

UFR de Mathematiques et Informatique

Master 2 MLDS 2017 – 2018

RAPPORT DE PROJET

VISUALISATION

Professeur: Nicoleta Rogovschi

Etudiant: BA Kalidou

NGUYEN Ngoc Tu

I. 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.

Methodes de visualisation

1. ACP

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 décorrélées 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 utilisée pour décrire et visualiser des données.

1. ADL

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. MDS

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 faible dimensions dont les distances inter-points correspondent aux dissimilarités dans les grandes dimensions.

1. 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. LLE

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. SOM

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.

II. Pre-traitement des donnees

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 classees en deux group. Le premier groupe a un effectif de 31 individus (MPM) et le deuxieme a 150 individus (AD). C’est un jeu de donnees decrit par 1626 variables quantitatives. La premier etape consiste a centrer et reduire toutes les variables car ces derniers ont des variabilites differentes.

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 microréseaux 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 classees en deux group. Le premier groupe a un effectif de 10 MD, 10 Mglio, 10 Rhab, 4 Ncer 8 PNET. C’est un jeu de donnees decrit par 1379 variables quantitatives. La premier etape consiste a centrer et reduire toutes les variables car ces derniers ont des variabilites differentes.

III. Visualisation des donnees: gordon et pomeroy

1. Gordon
2. ACP

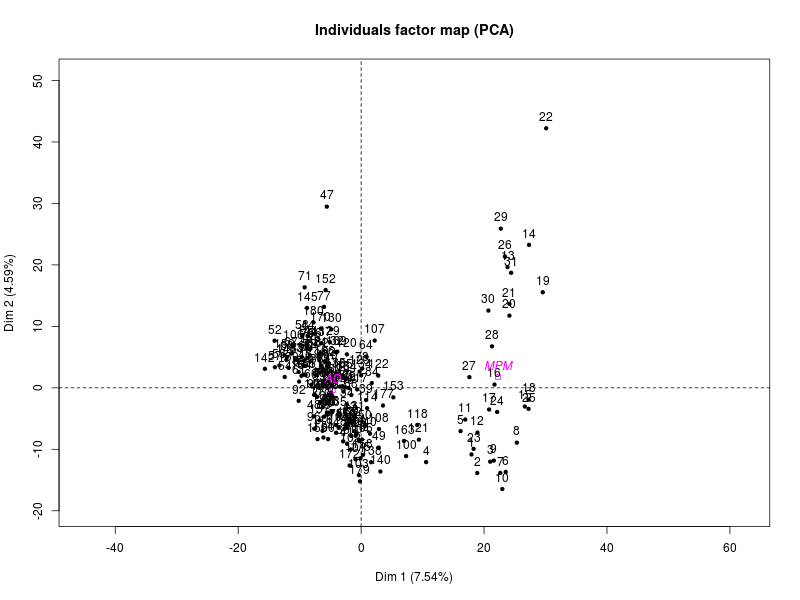


FIGURE 1 : Representation des individus sur 2 premiers axes factorielles de l’ACP

Avec l’ACP, les donnees sont centrees autour de l’origine des deux axes factorielles. Le premier axe qui discrime au mieux ce jeu de donnees. Avec ces deux axes factorielles on recupere environ 12% de l’inertie totale.

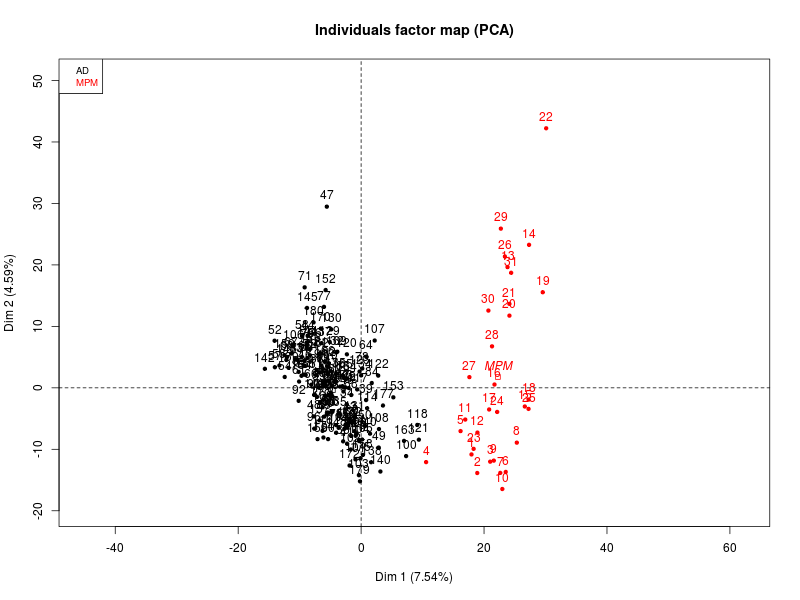


FIGURE 2 : Representation des individus avec des labels sur 2 premiers axes factorielles de l’ACP

En utilisant des labels des individus, l’ACP nous permet d’identifier l’existance de deux groupes separees par le premier axe. Etant donnee que l’ACP n’est pas une methode fait pour la classification, on ne s’attend pas a` ce que les classes soient bien separees. Neanmoins avec ce jeu de donnees, les classes sont bien visibles.

1. ADL

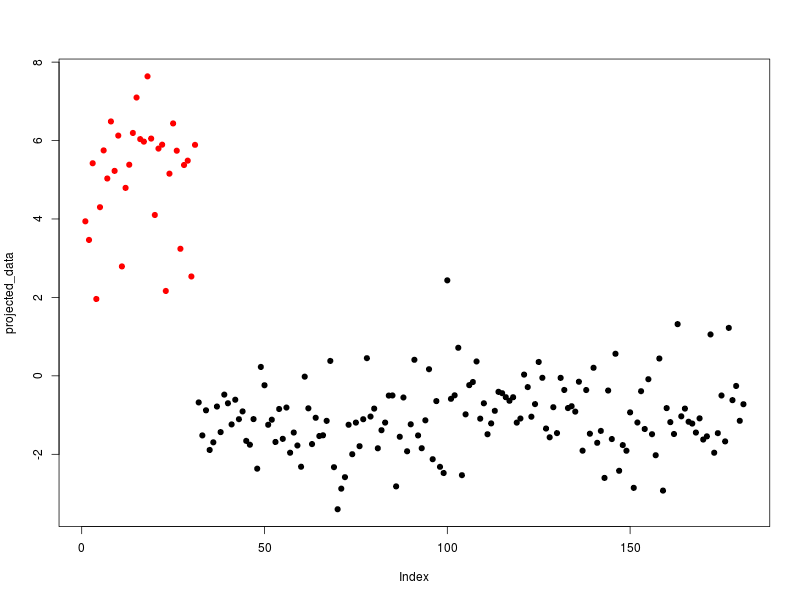


FIGURE 3 : Representation des individus sur l’axe factorielle de l’ADL

Avec l’ADL, le choix du nombre d’axes discriminants depend du nombre de classes. Dans ce jeu de donnees, 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 separees.

1. MDS

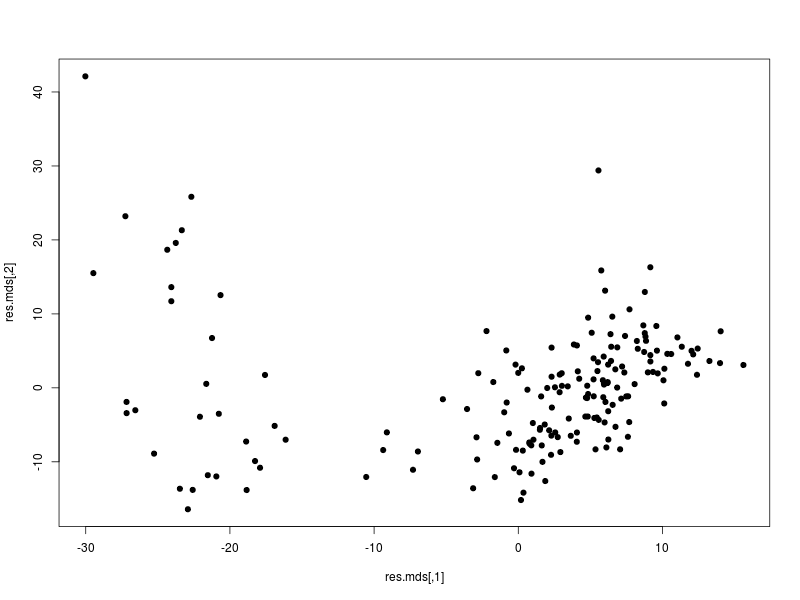


FIGURE 4 : Representation des individus sur deux axes factorielle du MDS

C’est une methode de non supervisee, aucune connaissance prealable des classes. Visuellement, la MDS offre une representation en deux groupes de densite differente. Il est impossible de savoir a` ce niveau si les individus ont été bien classees.

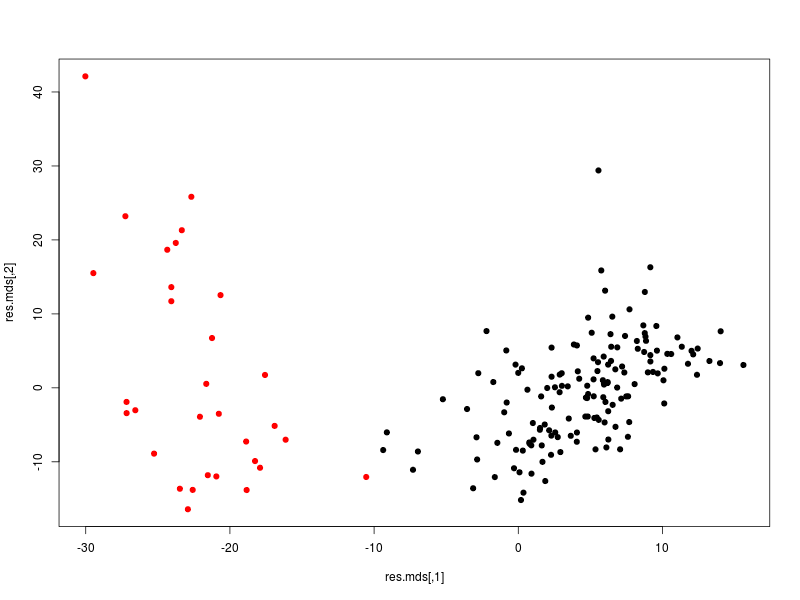


FIGURE 5 : Representation des individus avec des labels sur deux axes factorielle du MDS

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

1. Isomap

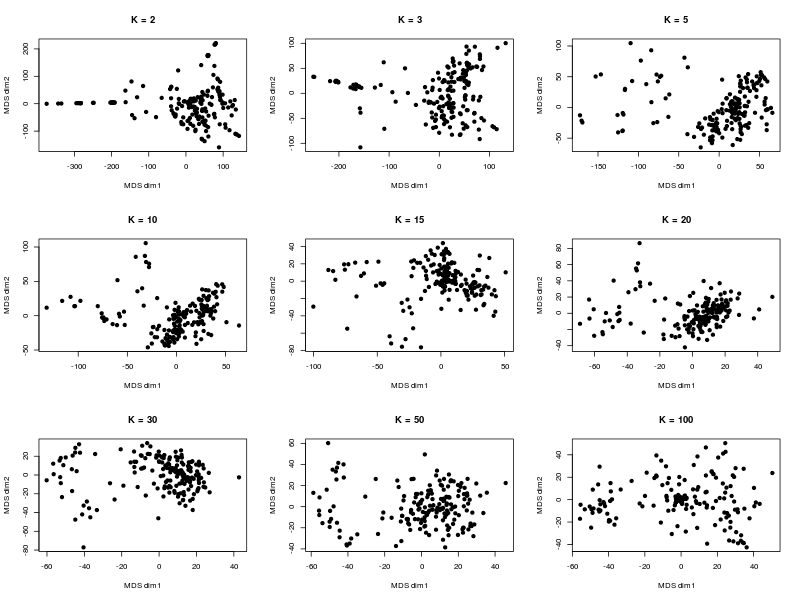


FIGURE 6 : Representation des individus en fonction de K plus proches voisins sur deux axes factorielle du MDS avec l’algorithme Isomap

En fonction des K, on observe une representation differente 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 heterogenes.

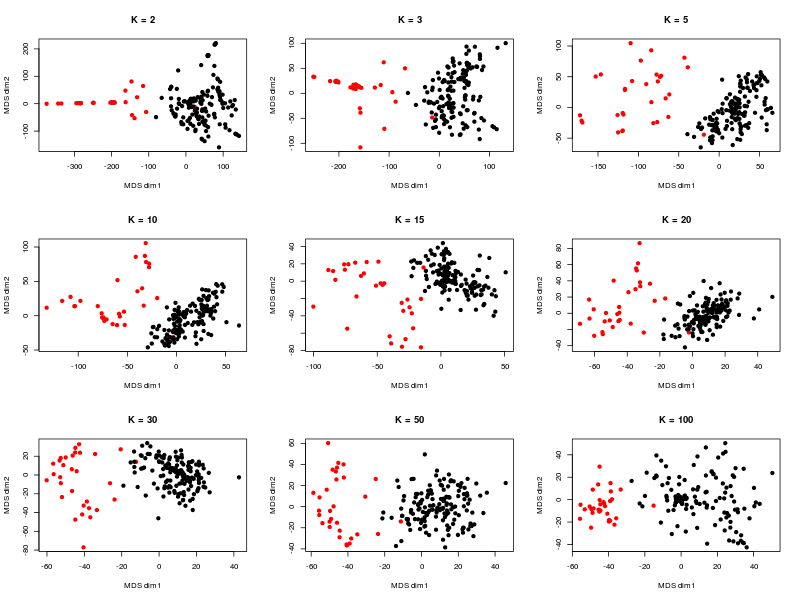


FIGURE 7 : Representation des individus labelises en fonction de K plus proches voisins sur deux axes factorielle du MDS avec l’algorithme Isomap

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

1. LLE

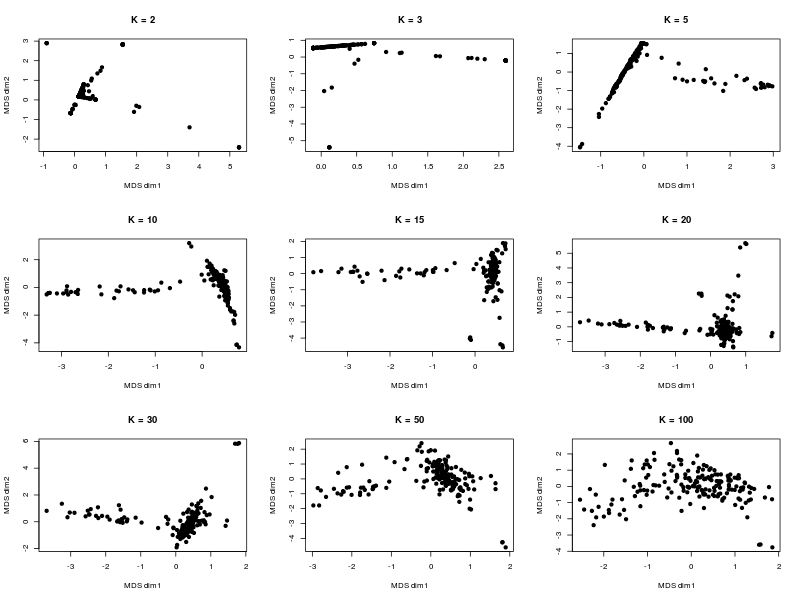


FIGURE 8 : Representation des individus en fonction de K plus proches voisins sur deux axes factorielle du MDS avec l’algorithme LLE

La forme des classe obtenue est differente. En fonction des K, on observe une representation differente 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 heterogenes.

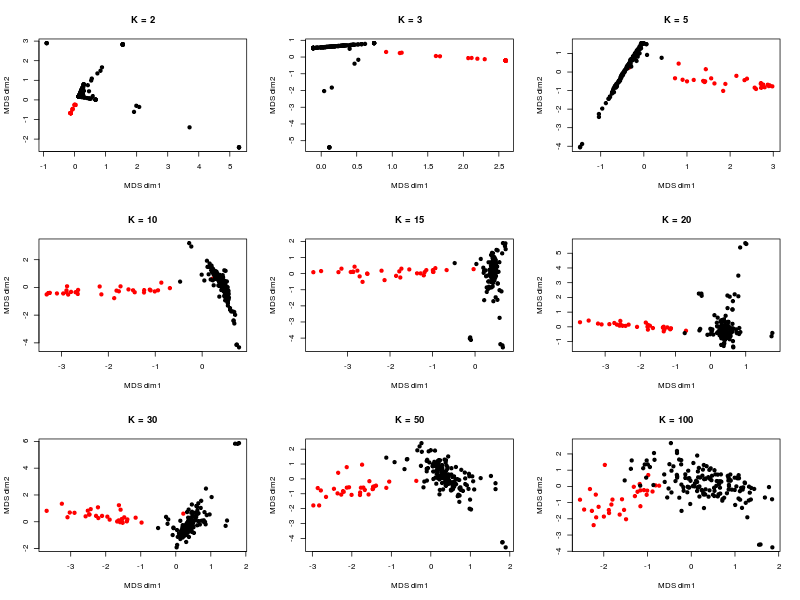


FIGURE 9 : Representation des individus labelises en fonction de K plus proches voisins sur deux axes factorielle du MDS avec l’algorithme LLE

Avec K = 5 et K = 10, tous les individus sont mis dans les bonnes classes excepte’ d’un individu mal classe’ .

1. SOM

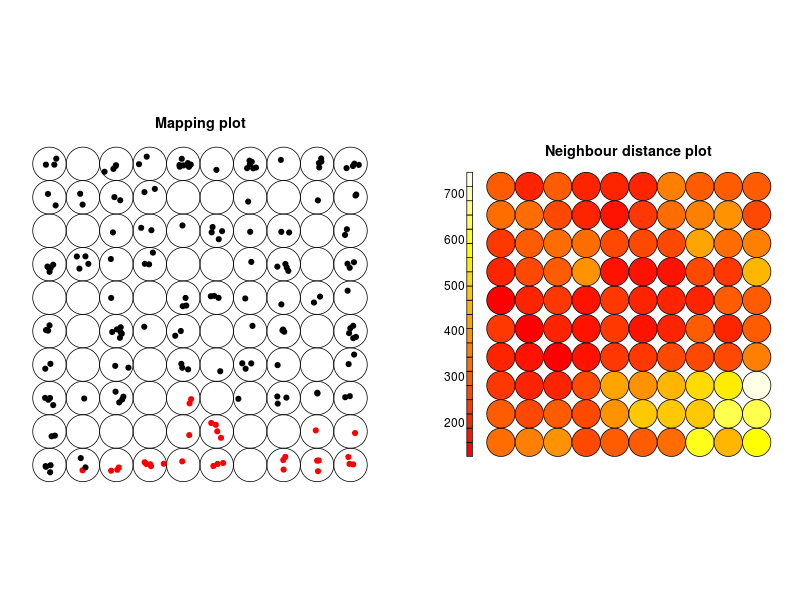


FIGURE 10 : Mapping plot et U-matrice

1. Pomeroy
2. ACP

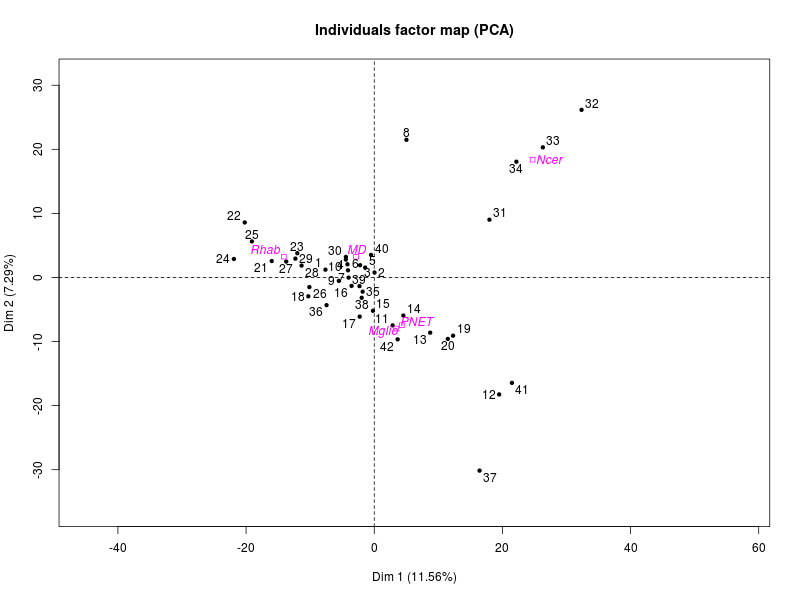


FIGURE 11 : Representation des individus sur 2 premiers axes factorielles de l’ACP

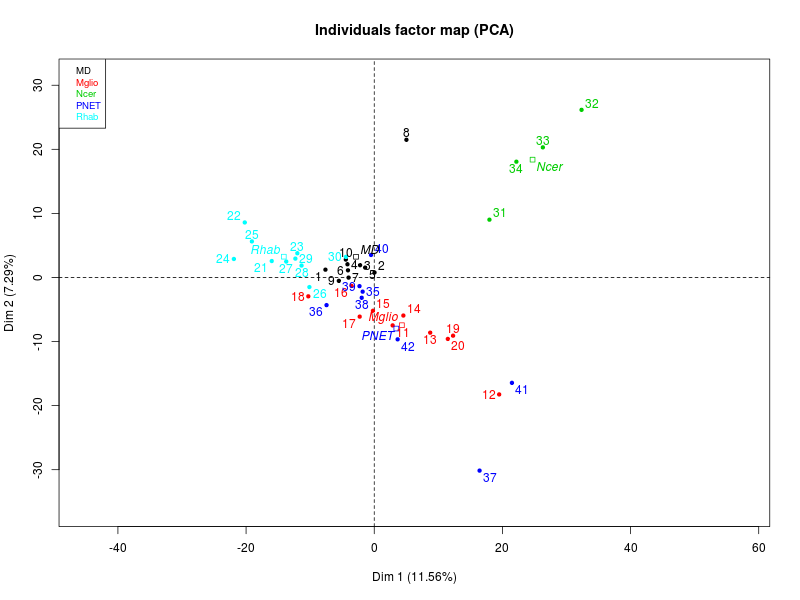


FIGURE 12 : Representation des individus avec des labels sur 2 premiers axes factorielles de l’ACP

1. ADL

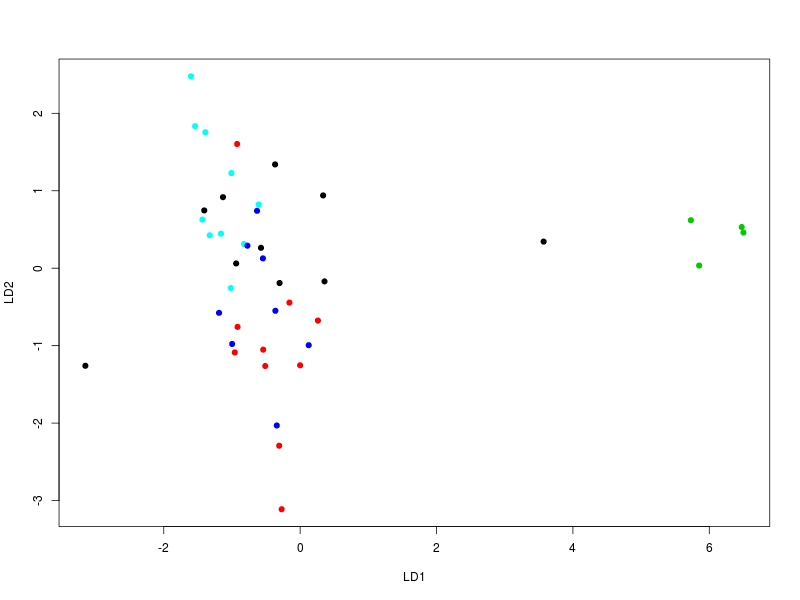


FIGURE 13 : Representation 2D des individus sur deux axes factorielle de l’ADL

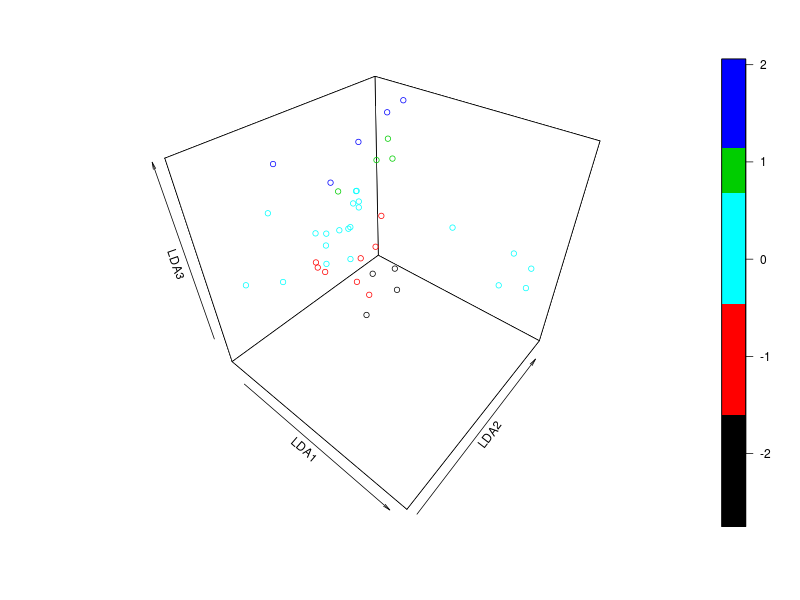


FIGURE 14 : Representation 3D des individus sur trois axes factorielle de l’ADL

1. MDS

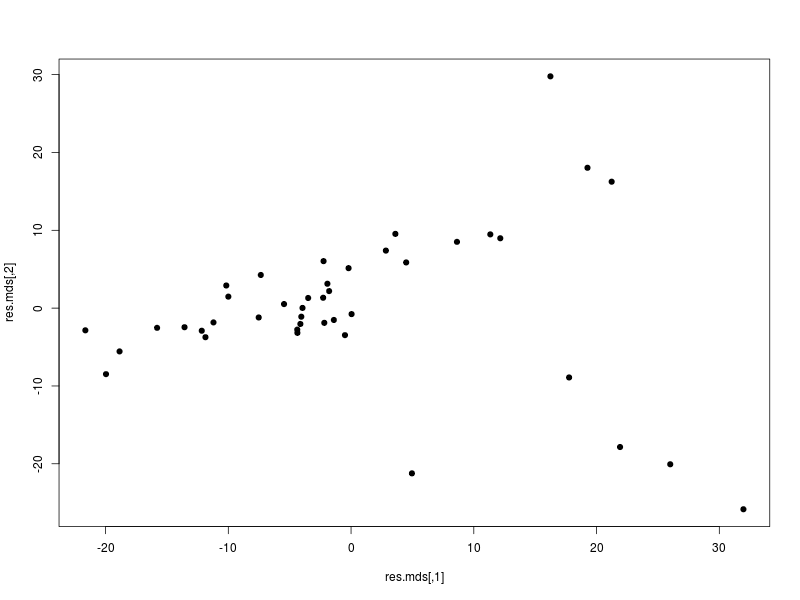


FIGURE 15 : Representation des individus sur deux axes factorielle du MDS

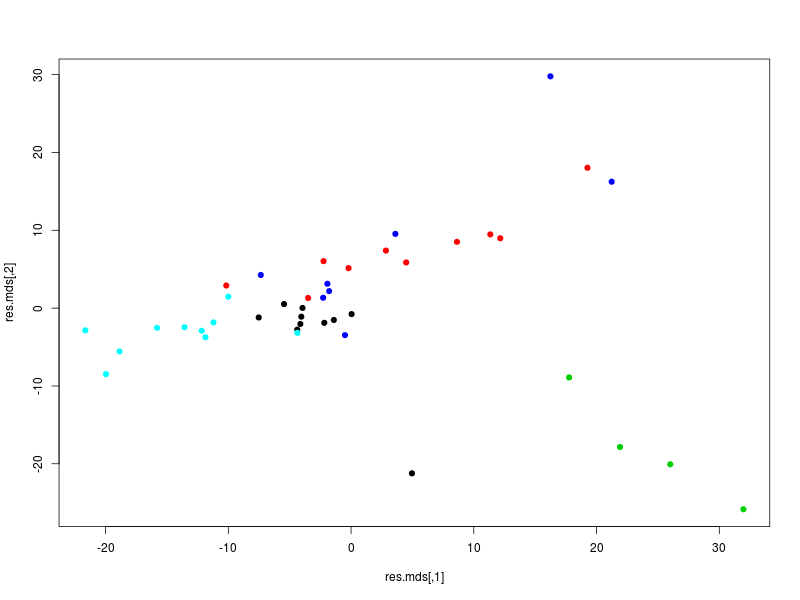


FIGURE 16 : Representation des individus labelises sur deux axes factorielle du MDS

1. Isomap

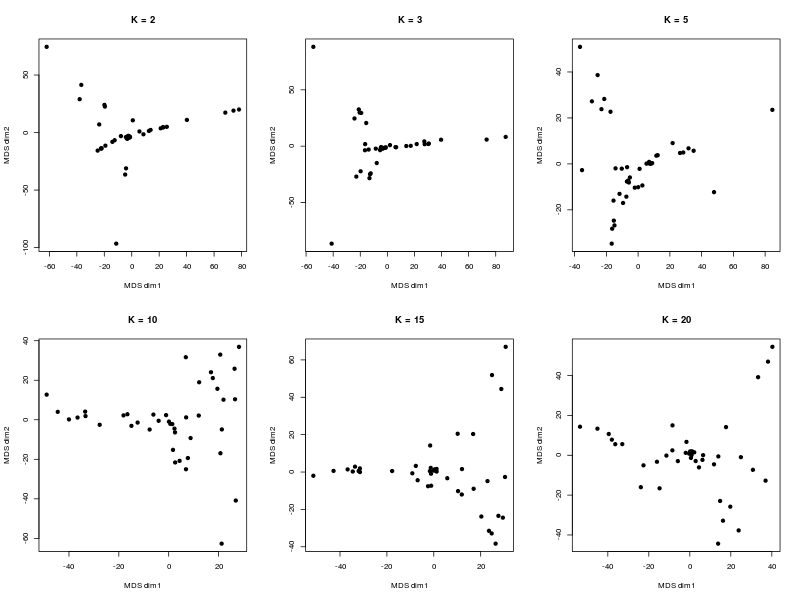


FIGURE 17 : Representation des individus en fonction de K plus proches voisins sur deux axes factorielle du MDS avec l’algorithme Isomap

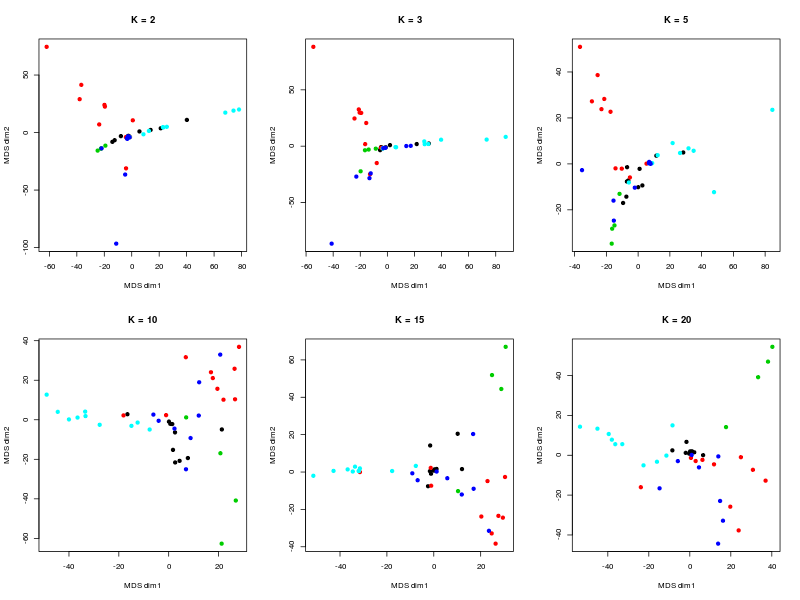


FIGURE 18 : Representation des individus labelises en fonction de K plus proches voisins sur deux axes factorielle du MDS avec l’algorithme Isomap

1. LLE

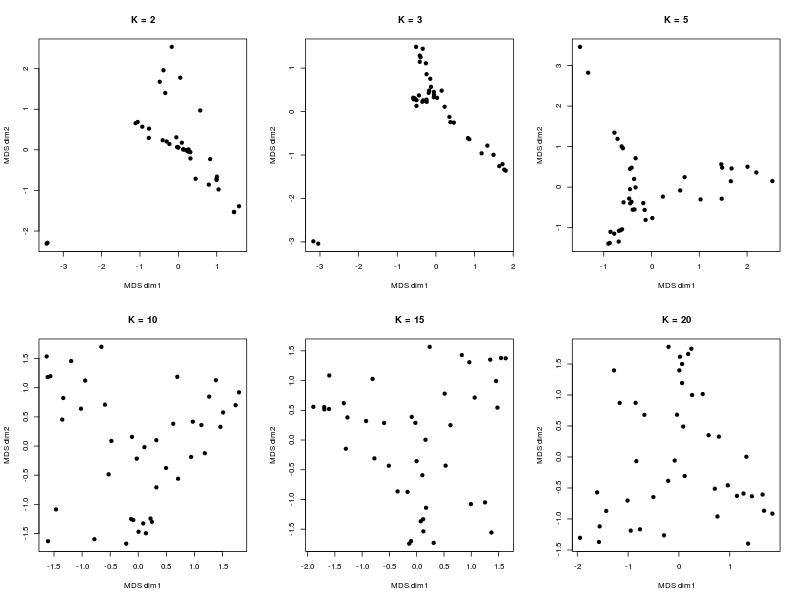


FIGURE 19 : Representation des individus en fonction de K plus proches voisins sur deux axes factorielle du MDS avec l’algorithme LLE

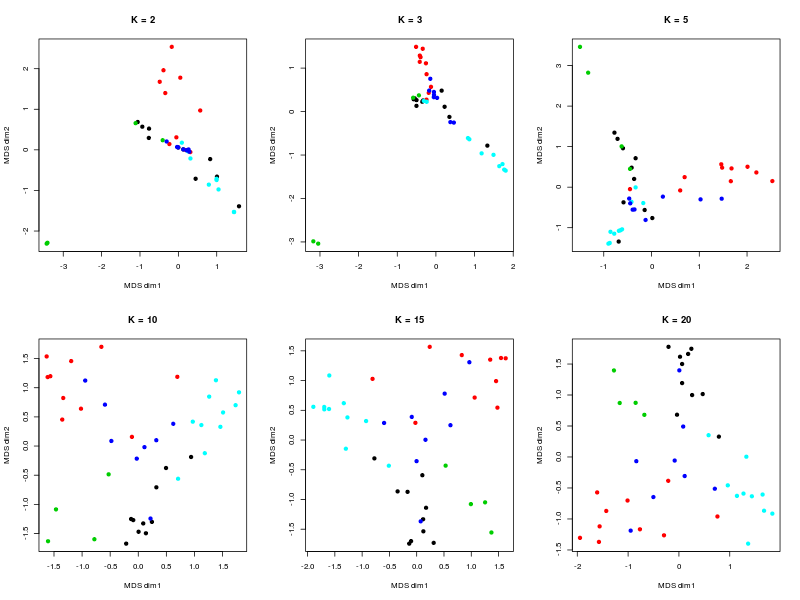


FIGURE 20 : Representation des individus labelises en fonction de K plus proches voisins sur deux axes factorielle du MDS avec l’algorithme LLE

1. SOM

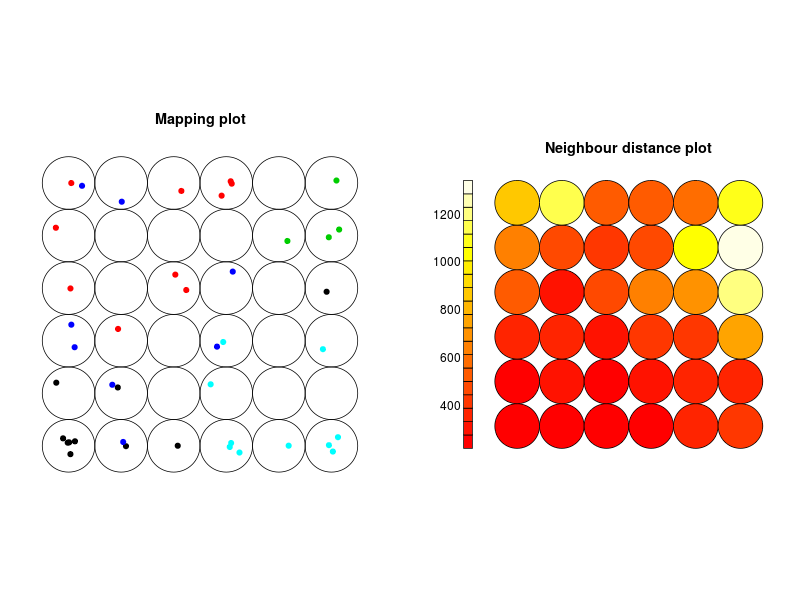


FIGURE 21 : Mapping plot et U-matrice

IV. Conclusion