UFR de Mathématiques et Informatique

Master 2 MLDS 2017 – 2018

RAPPORT DE PROJET VISUALISATION

Professeur : **Nicoleta Rogovschi**

Binôme : BA Kalidou - 21613863

NGUYEN Ngoc Tu - 21710268

1. Introduction

La **visualisation des données** est un ensemble de méthodes de représentation graphique, en deux ou trois dimensions, utilisant ou non de la couleur et des trames. Les moyens informatiques permettent de représenter des ensembles complexes de données, de manière plus simple, didactique et pédagogique.

Méthodes de visualisation

1.1. Analyse en Composantes Principales

L'**analyse en composantes principales** (**ACP** ou **PCA** en anglais pour *principal component analysis*) est une méthode de la famille de l'analyse des données et plus généralement de la statistique multivariée, qui consiste à transformer des variables liées entre elles (dites « corrélées » en statistique) en nouvelles variables décorrelées les unes des autres. Ces nouvelles variables sont nommées « composantes principales », ou axes principaux. Elle permet au praticien de réduire le nombre de variables et de rendre l'information moins redondante. L'ACP est majoritairement utilisé pour décrire et visualiser des données.

1.2. Analyse Discriminante Linéaire

C'est une méthode pour l'analyse de données de grande dimension dans le cas de l'apprentissage supervisé (les classes (les labels) sont disponibles dans l'ensemble de données). Elle trouve un espace à faible dimension optimal telle que, lorsque les points sont projetés, les données de différentes classes sont bien séparées. Elle est utile pour l'extraction des caractéristiques pour faciliter la classification supervisée.

1.3. Multi Dimensional Scaling

C'est un ensemble de techniques de réduction de la dimension qui projettent les distances entre les observations d'un espace à grandes dimensions dans un espace de petites dimensions. Il trouve une configuration des points dans un espace de faibles dimensions dont les distances inter-points correspondent aux dissimilarités dans les grandes dimensions.

1.4. Isomap

Isomap est un représentant des méthodes de cartographie isométrique, et étend l'échelle multidimensionnelle métrique (MDS) en intégrant les distances géodésiques imposées par un graphe pondéré. Pour être précis, la mise à l'échelle classique de la MDS métrique effectue un encastrement de faible dimension basé sur la distance par paires entre les points de données, qui est généralement mesurée en utilisant la distance euclidienne linéaire. Isomap se distingue par son utilisation de la distance géodésique induite par un graphe de voisinage intégré dans la mise à l'échelle classique. Ceci est fait pour incorporer la structure de collecteur dans l'intégration résultante. Isomap définit la distance géodésique comme étant la somme des poids de bord le long du chemin le plus court entre deux nœuds (calculé à l'aide de l'algorithme de Dijkstra, par exemple). Les n premiers vecteurs propres de la matrice de distance géodésique représentent les coordonnées dans le nouvel espace euclidien n-dimensionnel.

1.5. Locally Linear Embedding

L'inclusion locale linéaire (LLE) a été présentée à peu près au même moment que l'Isomap. Il a plusieurs avantages sur Isomap, y compris une optimisation plus rapide lorsqu'il est mis en œuvre pour tirer parti des algorithmes de matrice clairsemée, et de meilleurs résultats avec de nombreux problèmes. LLE commence

également par trouver un ensemble des voisins les plus proches de chaque point. Il calcule ensuite un ensemble de poids pour chaque point qui décrit le mieux le point comme une combinaison linéaire de ses voisins. Enfin, il utilise une technique d'optimisation basée sur le vecteur propre pour trouver l'encastrement de points de faible dimension, de sorte que chaque point est toujours décrit avec la même combinaison linéaire de ses voisins. LLE a tendance à mal gérer les densités d'échantillons non uniformes car il n'y a pas d'unité fixe pour empêcher les poids de dériver, car les différentes régions diffèrent en termes de densité d'échantillon. LLE n'a pas de modèle interne.

1.6. Self-Organizing Map

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.

2. Prétraitement des données

2.1. gordon

La distinction pathologique entre le mésothéliome pleural malin (MPM) et l'adénocarcinome (AD) du poumon peut être lourde en utilisant des méthodes établies. Les auteurs proposent qu'une technique simple, basée sur les niveaux d'expression d'un petit nombre de gènes, peut être utile dans le diagnostic précoce et précis du MPM et du cancer du poumon. Cette méthode est conçue pour distinguer avec précision les tissus génétiquement disparates en utilisant des rapports d'expression génique et des seuils choisis rationnellement.
Dans ce tableau, les individus sont classés en deux groupes. Le premier groupe a un effectif de 31 individus (MPM) et le deuxième a 150 individus (AD). C'est un jeu de données décrites par 1626 variables quantitatives. La première étape consiste à centrer et réduire toutes les variables car ces derniers ont des variabilités différentes.

2. pomeroy

Les tumeurs embryonnaires du système nerveux central (SNC) représentent un groupe hétérogène de tumeurs dont on sait peu de choses sur le plan biologique et dont le diagnostic, sur la seule base de l'aspect morphologique, est controversé. Les médulloblastomes (MD), par exemple, sont la tumeur cérébrale maligne la plus courante de l'enfance, mais leur pathogénie est inconnue, leur relation avec d'autres tumeurs embryonnaires du SNC est discutée et la réponse des patients au traitement est difficile à prévoir. Les auteurs ont abordé ces problèmes en mettant au point un système de classification basé sur les données d'expression génétique des micros réseaux d'ADN dérivées de 99 échantillons de patients. Ils démontrent que les médulloblastomes sont moléculairement distincts des autres tumeurs cérébrales, y compris les tumeurs neuroectodermiques primitives (PNET), les tumeurs atypiques tératoïdes / rhabdoïdes (Rhab) et les gliomes

malins (Mglio). Au sein de la classe des médulloblastomes (MD), ils ont également étudié l'hétérogénéité des classiques (C) desmoplasiques (D). Dans l'analyse, les tissus normaux ont également été considérés (Ncer). Dans ce tableau, les individus sont classés en deux groupes. Le premier groupe a un effectif de 10 MD, 10 Mglio, 10 Rhab, 4 Ncer 8 PNET. C'est un jeu de donnée décrit par 1379 variables quantitatives. La première étape consiste à centrer et réduire toutes les variables car ces derniers ont des variabilités différentes.

3. Visualisation des données : gordon et pomeroy

3.1. Gordon

3.1.1. ACP

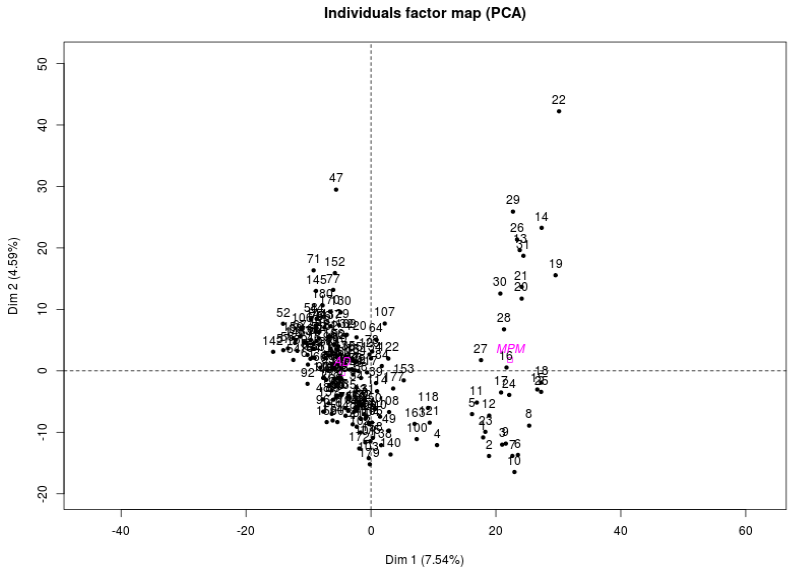

FIGURE 1 : Représentation des individus sur 2 premiers axes factorielles de l'ACP

Avec l'ACP, les données sont centrées autour de l'origine des deux axes factoriels. Le premier axe discrimine au mieux ce jeu de données. Avec ces deux axes factoriels on récupère environ 12% de l'inertie totale.

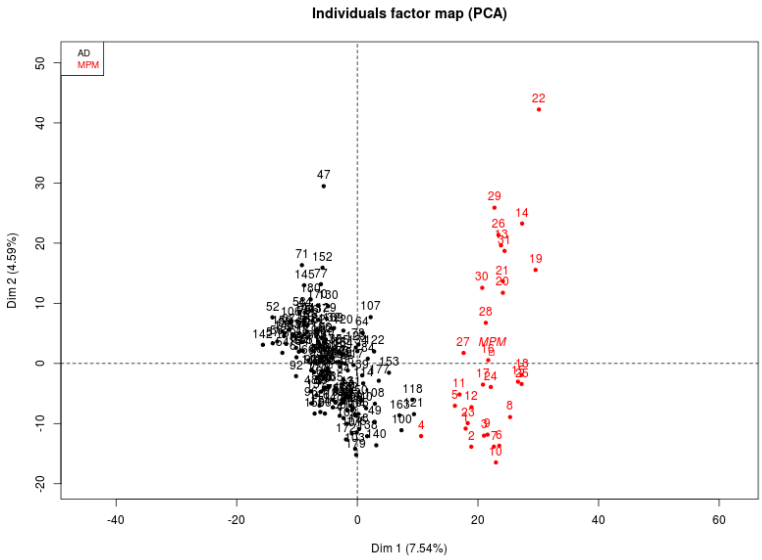

FIGURE 2 : Représentation des individus avec des labels sur 2 premiers axes factoriels de l'ACP

En utilisant des labels des individus, l'ACP nous permet d'identifier l'existence de deux groupes séparés par le premier axe. Etant donné que l'ACP n'est pas une méthode fait pour la classification, on ne s'attend pas à ce que les classes soient bien séparées. Néanmoins avec ce jeu de données, les classes sont bien visibles.

3.1.2. ADL

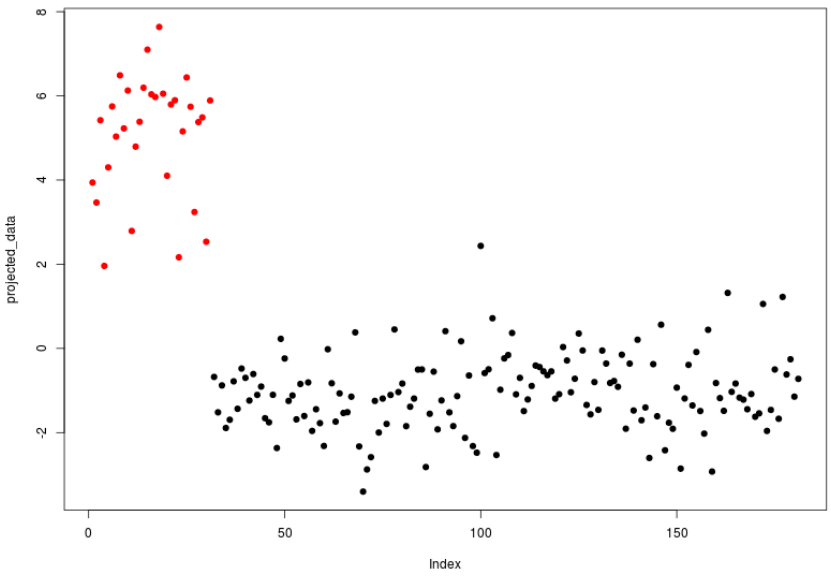

FIGURE 3 : Représentation des individus sur l'axe factorielle de l'ADL

Avec l'ADL, le choix du nombre d'axes discriminants dépend du nombre de classes. Dans ce jeu de données, des individus sont mis dans deux classes, donc un seul axe discriminant suffit pour faire la classification. Lorsqu'on regarde l'axe vertical, les deux groupes d'individus sont bien séparées.

3.1.3. MDS

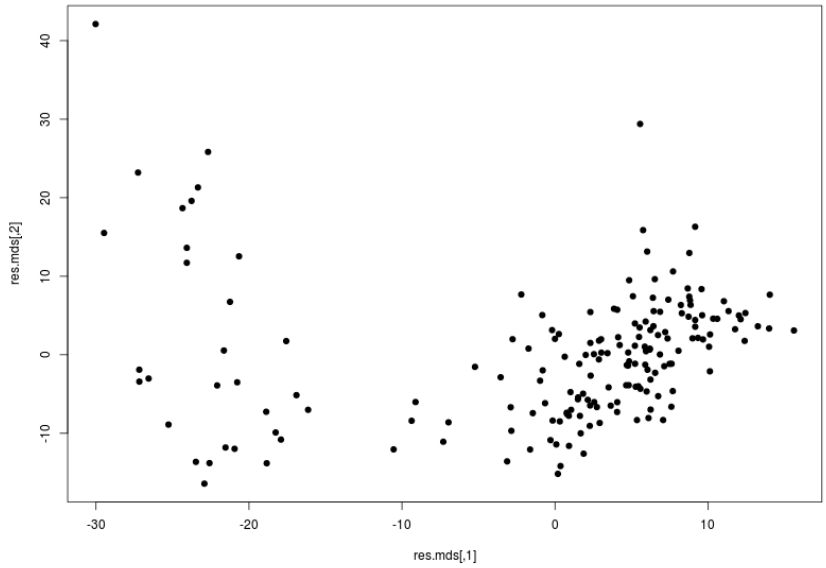

FIGURE 4 : Représentation des individus sur deux axes factorielle du MDS

C'est une méthode de non supervisée, aucune connaissance préalable des classes. Visuellement, la MDS offre une représentation en deux groupes de densité différente. Il est impossible de savoir a` ce niveau si les individus ont été bien classés.

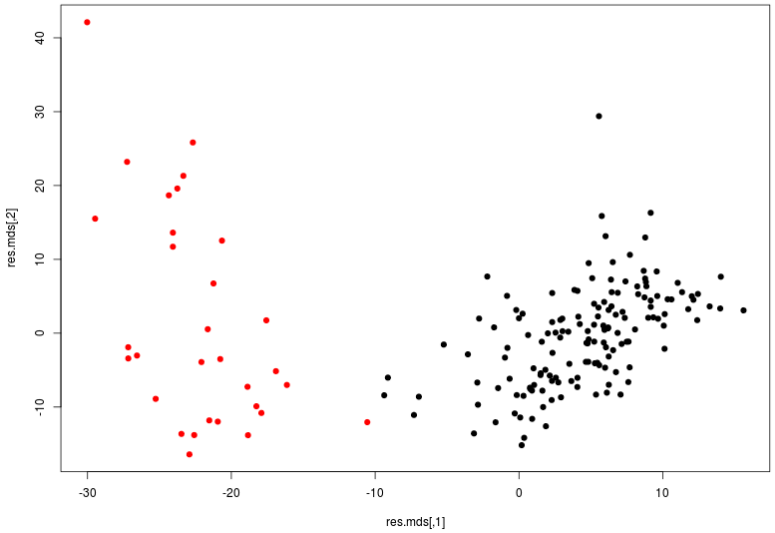

FIGURE 5 : Représentation des individus avec des labels sur deux axes factorielle du MDS

Avec des labels, on observe bien une bonne classification sur le jeu de données gordon. Les classes sont parfaitement séparées suivant l'axe horizontal.

3.1.4. Isomap

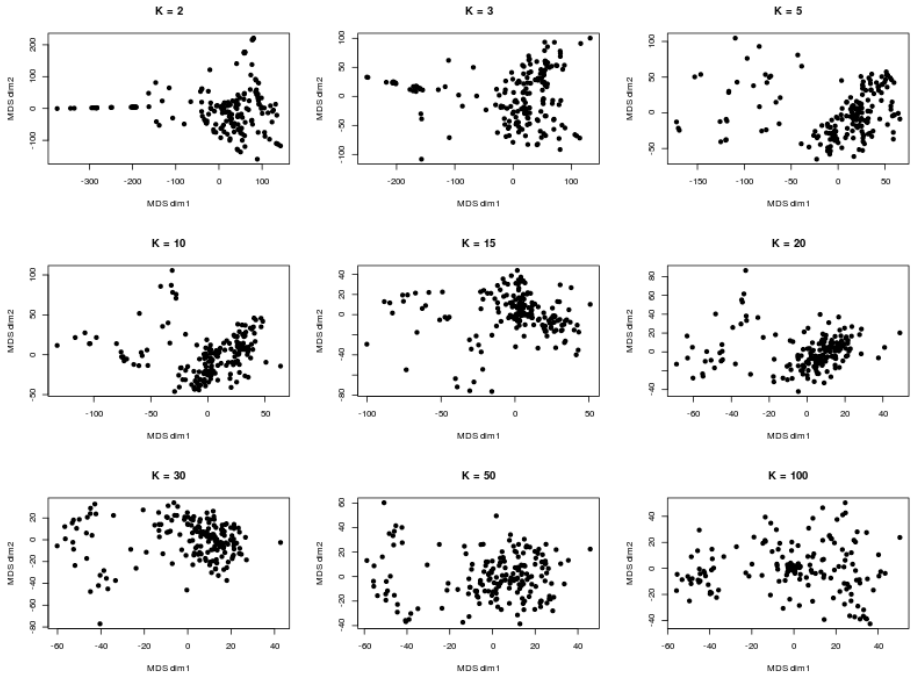

FIGURE 6 : Représentation des individus en fonction de K plus proches voisins sur deux axes factorielle du MDS avec l'algorithme Isomap

En fonction des K, on observe une représentation différente des individus. Par exemple, avec K = 2, ce sera difficile de faire la classification. La meilleure classification est obtenue lorsqu'on choisit les 5 plus proches voisins. Plus on augmente le K, plus les classes sont hétérogènes.

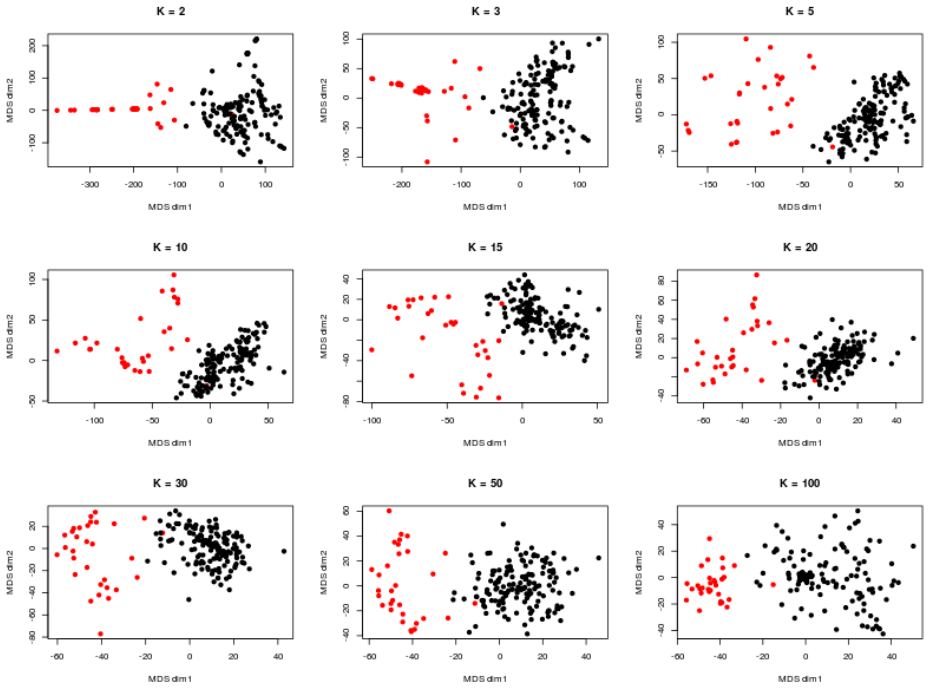

FIGURE 7 : Représentation des individus labélisés en fonction de K plus proches voisins sur deux axes factorielle du MDS avec l'algorithme Isomap

Avec K = 5, tous les individus sont mis dans les bonnes classes excepte' d'un individu mal classé.

3.1.5. LLE

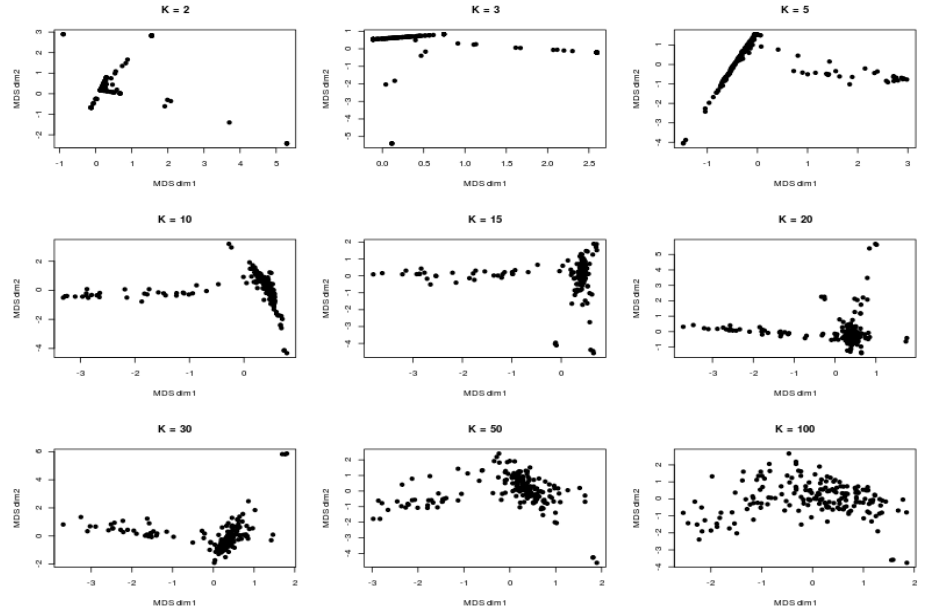

FIGURE 8 : Représentation des individus en fonction de K plus proches voisins sur deux axes factorielle du MDS avec l'algorithme LLE

La forme de la classe obtenue est différente. En fonction des K, on observe une représentation différente des individus. Par exemple, avec K = 100, ce sera difficile de faire la classification. La meilleure classification est obtenue lorsqu'on choisit les 5 ou 10 plus proches voisins. Plus on augmente le K, plus les classes sont hétérogènes.

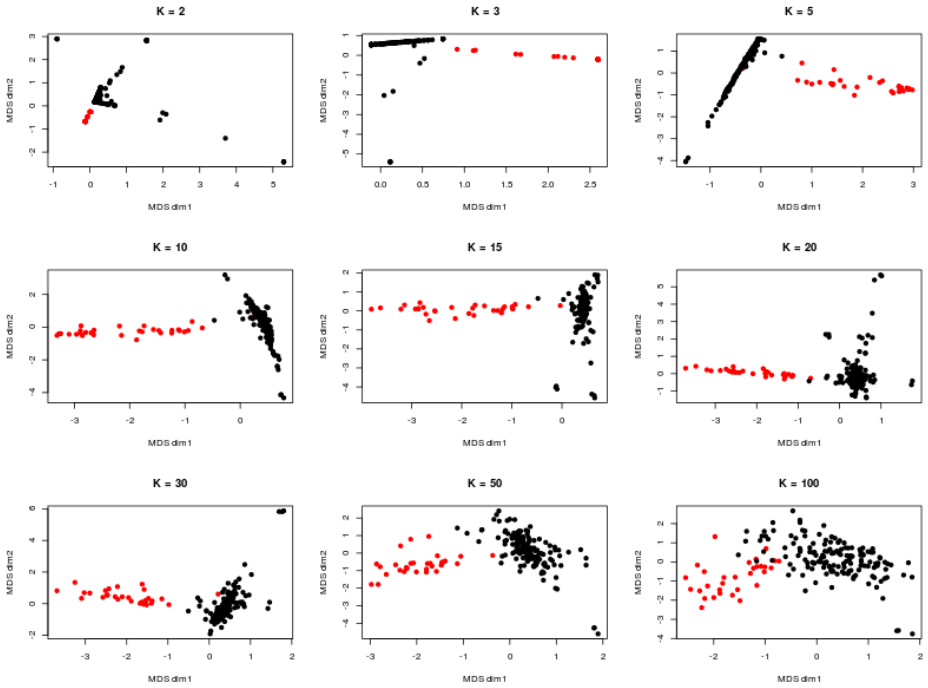

FIGURE 9 : Représentation des individus labélisés en fonction de K plus proches voisins sur deux axes factorielle du MDS avec l'algorithme LLE

Avec K = 5 et K = 10, tous les individus sont mis dans les bonnes classes exceptées d'un individu mal classés.

3.1.6. SOM

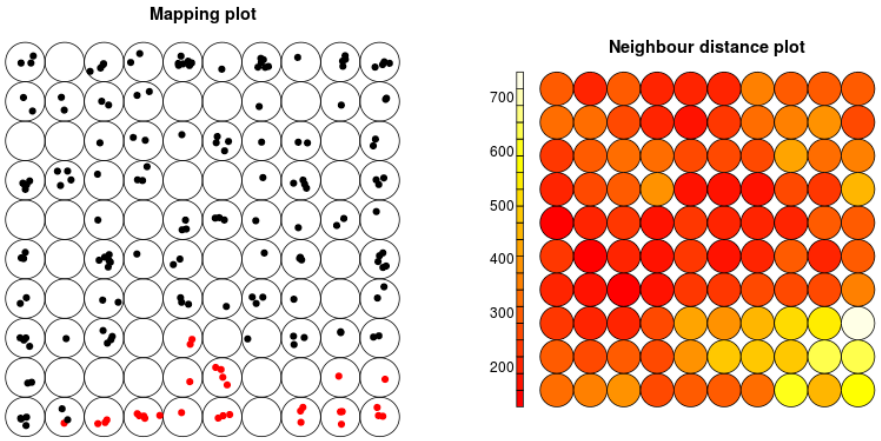

FIGURE 10 : Mapping plot et U-matrice

Avec le Mapping plot on observe les données semblables dans les mêmes neurones ou des neurones proches. On distingue facilement l'existence de deux classes visibles à partir d'U-matrice. Avec le U-matrice on observe une frontière qui sépare ces deux classes

3.2. Pomeroy

3.2.1. ACP

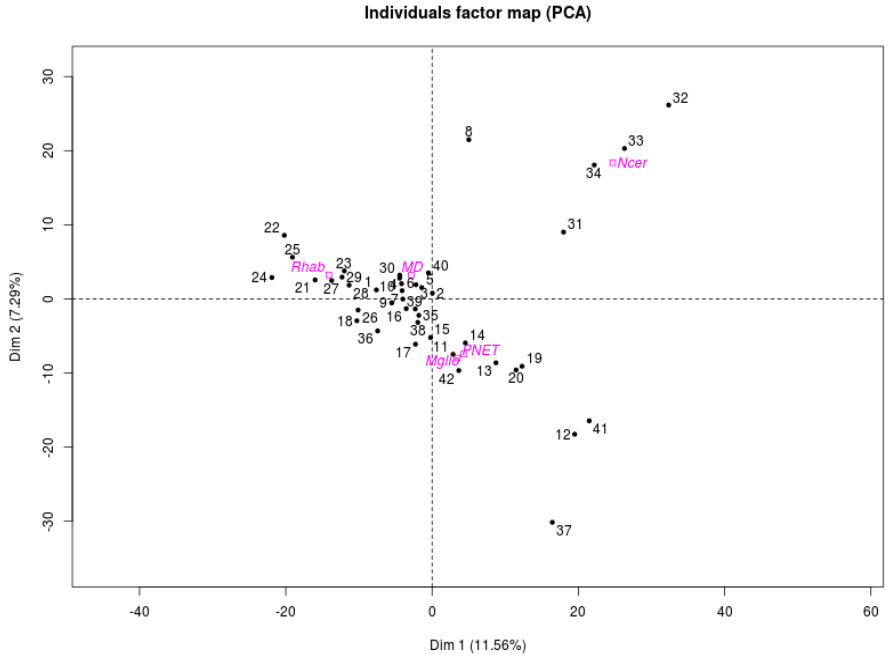

FIGURE 11 : Représentation des individus sur 2 premiers axes factoriels de l'ACP

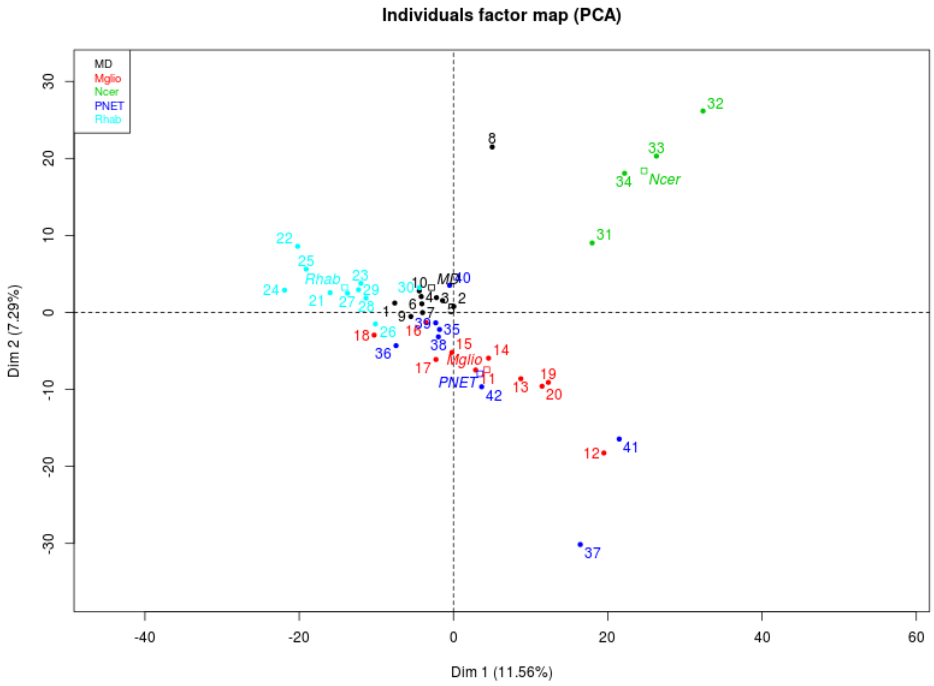

FIGURE 12 : Représentation des individus avec des labels sur 2 premiers axes factoriels de l'ACP

Aucun axe discrimine ce jeu de données ce qui est normal du fait que le rôle de l'ACP est de réduire les dimensions. Avec ces deux axes factoriels on récupère environ 18% de l'inertie totale.

3.2.2. ADL

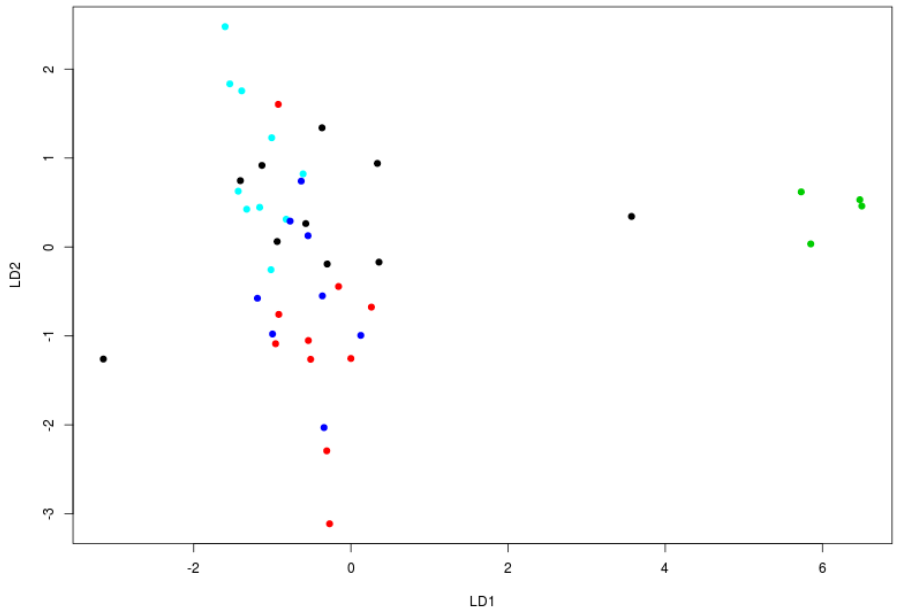

FIGURE 13 : Représentation 2D des individus sur deux axes factorielle de l'ADL

L'ADL à 2 axes identifie bien la classe représentée en vert par rapport aux autres classes qui sont mélangées entre elles. Avec ce jeu de données l'ADL à deux dimensions fait une mauvaise classification.

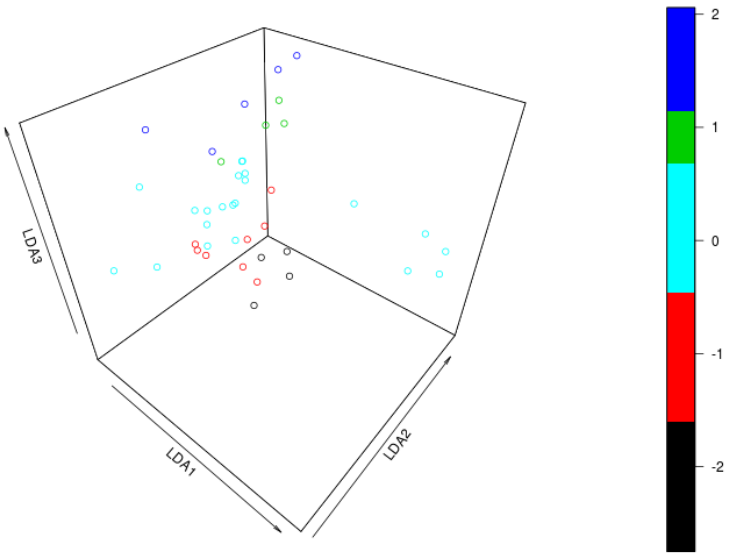

FIGURE 14 : Représentation 3D des individus sur trois axes factorielle de l'ADL

En 3D la segmentation des classes est beaucoup plus visible mais l'interprétation reste fastidieuse.

3.2.3. MDS

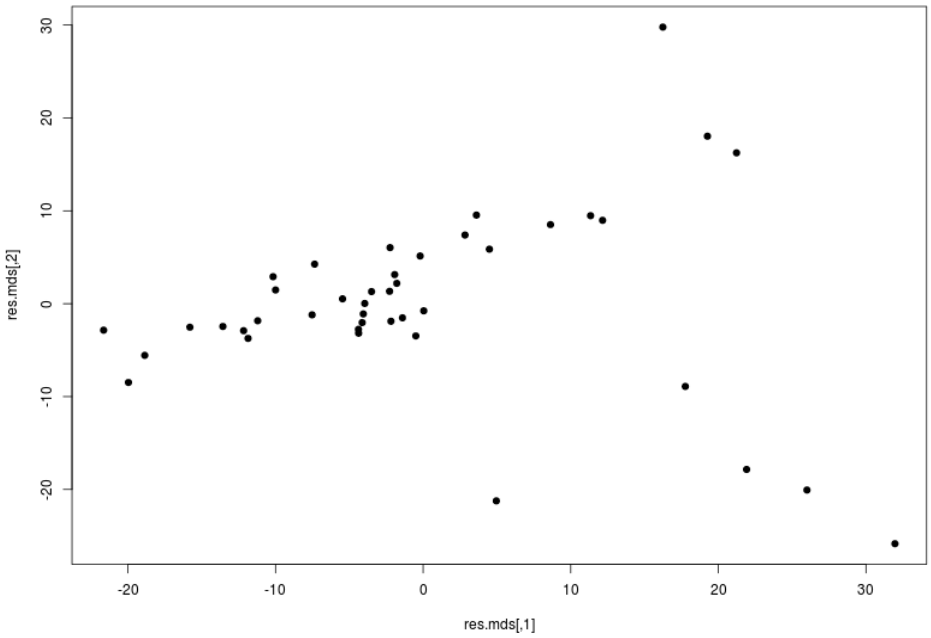

FIGURE 15 : Représentation des individus sur deux axes factorielle du MDS

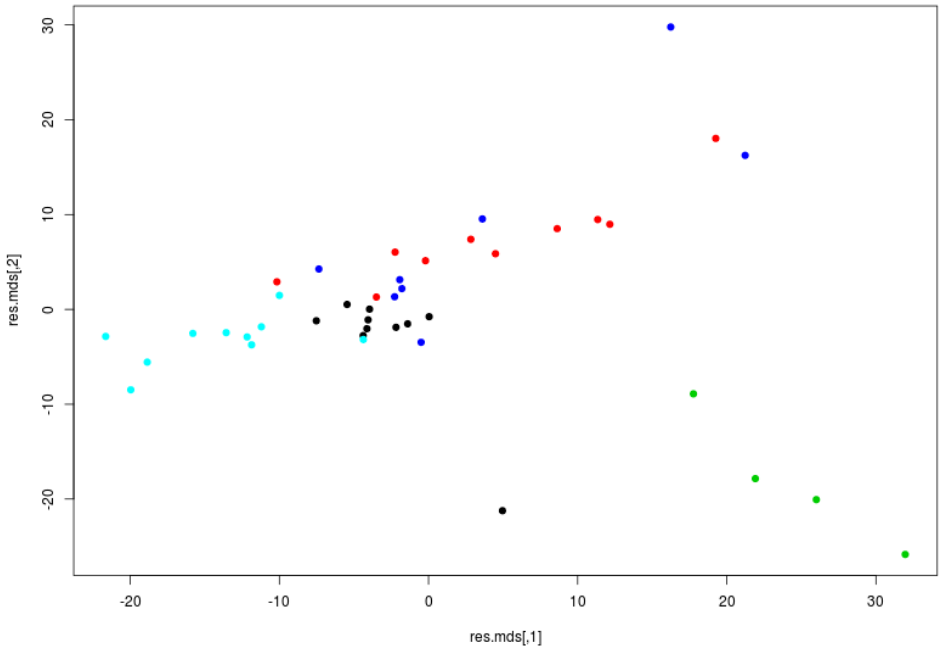

FIGURE 16 : Représentation des individus labélisés sur deux axes factorielle du MDS

Avec cette méthode on parvient à identifier de manières séparées deux classes d'individus est les autres sont mélangées.

3.2.4. Isomap

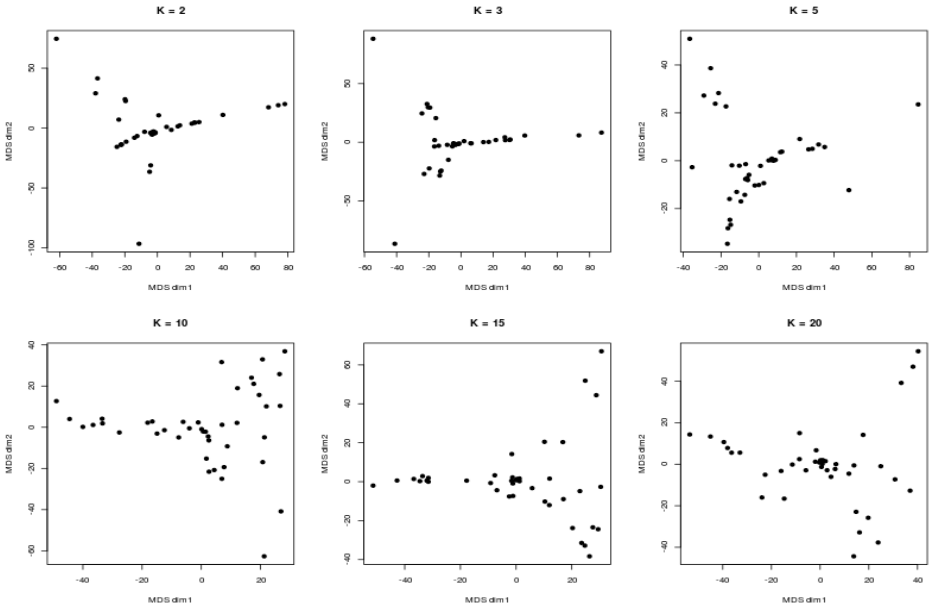

FIGURE 17 : Représentation des individus en fonction de K plus proches voisins sur deux axes factorielle du MDS avec l'algorithme Isomap

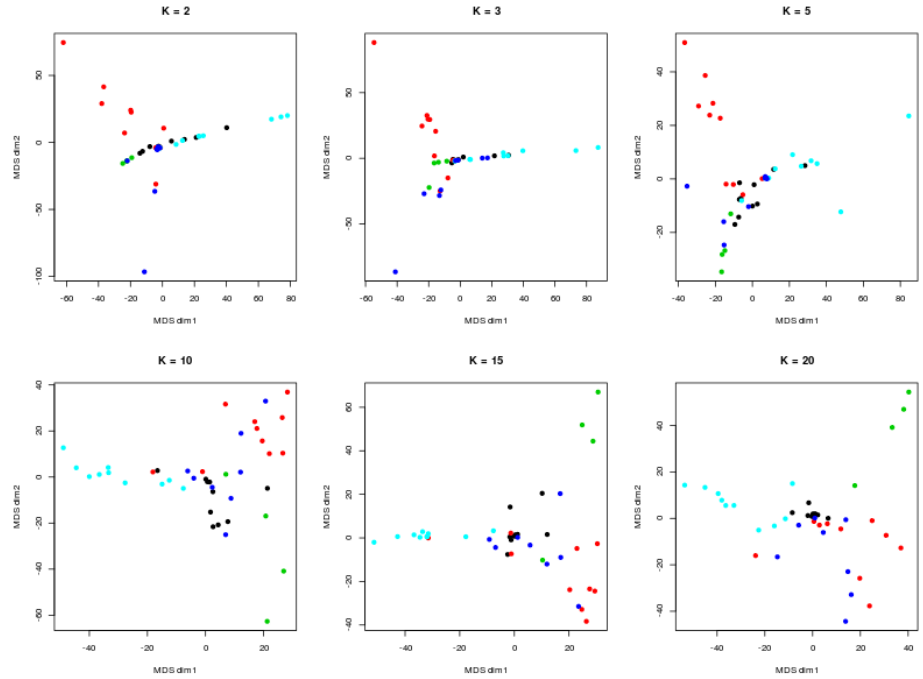

FIGURE 18 : Représentation des individus labélisés en fonction de K plus proches voisins sur deux axes factorielle du MDS avec l'algorithme Isomap

On obtient des visualisations différentes en fonction du nombre de K plus proches voisins la meilleure visualisation est obtenue lorsque K=5. On peut noter que pour n'importe quelle valeur de K toutes les classes ne sont pas indentifiables de manières séparées.

3.2.5. LLE

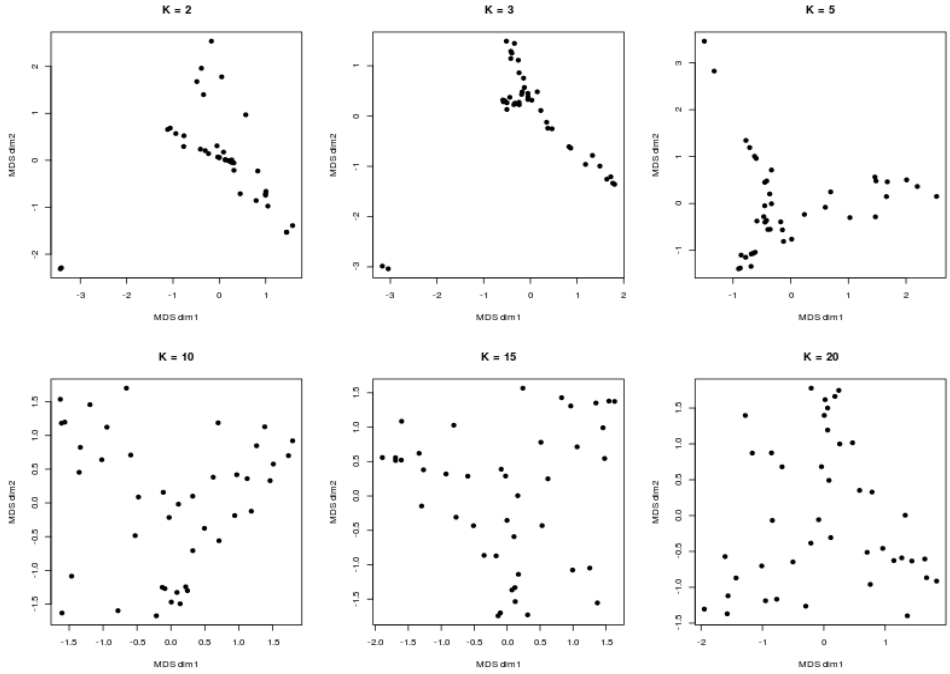

FIGURE 19 : Représentation des individus en fonction de K plus proches voisins sur deux axes factorielle du MDS avec l'algorithme LLE

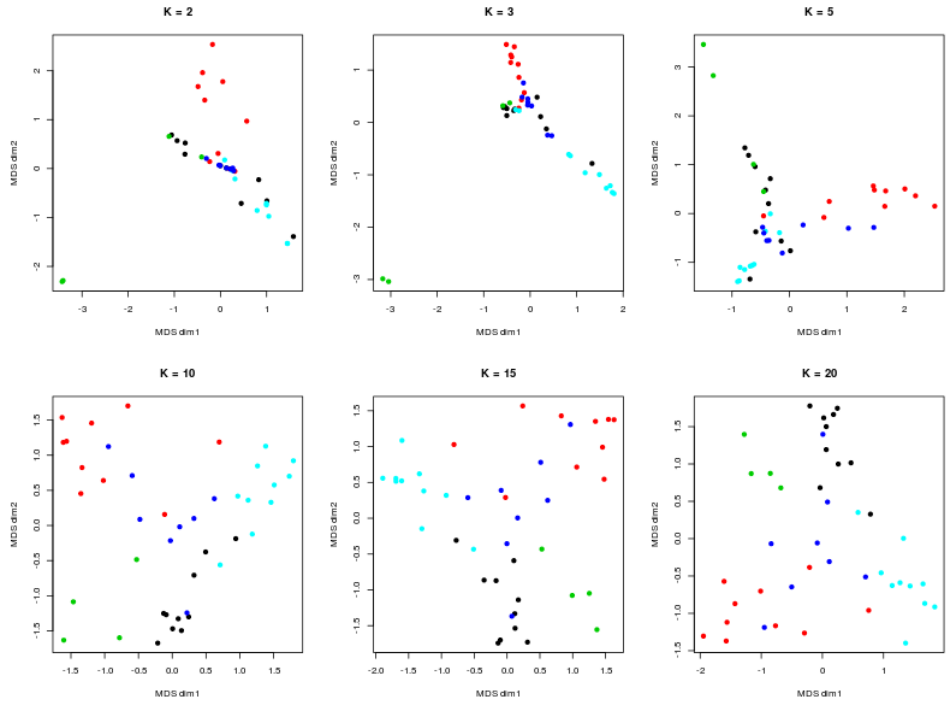

FIGURE 20 : Représentation des individus labélisés en fonction de K plus proches voisins sur deux axes factorielle du MDS avec l'algorithme LLE

Cette méthode ne permet pas de faire la classification avec ce jeu de données. Quel que soit le nombre de plus proches voisins choisit les données sont mélangées.

3.2.6. SOM

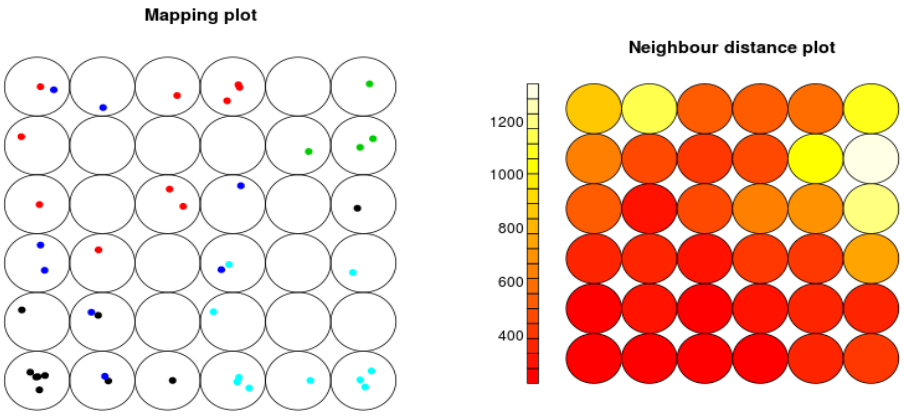

FIGURE 21 : Mapping plot et U-matrice

Avec le Mapping plot on voit que des individus de de classes différentes sont mis dans le même neurone ou des neurones très proches.

4. Conclusion

Plusieurs algorithms d'apprentissage existent et leurs performances dépendent de plusieurs phénomènes à savoir la forme et le nombre de classe, la distribution des classes dans l'espace.

Avec les techniques de visualition on parvient à des répresentations souvent visuellement interprétables.