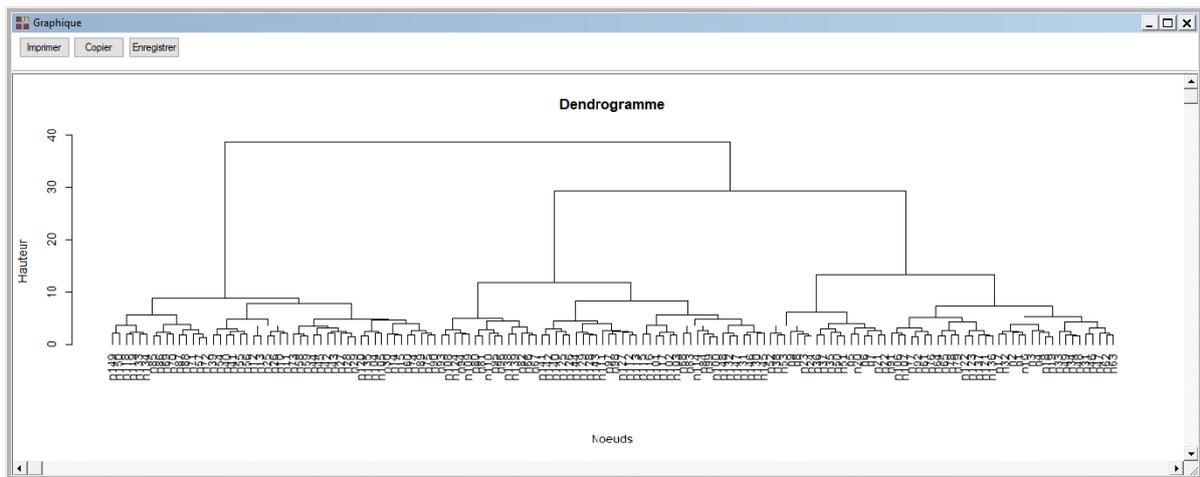
Carte thermique individuelle



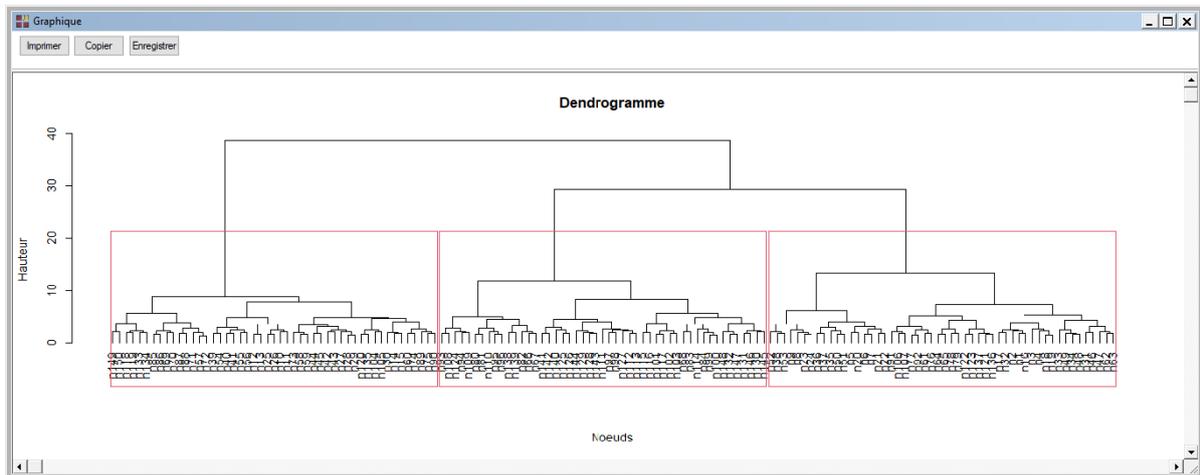
Ensemble des cartes thermiques



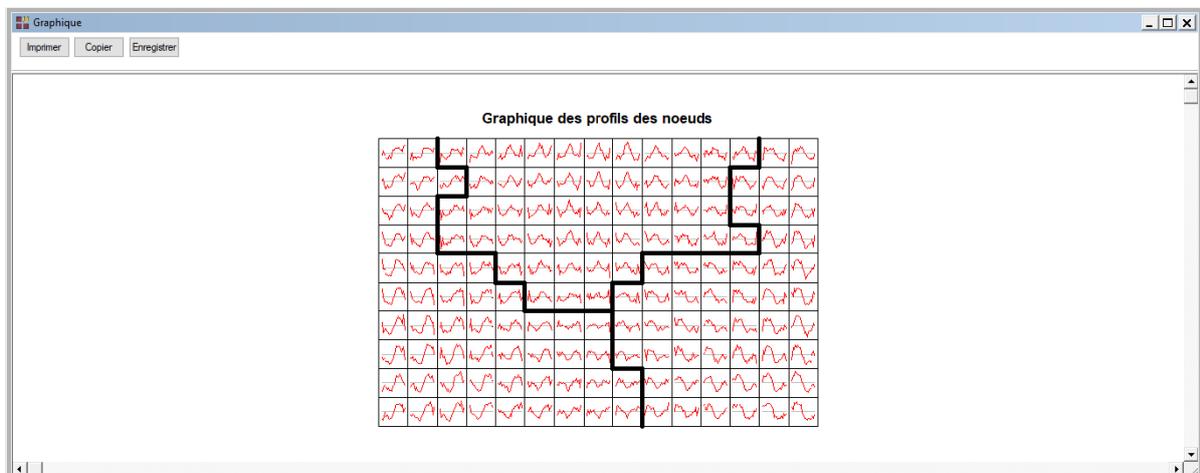
Dendrogramme (Classification ascendante hiérarchique)



Trois classes sont demandées.



Les frontières entre les classes formées sont alors affichées dans tous les graphiques de la grille, comme montré ci-dessous dans le graphique des profils dans les noeuds.



Les variables internes créées par la procédure

Voici la liste des variables internes créées par la procédure.

<i>Variable</i>	<i>Contenu</i>
progression	Progression de l'apprentissage
profils	Profils des noeuds
noeudobs	Nœuds des observations
effectifs	Effectifs dans les nœuds
distances	Distances aux profils

Références

Documentation du package R – 'kohonen' (2019)

<https://cran.r-project.org/web/packages/kohonen/kohonen.pdf>

Exemple 1

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

Exemple 2

Breiman,L., Friedman,J.H., Olshen,R.A., & Stone,C.J. (1984). Classification and Regression Trees. Wadsworth International Group: Belmont, California. (pages 43-49).

<https://archive.ics.uci.edu/ml/datasets/Waveform+Database+Generator+%28Version+2%29>