

UNIWIN VERSION 10.0.0

CLASSIFICATION PAR LA METHODE DBSCAN

Révision : 30/06/2024

Définition.....	1
Définitions.....	2
Données manquantes	3
Entrée des données	3
Exemple 1 : Fichier DBSCAN1	4
L'option Rapports	6
L'option Graphiques	7
Exemple 2 : Fichier DBSCAN2	10
Exemple 3 : Fichier DBSCAN3	13
Exemple 4 : Fichier DBSCAN4	14
Exemple 5 : Fichier IRIS	16
Les variables internes créées par la procédure	18
Références	19

Définition

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) est un algorithme de partitionnement de données proposé en 1996 par Martin Ester, Hans-Peter Kriegel, Jörg Sander et Xiaowei Xu.

Les classes formées correspondent à des régions denses dans l'espace des données séparées par des régions de plus faibles densités de points. L'algorithme DBSCAN repose sur cette notion intuitive de classes et de bruit. Il possède plusieurs avantages par rapport aux autres techniques de partitionnement, notamment sa capacité à créer un nombre de classes non défini a priori, à reconnaître des classes non convexes et à isoler les points suspects.

La procédure affiche un rapport indiquant pour chaque observation sa classe, sa distance au point central, son coefficient de silhouette ainsi qu'une synthèse de la classification. Les graphiques des distances K-NN, des distances d'accessibilité par densité, des coefficients de silhouette et des nuages des points des classes avec ou sans enveloppes sont proposés.

Cette procédure est basée sur le package R 'dbscan'.

Définitions

Epsilon de voisinage

L'épsilon de voisinage d'un point p est l'ensemble des points qui sont dans un rayon epsilon autour de p .

Si epsilon est trop faible, alors aucun point n'est voisin d'un autre. Il n'y a que des points suspects. Si epsilon est trop grand, tous les points sont voisins entre eux et il n'y a qu'une seule classe.

Nombre minimum de points

Des règles usuelles consistent à définir le nombre minimum de points comme soit égal au nombre de colonnes de données +1, soit à deux fois le nombre de colonnes de données.

Si le nombre de points est faible, tous les voisinages sont denses. Il suffit d'un unique voisin commun pour relier deux classes. Si le nombre de points est grand, peu de voisinages sont denses. Beaucoup de points seront des points frontières ou suspects.

Point central

Un point est un point central (cœur) si dans son epsilon de voisinage il y a au minimum un nombre défini de points.

Point frontière

Un point est un point frontière s'il n'est pas un point central mais est dans le voisinage d'un point central. Il est possible de ne pas les inclure dans l'analyse et donc de les considérer comme des points suspects.

Point suspect

Les autres points sont dits points suspects (bruit).

Point directement accessible par densité

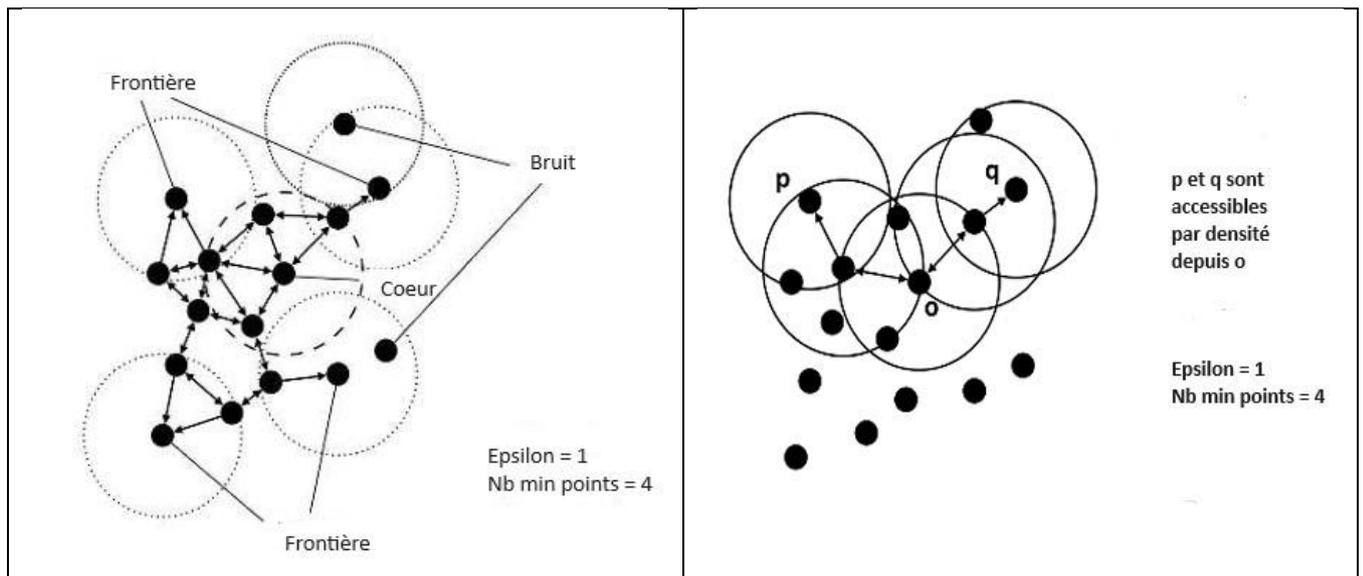
Un point q est directement accessible par densité depuis un point p si le voisinage de p est dense et que q est dans le voisinage de p .

Point accessible par densité

Un point q est accessible par densité s'il existe une suite ordonnée de points

p_1, \dots, p_n telle que $p_1=p$ et $p_n=q$

et que pour tout i , p_{i+1} est directement accessible depuis p_i .



Données manquantes

Dans cette procédure, les données manquantes ne sont pas permises.

Entrée des données

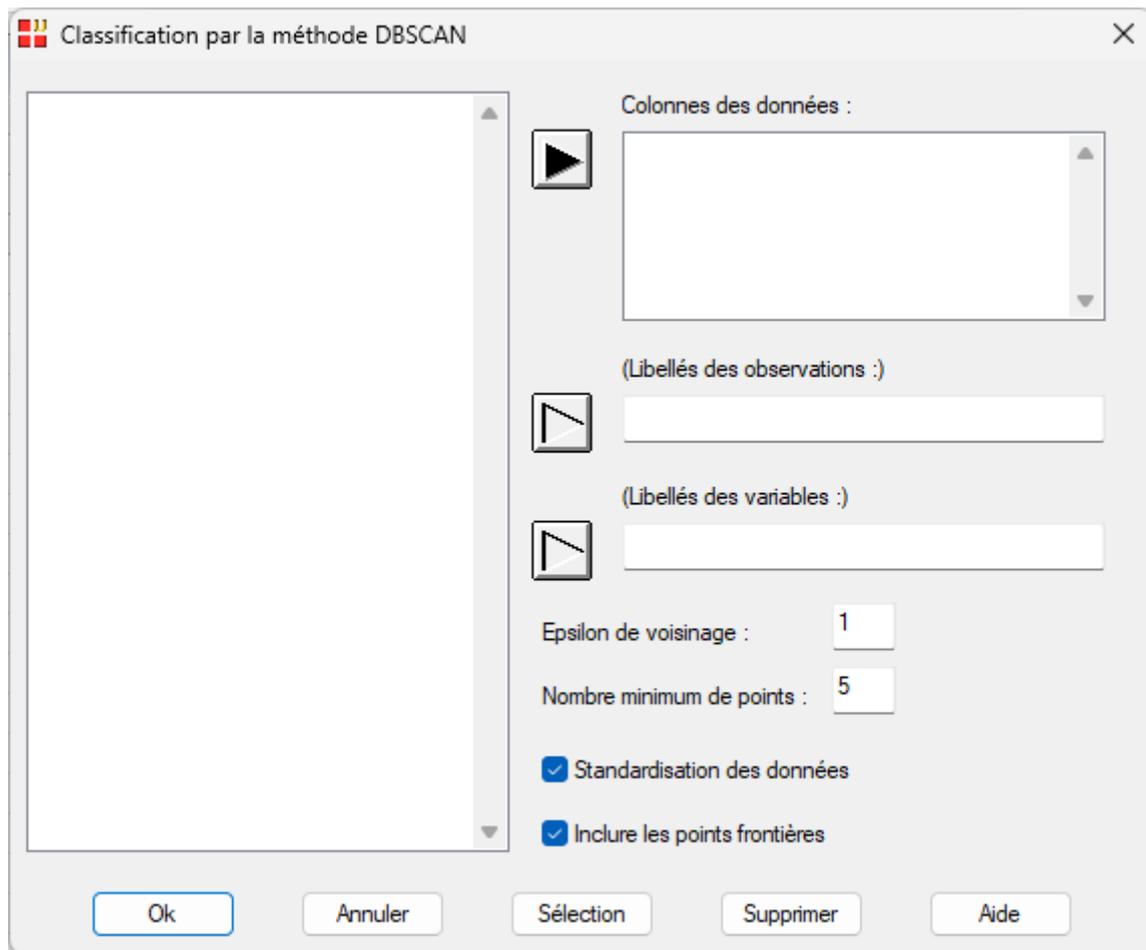
Cliquons sur l'icône DBSCAN dans le ruban Décrire. La boîte de dialogue montrée ci-après s'affiche.

Cette boîte de dialogue permet de définir les colonnes des données (variables quantitatives) utilisées dans l'analyse.

Elle permet, en option, d'indiquer les noms des variables contenant les libellés des observations et les libellés des variables.

Une analyse sur les données d'origine ou sur les données centrées et réduites peut être réalisée.

L'épsilon de voisinage, le nombre minimum de points et l'inclusion ou non des points frontières peuvent être précisés.



Exemple 1 : Fichier DBSCAN1

Nous utiliserons le fichier DBSCAN1 pour illustrer ce premier exemple.

Ce fichier contient 4 ensembles, se recouvrant partiellement, de 100 observations générées à partir de lois gaussiennes légèrement bruitées.

Cliquons sur l'icône DBSCAN dans le ruban Décrire.

Sélectionnons les variables 'X' et 'Y' comme colonnes de données.

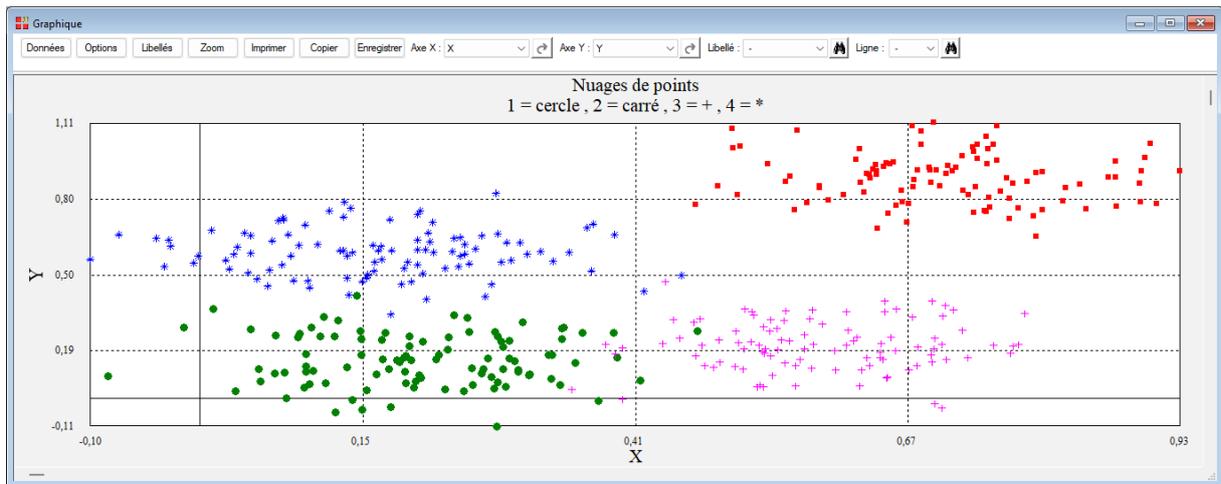
Décochons 'Standardisation des données'.

Le nombre minimum de points choisi est égal au nombre de colonnes de données + 1, donc ici 3.

Concernant la valeur de l'épsilon de voisinage, nous la fixons à 0,06 (voir le graphique des distances K-NN).

Incluons les points frontières.

Visualisons les données dans un graphique 2D.



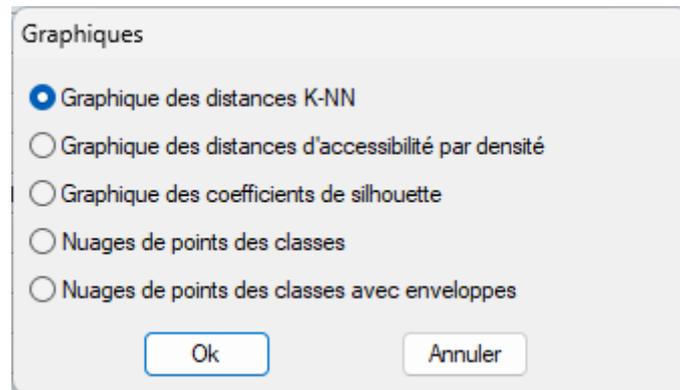
Après avoir renseigné cette boîte de dialogue, UNIWIN débute le calcul de l'Analyse DBSCAN. Après quelques instants, l'écran suivant s'affiche :

	1	2	3	4	5	6	7	8
1								
2	(C) UNIWIN version 10.0.0							
3								
4	DATE : 19/04/2024							
5	ORDINATEUR : LAPTOP-LEGBLO77							
6	UTILISATEUR : cchar							
7	FICHIER(S) DE DONNEES OUVERT(S) : DBSCAN1.SGD							
8								
9	RESULTATS DE L'ANALYSE DBSCAN							
10								
11	Sélection :							
12	Aucune							
13								
14	Nombre d'observations : 400							
15								
16	Colonnes des données :							
17	X							
18	Y							
19								
20	Données non standardisées							
21	Epsilon de voisinage : 0,06							
22	Rapport Explorateur							

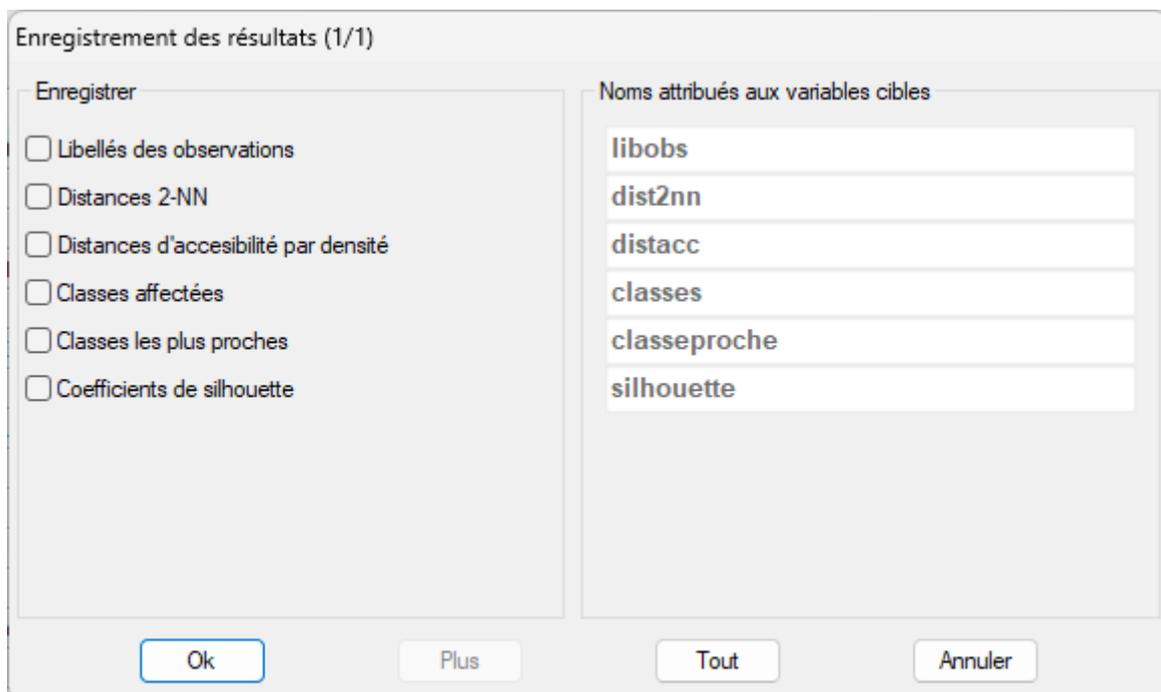
La barre d'outils 'Rapports et Graphiques' permet par l'icône 'Données'  de rappeler la boîte de dialogue d'entrée des données.

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.

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.

Le premier tableau affiche pour chaque observation sa classe d'affectation, sa distance d'accessibilité par densité, sa classe la plus proche et son coefficient de silhouette.

Observation	Classe affectée	Accessibilité par densité	Classe la plus proche	Coefficient de silhouette
o1	1	0,00000	7	-0,08517
o2	2	0,04036	6	-0,36463
o3	1	0,03285	5	0,29511
o4	1	0,05502	7	-0,65119
o5	1	0,03282	7	-0,39376
o6	2	0,03749	6	0,37951
o7	1	0,04409	5	0,07438
o8	3	0,02618	7	-0,27169
o9	1	0,03151	7	0,00501
o10	2	0,03369	6	0,42166
o11	1	0,01856	5	0,21055
o12	3	0,03540	7	-0,00856

Le deuxième tableau affiche une synthèse de la classification. Trois grandes classes sont formées ainsi que quatre petites classes. Quinze observations suspectes sont non affectées.

Classe affectée	Effectif	Coefficient de silhouette
0	15	-0,57616
1	191	0,02509
2	92	0,29835
3	90	0,09450
4	3	0,80423
5	3	0,73538
6	3	0,85572
7	3	0,66013

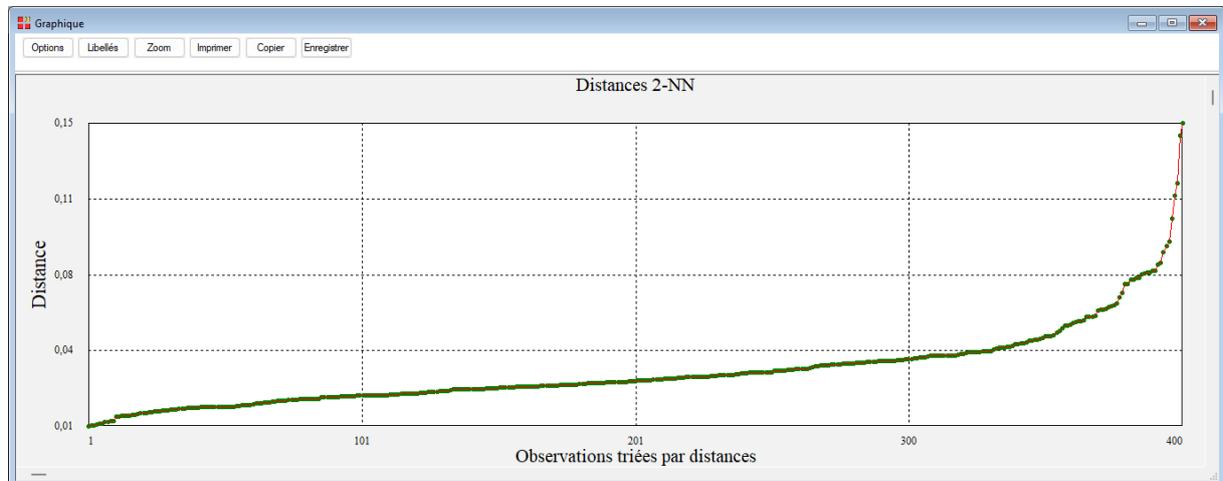
L'option Graphiques

- Graphique des distances K-NN

Cette option affiche l'évolution des distances aux K plus proches voisins.

L'idée est que les observations situées à l'intérieur des classes auront une petite distance car elles sont proches d'autres observations de la même classe, tandis que les observations suspectes seront isolées et auront des distances plutôt grandes.

Le nombre minimum de points pour DBSCAN incluant le point de données, ce qui n'est pas le cas dans K-NN, le nombre de plus proches voisins est égal au nombre minimum de points -1.



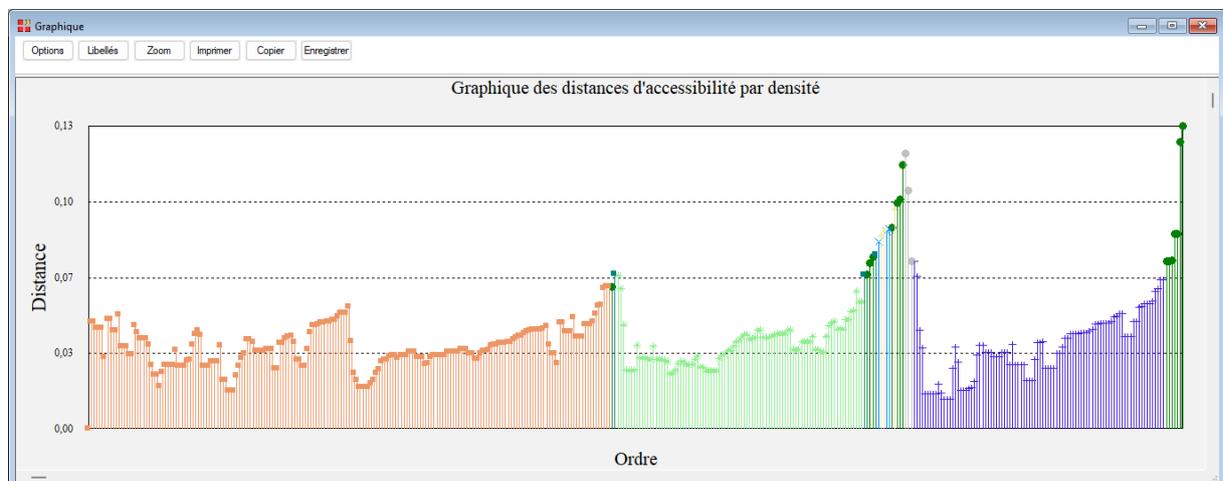
La recherche d'un coude dans ce graphique permet de choisir une valeur pour l'épsilon de voisinage, ici probablement aux alentours de 0,06.

- Graphique des distances d'accessibilité par densité

Ce graphique est basé sur l'algorithme OPTICS (Ordering Points To Identify the Clustering Structure.). Il utilise un epsilon de voisinage égal à la plus grande distance K-NN et un nombre minimum de points égal à 5 par défaut.

OPTICS ordonne les observations de sorte que des observations proches dans l'espace soient voisines dans le graphique. Cela est analogue à l'ordre obtenu par une classification ascendante hiérarchique utilisant la méthode du lien simple (saut minimum).

Les vallées dans ce graphique représentent des zones de fortes densités tandis que les sommets indiquent des observations séparant les classes. Plus les sommets sont élevés, plus les classes sont bien séparées.

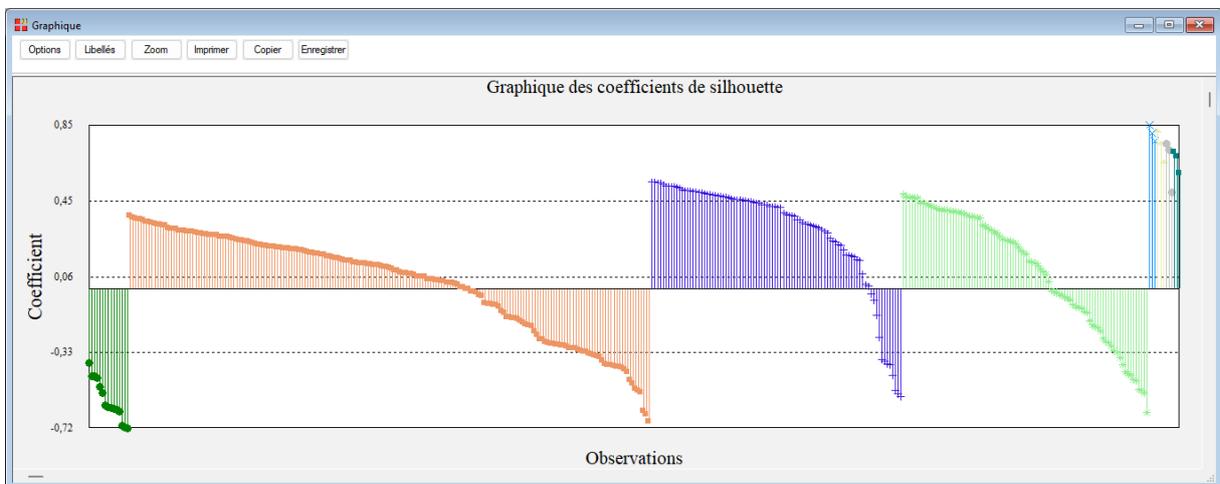


- Graphique des coefficients de silhouette

Le coefficient de silhouette est une mesure de la qualité du partitionnement. Il varie entre -1 et +1.

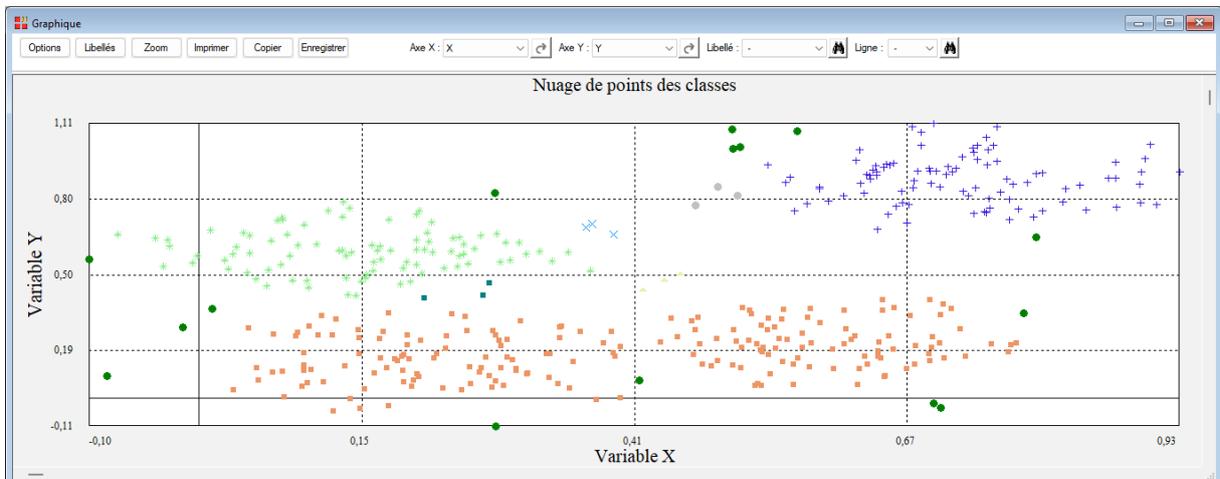
Pour chaque observation, le coefficient de silhouette est la différence entre la distance moyenne avec les observations du même groupe et la distance moyenne avec les observations des autres groupes.

Si cette différence est négative, l'observation est en moyenne plus proche du groupe voisin que du sien : elle est donc mal classée. À l'inverse, si cette différence est positive, l'observation est en moyenne plus proche de son groupe que du groupe voisin : elle est donc bien classée.



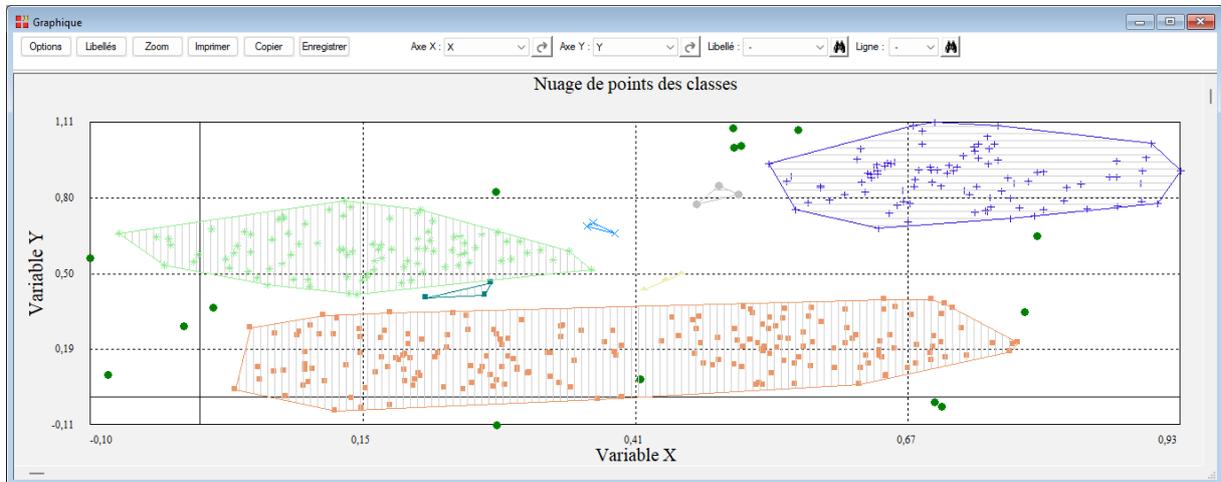
- Nuage de points des classes

Le nuage de points des classes affiche les classes formées. Clairement dans cet exemple, les deux classes en haut du graphique ont bien été reconnues mais les deux classes en bas du graphique ont été fusionnées.



- Nuage de points des classes avec enveloppes

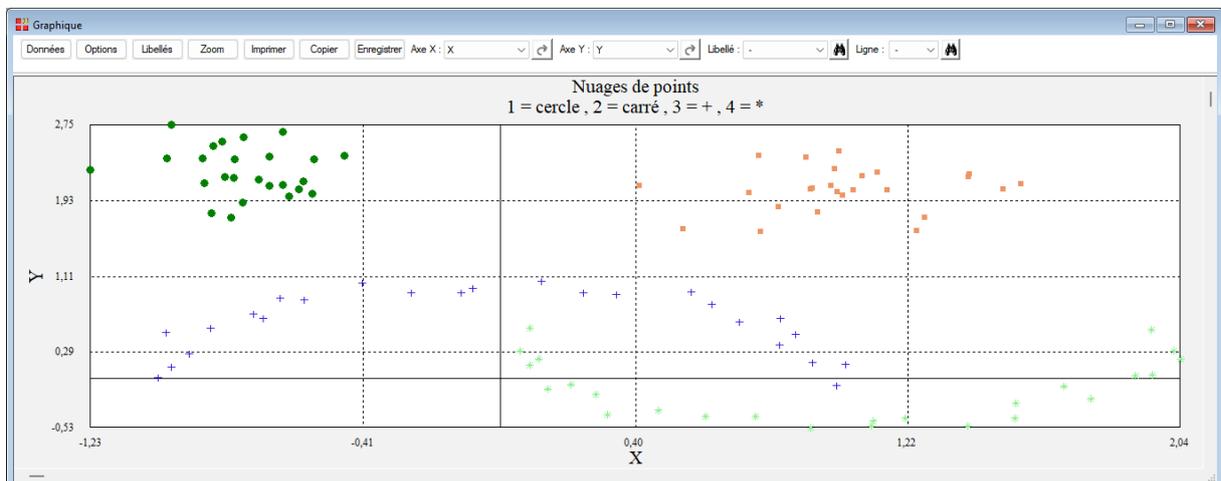
Ce graphique complète le précédent en lui ajoutant les enveloppes convexes des nuages de points des différentes classes formées.



Exemple 2 : Fichier DBSCAN2

Nous utiliserons le fichier DBSCAN2 pour illustrer ce deuxième exemple.

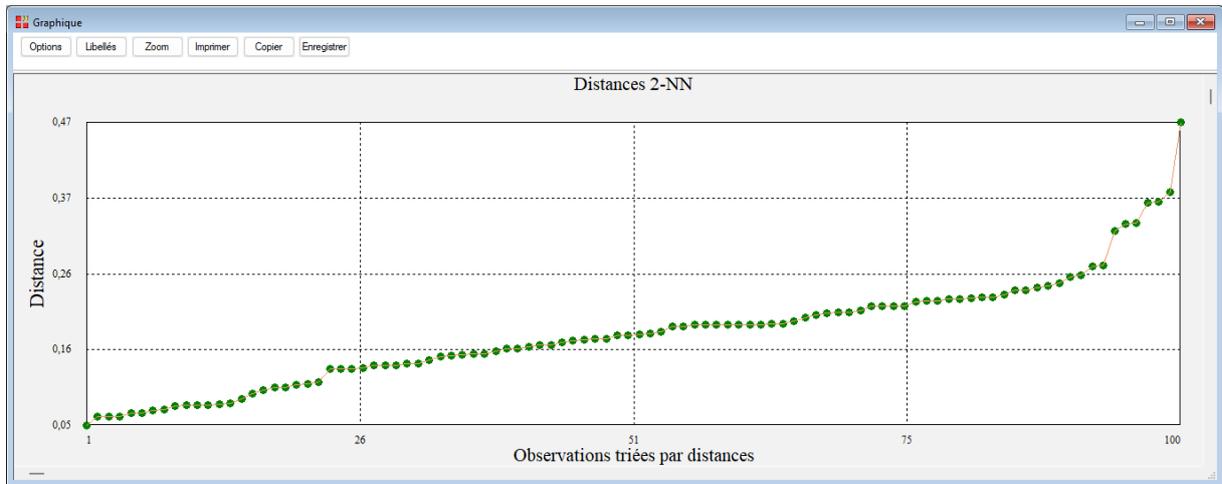
Ce fichier contient 100 observations mesurées sur 2 variables et est constitué de 4 ensembles de données.



Cliquons sur l'icône DBSCAN dans le ruban Décrire.

Sélectionnons les variables 'X' et 'Y' comme colonnes de données, décochons 'Standardisation des données' et conservons les valeurs par défaut de l'épsilon de voisinage et du nombre minimum de points. Incluons les points frontières.

Visualisons le graphique des distances K-NN. Celui-ci indique qu'une valeur d'épsilon d'environ 0,3 semble adéquate.

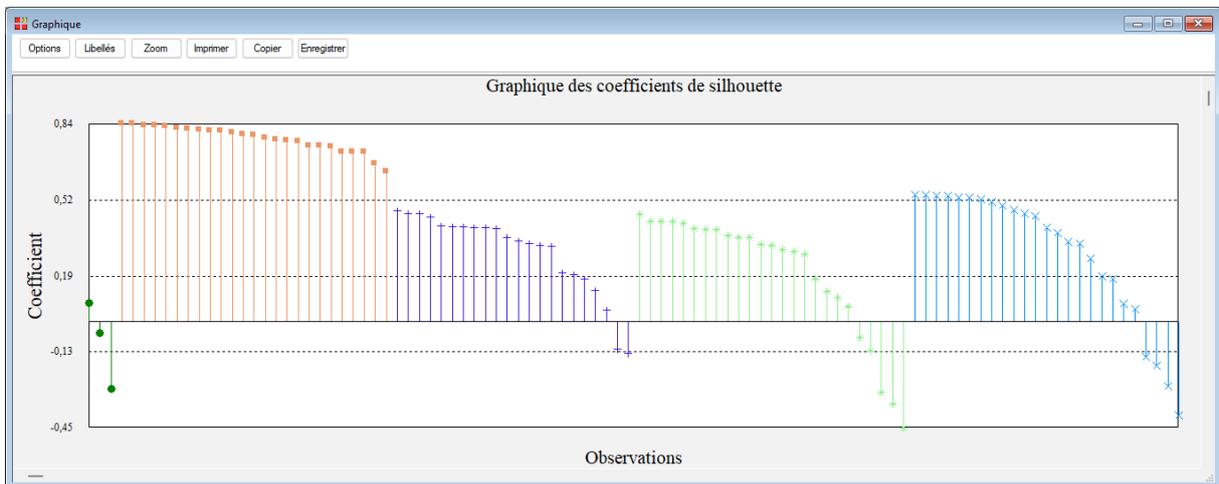
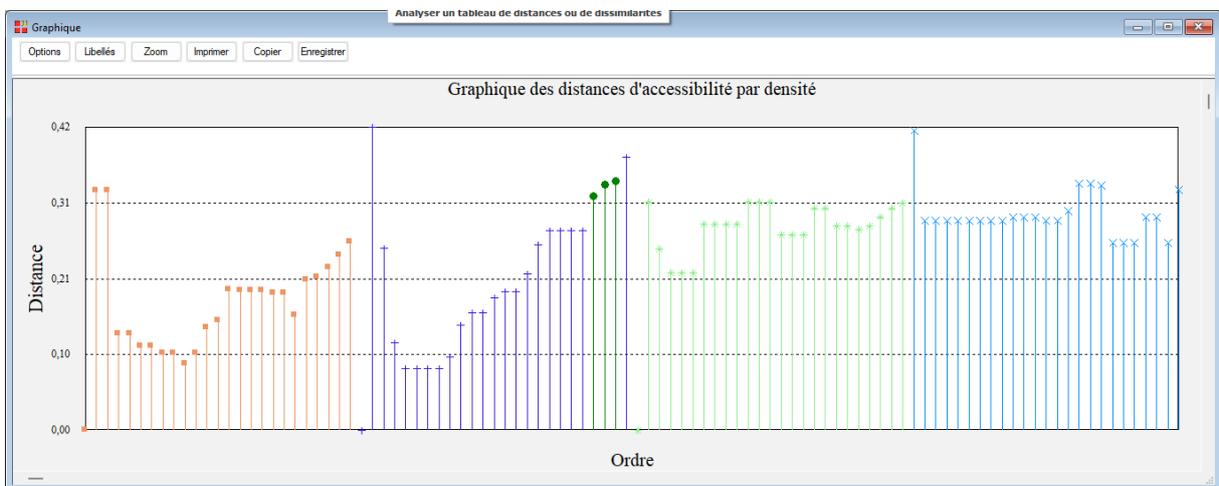
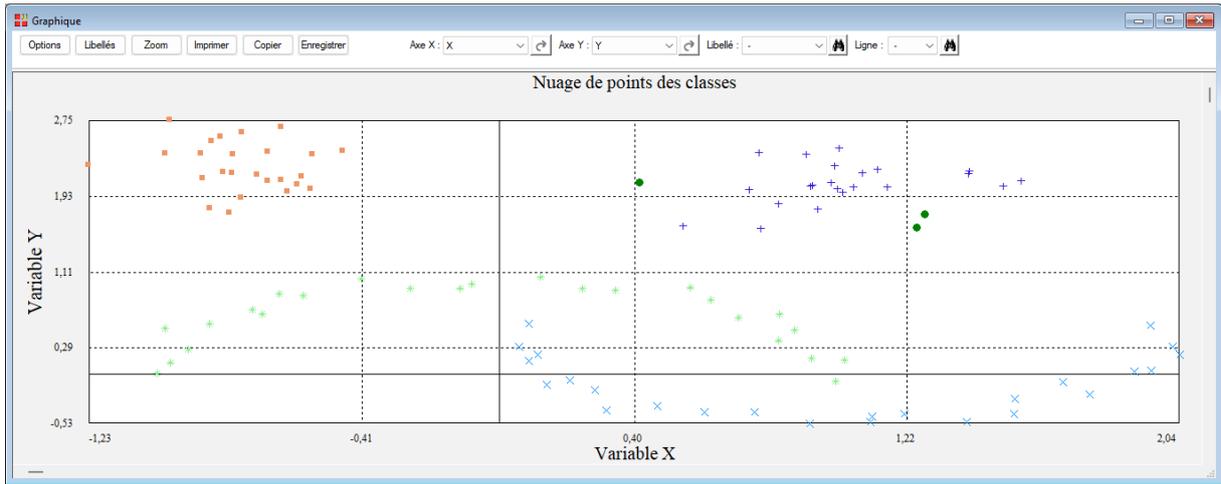


Cliquons sur l'icône dans la barre d'outils pour rappeler la boîte de dialogue d'entrée des données, entrons cette valeur pour l'épsilon de voisinage et exécutons à nouveau l'analyse.

La synthèse de la classification indique que 4 classes sont formées et que 3 observations sont suspectes.

Classe affectée	Effectif	Coefficient de silhouette
0	3	-0,08843
1	25	0,77842
2	22	0,29143
3	25	0,20435
4	25	0,29118

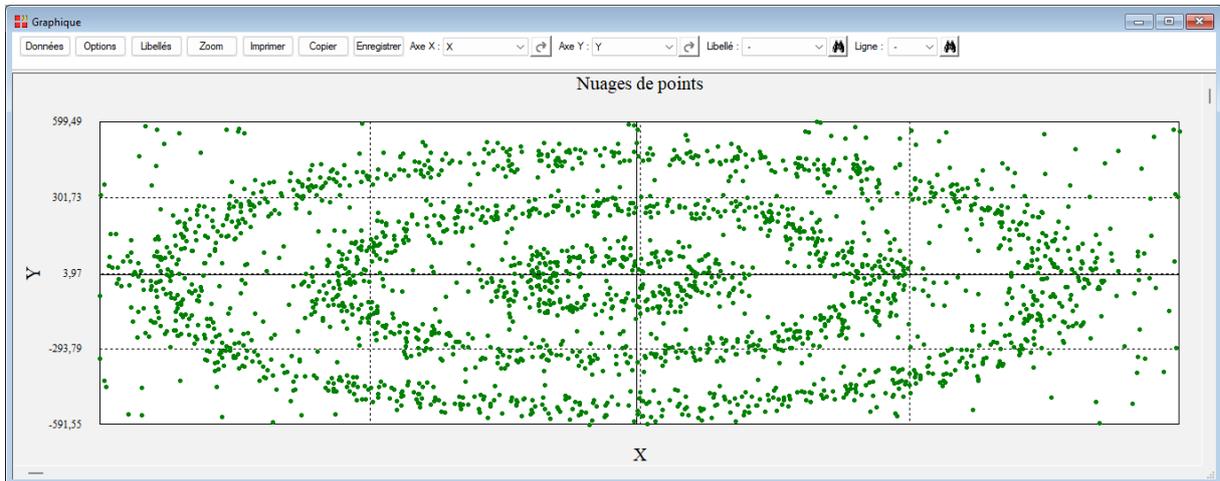
Visualisons le nuage des points et les graphiques des distances d'accessibilité par densité et des coefficients de silhouette.



Exemple 3 : Fichier DBSCAN3

Nous utiliserons le fichier DBSCAN3 pour illustrer ce troisième exemple.

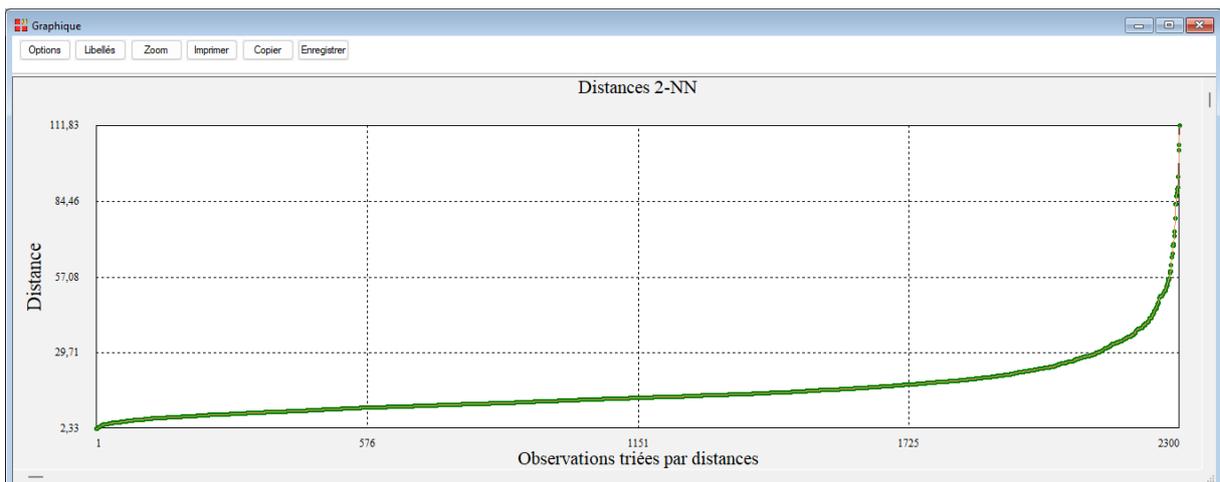
Ce fichier contient 2300 observations mesurées sur 2 variables et est constitué principalement de 3 grands ensembles de données.



Cliquons sur l'icône DBSCAN dans le ruban Décrire.

Sélectionnons les variables 'X' et 'Y' comme colonnes de données, décochons 'Standardisation des données' et conservons les valeurs par défaut de l'épsilon de voisinage et du nombre minimum de points. Inclons les points frontières.

Visualisons le graphique des distances K-NN. Celui-ci indique qu'une valeur d'épsilon d'environ 32 semble adéquate.



Cliquons sur l'icône dans la barre d'outils pour rappeler la boîte de dialogue d'entrée des données, entrons cette valeur pour l'épsilon de voisinage et exécutons à nouveau l'analyse.

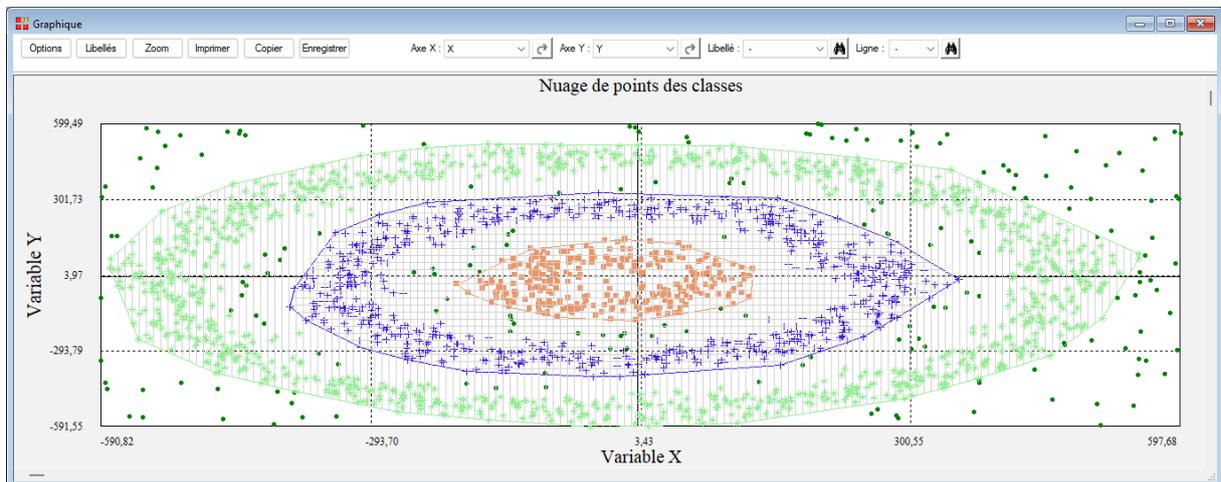
La synthèse de la classification indique que de nombreuses petites classes sont formées. Augmentons le nombre minimum de points de 3 à 5. Nos trois grands ensembles de points sont alors bien détectés

Rapports et Graphiques

Rapport DBSCAN
 Classification des observations
 Synthèse de la classification

	1	2	3	4	5	6	7	8
1								
2	SYNTHESE DE LA CLASSIFICATION DES OBSERVATIONS							
3	Nombre de classes formées : 3							
4	Il y a 179 observations non affectées.							
5								
6								
7		Classe affectée	Effectif	Coefficient de silhouette				
8	0		179	-0,24842				
9	1		308	0,54889				
10	2		736	-0,19650				
11	3		1077	-0,20904				
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								

Rapport Explorateur /



Exemple 4 : Fichier DBSCAN4

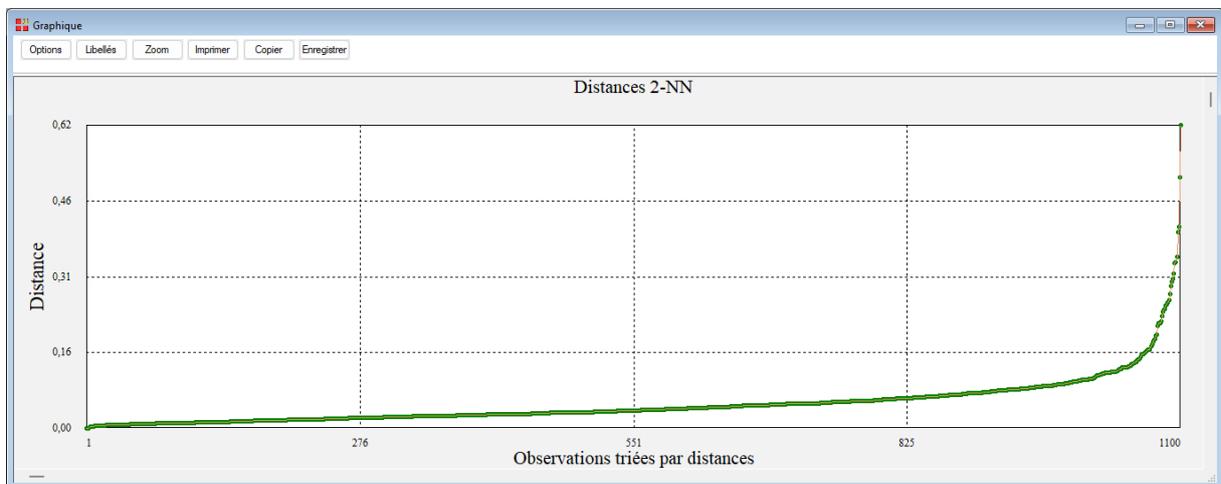
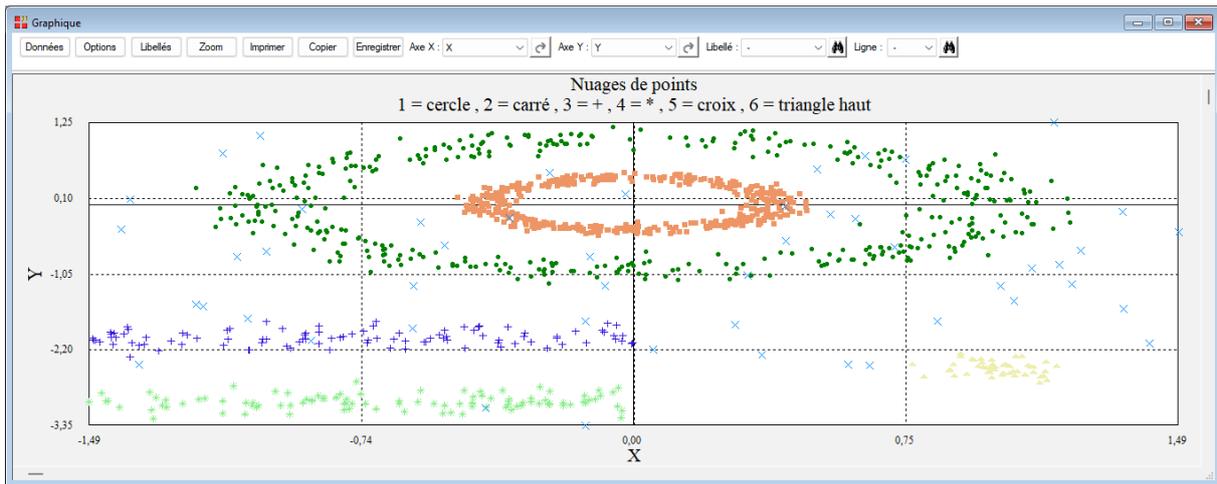
Nous utiliserons le fichier DBSCAN4 pour illustrer ce quatrième exemple.

Ce fichier contient 1100 observations mesurées sur 2 variables et est constitué de 6 ensembles de données.

Cliquons sur l'icône DBSCAN dans le ruban Décrire.

Sélectionnons les variables 'X' et 'Y' comme colonnes de données, décochons 'Standardisation des données' et conservons les valeurs par défaut de l'épsilon de voisinage et du nombre minimum de points. Incluons les points frontières.

Visualisons le graphique des distances K-NN. Celui-ci indique qu'une valeur d'épsilon d'environ 0,15 semble adéquate.



Cliquons sur l'icône  dans la barre d'outils pour rappeler la boîte de dialogue d'entrée des données, entrons cette valeur pour l'épsilon de voisinage, exécutons à nouveau l'analyse et affichons la synthèse de la classification.

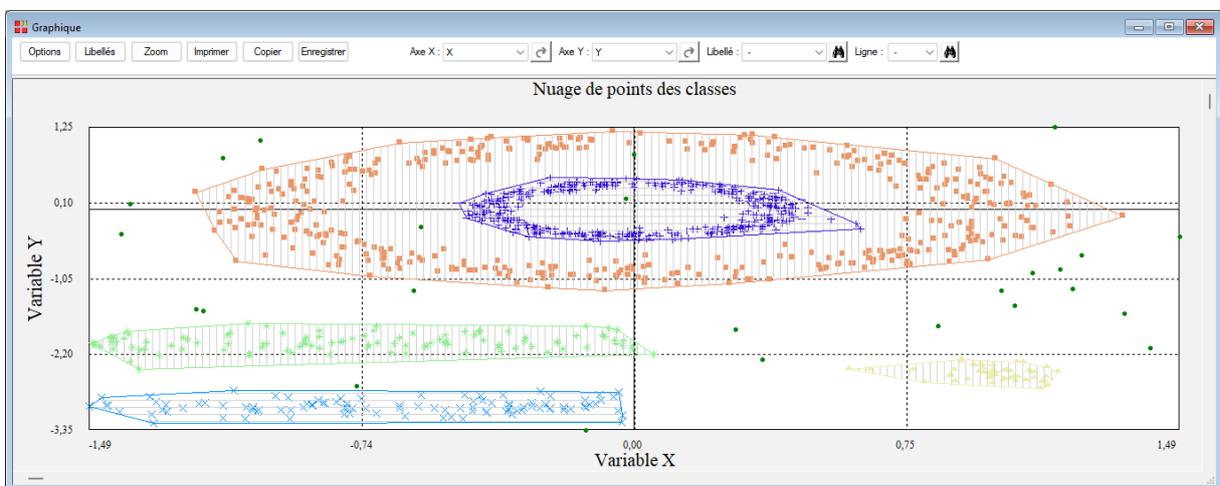
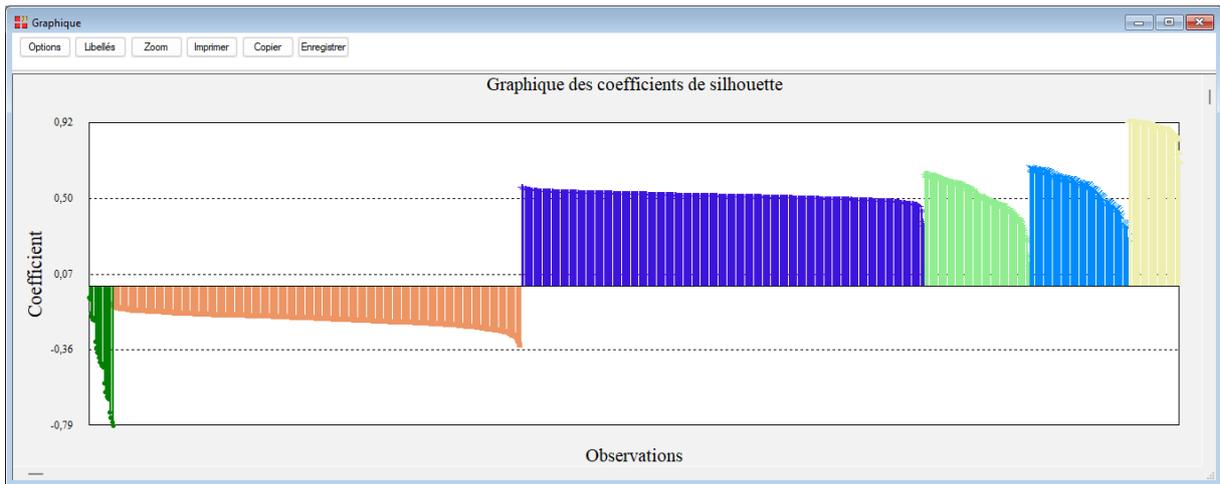
Rapports et Graphiques

Rapport DBSCAN

- Classification des observations
- Synthèse de la classification

	1	2	3	4	5	6	7	8
1								
2	SYNTHÈSE DE LA CLASSIFICATION DES OBSERVATIONS							
3	Nombre de classes formées : 5							
4	Il y a 25 observations non affectées.							
5								
6								
7	Classe affectée	Effectif	Coefficient de silhouette					
8	0	25	-0,44491					
9	1	411	-0,19228					
10	2	406	0,50447					
11	3	106	0,50352					
12	4	100	0,55203					
13	5	52	0,88846					
14								
15								
16								
17								
18								
19								
20								
21								
22								

Rapport Explorateur /



Exemple 5 : Fichier IRIS

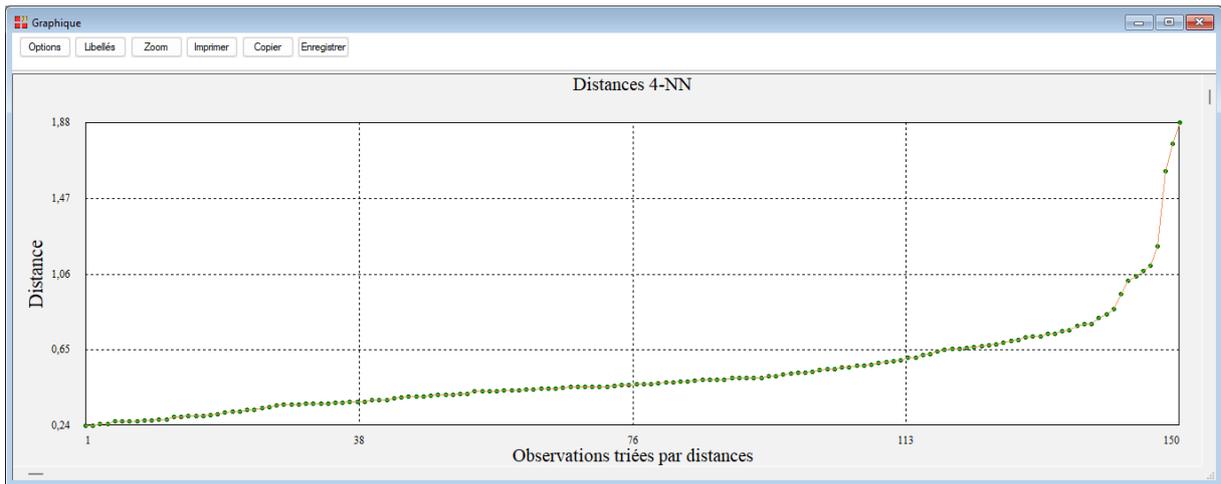
Nous utiliserons le fichier IRIS pour illustrer ce cinquième exemple.

Ce fichier contient pour 150 iris de 3 espèces (Setosa, Versicolor, Virginica) différentes les mesures des quatre caractéristiques suivantes exprimées en millimètres : longueur du sépale, largeur du sépale, longueur du pétale et largeur du pétale

Sélectionnons les quatre variables 'lonsepal', 'larsepal', 'lonpetal' et 'larpetal' comme colonnes de données, cochons 'Standardisation des données' et conservons les valeurs par défaut de l'épsilon de voisinage et du nombre minimum de points. Incluons les points frontières.

Visualisons le graphique des distances K-NN.

Celui-ci indique qu'une valeur d'épsilon d'environ 0,85 semble adéquate.



Visualisons le rapport et les graphiques.

Rapports et Graphiques

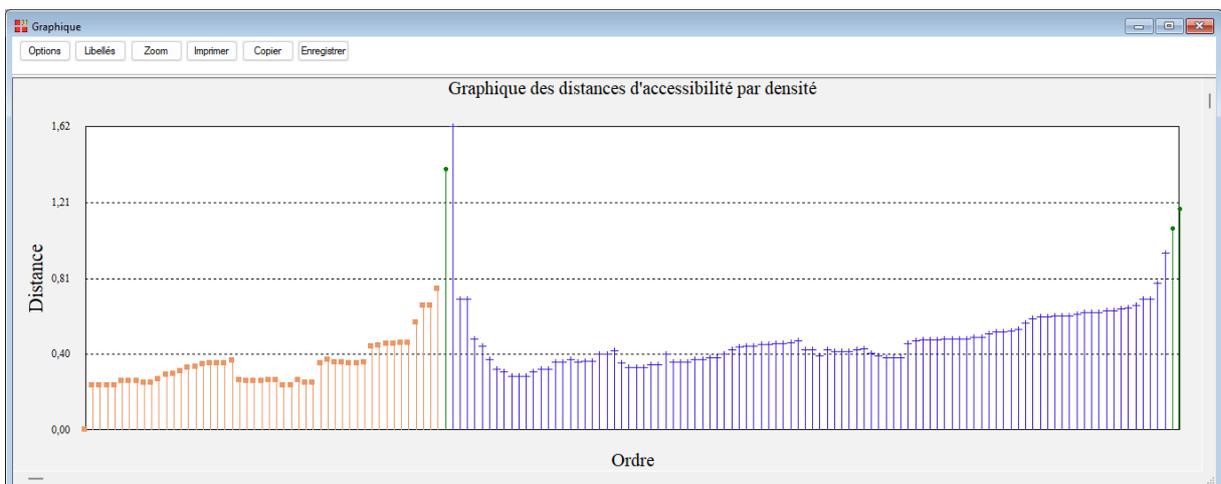
Rapport DBSCAN

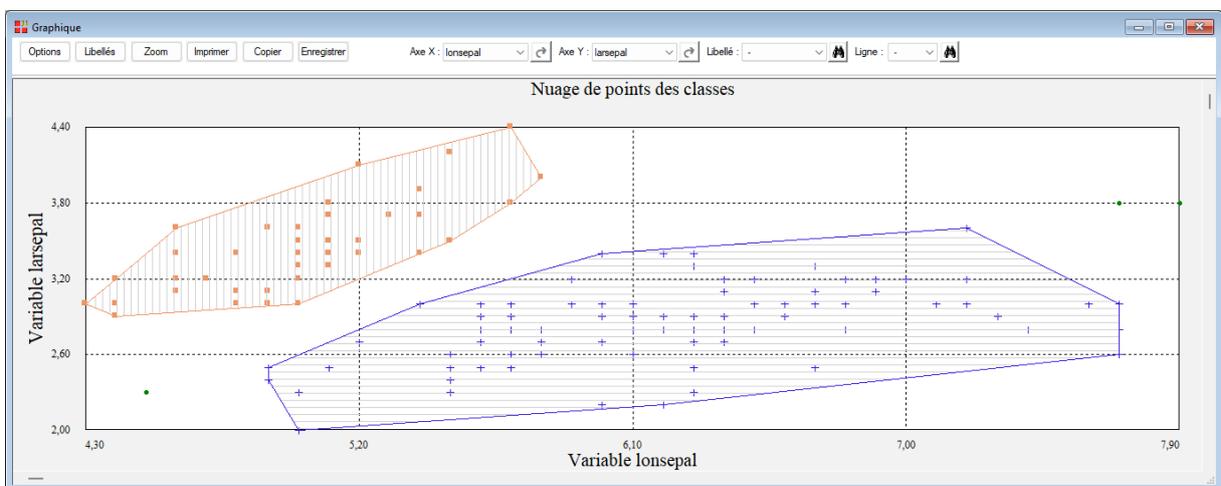
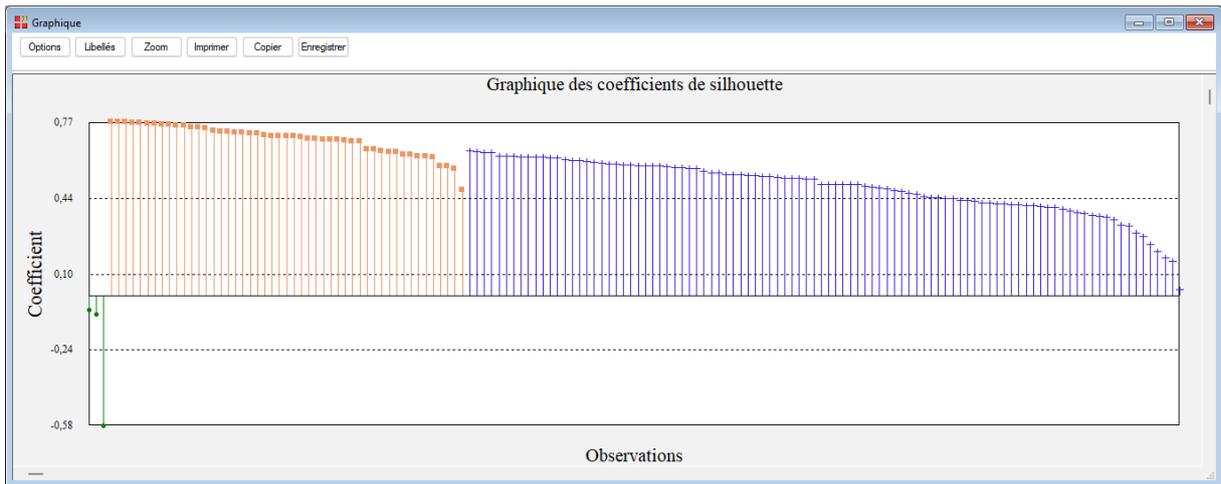
- Classification des observations
- Synthèse de la classification

	1	2	3	4	5	6	7	8
1								
2	SYNTHÈSE DE LA CLASSIFICATION DES OBSERVATIONS							
3	Nombre de classes formées : 2							
4	Il y a 3 observations non affectées.							
5								
6								
7	Classe affectée							
8	0		Effectif	3		Coefficient de silhouette	-0,23997	
9	1		49			0,69546		
10	2		98			0,48353		
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								

Rapport Explorateur

Seules deux des trois classes (espèces) d'iris sont détectées. Les classes Versicolor et Virginica sont confondues.





Les variables internes créées par la procédure

Voici la liste des variables internes créées par la procédure.

<i>Variable</i>	<i>Contenu</i>
libobs	Libellés des observations
distknn	Distances K-NN
distacc	Distances d'accessibilité par densité
classes	Affectations aux classes
classeproche	Classes les plus proches
silhouette	Coefficients de silhouette

Références

Documentation du package R 'dbscan' (2023)

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

Hahsler M, Piekenbrock M, Doran D (2019). "dbscan: Fast Density-Based Clustering with R." - Journal of Statistical Software – 91 (1), 1-30.

Ankerst M, Breunig MM, Kriegel HP, Sander J (1999). "OPTICS: ordering points to identify the clustering structure." In ACM Sigmod Record, volume 28, pp. 49–60. ACM

Reconnaissance des formes et méthodes neuronales - Classification automatique par densité, Nicolas Audebert (2022), Conservatoire National des Arts et Métiers, Paris, France