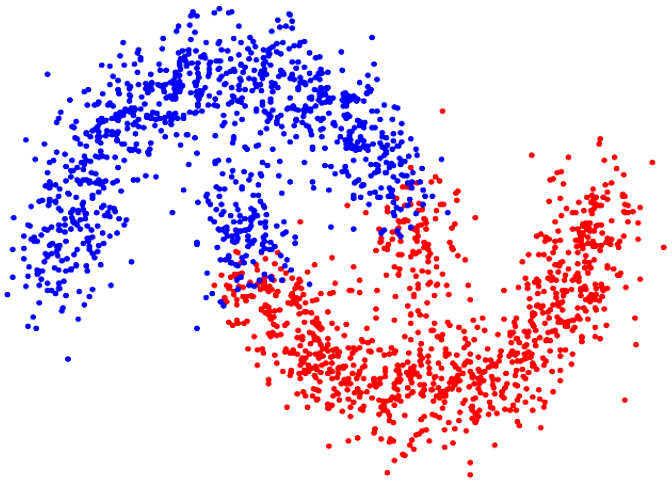
Master 2 Data Science

Université Lyon 1

Projet Analyse Factorielle



réalisé par
Florian Robinet et Hugo Duportet

Table des matières

1 Primar Tumor

1.1 Jeux de données

Le jeu de données comporte 961 observations. Pour chaque individus on observe une variable de réponse, une variable non-prédicative et quatre variables prédicatives.

1.1.1 Variable quantitative

Il y a une variable quantitative qui est l'âge. On procède à une discrétisation de cette variable afin de pouvoir l'utiliser dans l'analyse. Celle-ci ce fait par rapport aux quartiles, on obtient :

- **Catégorie 1 :** moins de 29 ans
- **Catégorie 2 :** 30-41 ans
- **Catégorie 3 :** 42-50 ans
- **Catégorie 4 :** Plus de 50 ans

1.1.2 Variables qualitatives

La variable de réponse concerne la gravité de la tumeur : bénin (0) ou malin (1). La distribution est relativement équilibrée avec 516 bénins et 445 malins.

Les variables prédicatives sont au nombre de quatre :

- **Critère BI-RADS :** 0 biopsie incomplète et 2,3,4,5,6 (gravité croissante)
- **Forme de la tumeur :** rond (1), ovale (2), lobulaire (3) ou irrégulier (4)
- **Marge de la tumeur :** circonscrite (1), macrolobulée (2), obscure (3), définie (4) ou spiculée (5)
- **Densité de la tumeur :** forte (1), moyen (2) faible (3) ou contenant de la graisse (4)

1.1.3 Variables manquantes

Il existe de nombreuses données manquantes, au vu du nombre d'observations, on choisit de travailler sur un tableau sans données manquantes.

1.2 Les variables qualitatives sont-elles corrélées ?

1.2.1 Méthode AFCM

Méthode factorielle de réduction de dimension pour l'exploration statistique de données qualitatives complexes. Cette méthode est une généralisation de l'Analyse Factorielle des Correspondances, permettant de décrire les relations entre p variables qualitatives simultanément observées sur n individus.

On considère maintenant p variables qualitatives ($p3$) notées $X^j; j = 1, ..., p$, possédant respectivement c_j modalités, avec $c = \sum_{j=1}^{p} c_j$.

On suppose que ces variables sont observées sur les mêmes n individus, chacun affecté du poids 1/n.

Soit $X = [X_1|...|X_p]$ le tableau disjonctif complet des observations ($Xestnc$) et $B = X'X$ le tableau de Burt correspondant (B est carré d'ordre c, symétrique).

DEFINITION : On appelle Analyse Factorielle des Correspondances Multiples (AFCM) des variables $(X^1, ..., X^p)$ relativement à l'échantillon considéré, l'AFC réalisée soit sur la matrice X soit sur la matrice B.

1.2.2 Résultats

Pour étudier les les corrélation on réalise l'AFCM et on représente graphiquement les résultats.

Dans un premier temps on représente le cercle de corrélation des variables. On peut donc étudier les corrélations entre celles-ci dans l'espace projeté.

Ce qui est intéressant ici c'est que le premier axe de l'AFCM coupe quasi parfaitement la variable de réponse Sévérité. Cela permet une facilité de lecture sur l'influence de l'ensemble des autres variables concernant la variable de réponse. On peut voir par exemple que le fait d'être âgé (plus de 50 ans), d'avoir une tumeur de forme irrégulière, et d'avoir une biopsie de niveau 5 est fortement corrélé avec le fait que cette tumeur soit malignes.

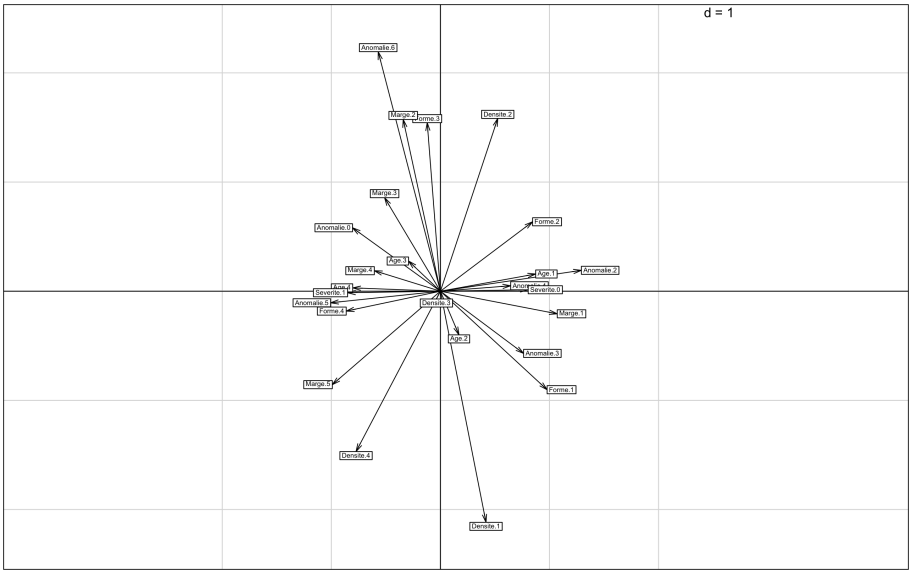


FIGURE 1 – Représentation de l'influence des variables dans la projection

Pour obtenir légèrement plus de précision il est possible de réaliser un boxplot de chaque valeur pour chaque attribut projeté sur le premier axe.

C'est ce que nous avons représenté sur la figure suivante. Comme précédemment on voit bien que la variable de réponse est expliquée par le premier axe. On peut voir que la population est âgée plus elle est à risque. La gravité de la tumeur est croissante parallèlement à la classe de la forme. Concernant la marge seule la marge circonscrite est corrélée positivement à la non dangerosité de la tumeur. Pour la densité, si la tumeur contient de la graisse, elle est corrélée positivement à la gravité de la tumeur.

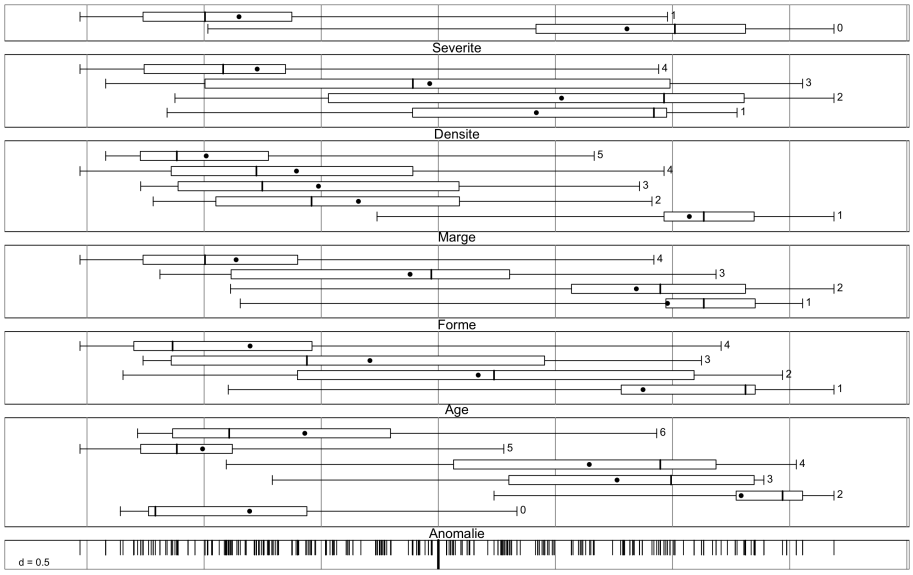


FIGURE 2 – Boxplot des modalités selon le premier axe

Il est aussi possible de réaliser une représentation graphique des individus dans l'espace projeté coloré en fonction de la valeur de l'attribut. Ce type de représentation permet une lecture plus aisé des résultats de l'AFCM.

On peut donc voir dans le graphique suivant la représentation des classes dans l'espace projeté. Comme le premier axe sépare la sévérité de la tumeur, tous les éléments à droite de l'axe des ordonnés seront corrélés positivement avec le fait que la tumeur soit bénigne et ceux à droite avec le fait que la tumeur soit maligne. De même il est possible d'observer d'autre types de corrélations entre les différentes valeurs des attributs. Par exemple on peut voir que la marge circonscrite est corrélée positivement avec une forme ronde ou ovale de la tumeur.

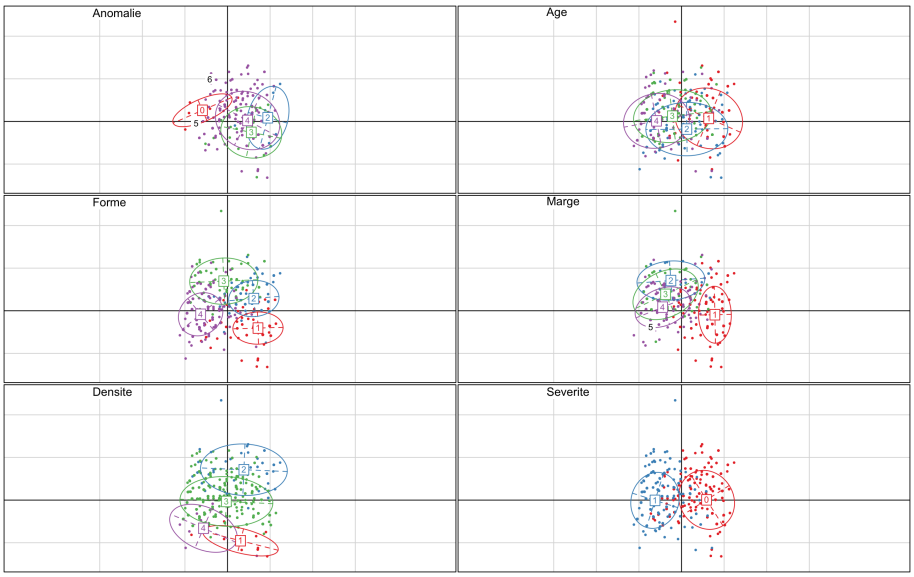


FIGURE 3 – Représentation par classes des données projetées

1.3 La gravité du cancer est-elle décrite par les caractéristiques obtenues à la suite de la biopsie ?

On étudie enfin plus en détail les corrélations entre le résultat de la biopsie et la gravité de la tumeur, en représentant en parallèle la classification par sévérité de la tumeur et par classe d'anomalie.

On voit dans la figure suivante que les tumeurs malignes vont être corrélées positivement avec avec les classes 6, 5 et 0 de la biopsie. A l'inverse les tumeurs bénignes vont être corrélées positivement avec les classes 2, 3 et 4 de la biopsie. Ce résultat parait cohérent dans le sens ou le niveau de gravité traduit par la biopsie et croissant. Pour la catégorie 0, il s'agit selon la classification BI-RADS des biopsies dont l'investigation est incomplète, on peut l'interpréter comme l'origine d'une complication à investiguer.

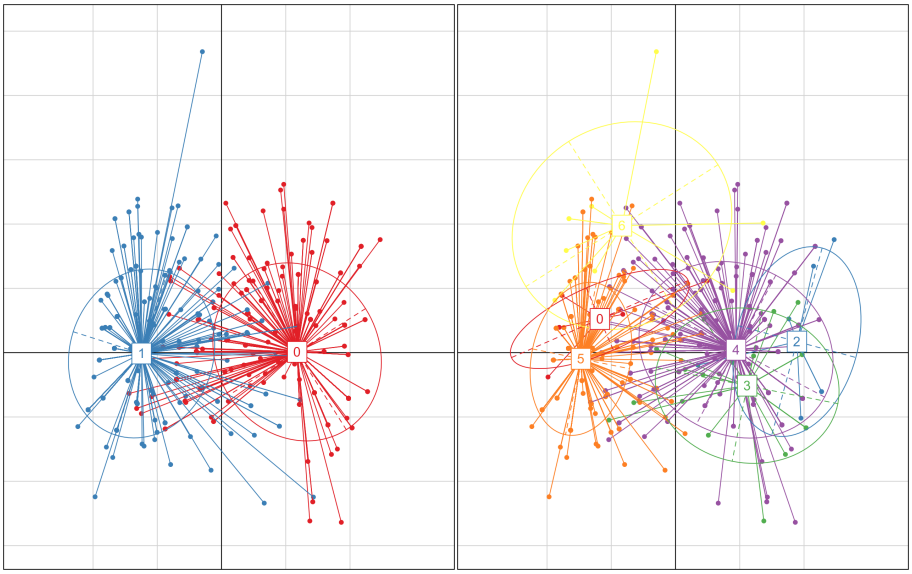


FIGURE 4 – Classification de la severité et des resultats de la biopsie

2 Stone Flakes

2.1 Jeux de données

L'ensemble de données traite de l'étude de la partie la plus ancienne de l'humanité. Les hommes préhistoriques fabriquaient des outils de pierre en frappant sur de la pierre brute, et créant ainsi des éclats (flakes) de pierres, résidus du processus d'artisanat. Il est difficile pour les archéologues de trouver des outils relatifs à cette époque mais trouve cependant des éclats de pierres en plus grand nombre. L'ensemble de données *StoneFlakes* concerne ces éclats.
Le jeu de données comporte 79 observations de regroupements d'éclats de pierre. Il y a 8 variables quantitatives et 7 variables qualitatives. La variable cible du jeu de données est la variable groupe, qui représente l'appartenance à un groupe caractéristique d'une période temporelle de la préhistoire. On notera que le groupe homo sapiens n'a que 3 représentants et qu'il ne représente peut être pas très fidèlement la réalité. Le groupe Néandertal et Technique Levallois correspondent à une même époque (paléolithique moyen) et se distingue sur leur technique de taille des pierres ou silex. La technique Levallois est une méthode de débitage particulièrement avancée de la pierre employée au cours de cette époque.

2.1.1 Variables quantitatives

- **LBI :** Indice de largeur/longueur de la surface du plan de frappe
- **RTI :** Épaisseur relative de la surface du plan de frappe
- **WDI :** Indice de largeur/profondeur du plan de frappe
- **FLA :** Angle d'écaillage (l'angle entre la surface de frappe et la surface de séparation)
- **PSF :** Plan de frappe primaire/rudimentaire (oui/non, fréquence relative)
- **FSF :** Plan de frappe facettés/complexe (oui/non, fréquence relative)
- **ZDF1 :** Surface plane totalement travaillée (oui/non, fréquence relative)
- **PROZD :** Proportion de surface plane travaillée (continu)

2.1.2 Variables qualitatives

- **group :** Groupe défini par les archéologistes (1= paléolithique inferieur, Homo ergaster, plus vieux ; 2=Méthode Levallois ; 3=paléolithique moyen, Homme de Néandertal ; 4=Homo sapiens, plus jeune)
- **age :** Age des artefacts retrouvés (très approximatif)
- **dating :** Mode de datation (geological=plus précis, typological)
- **mat :** Constituon de la pierre (1=silex, 2=autres)
- **region :** Region (mit=Central Germany, d=Non-Central Germany, eur=Europe without Germany
- **site :** site (1=carrière de gravier, 0=autre)
- **number :** nombres d'éclats retrouvés par regroupement

2.2 Does the data reflect the technological progress during several hundred thousand years ?

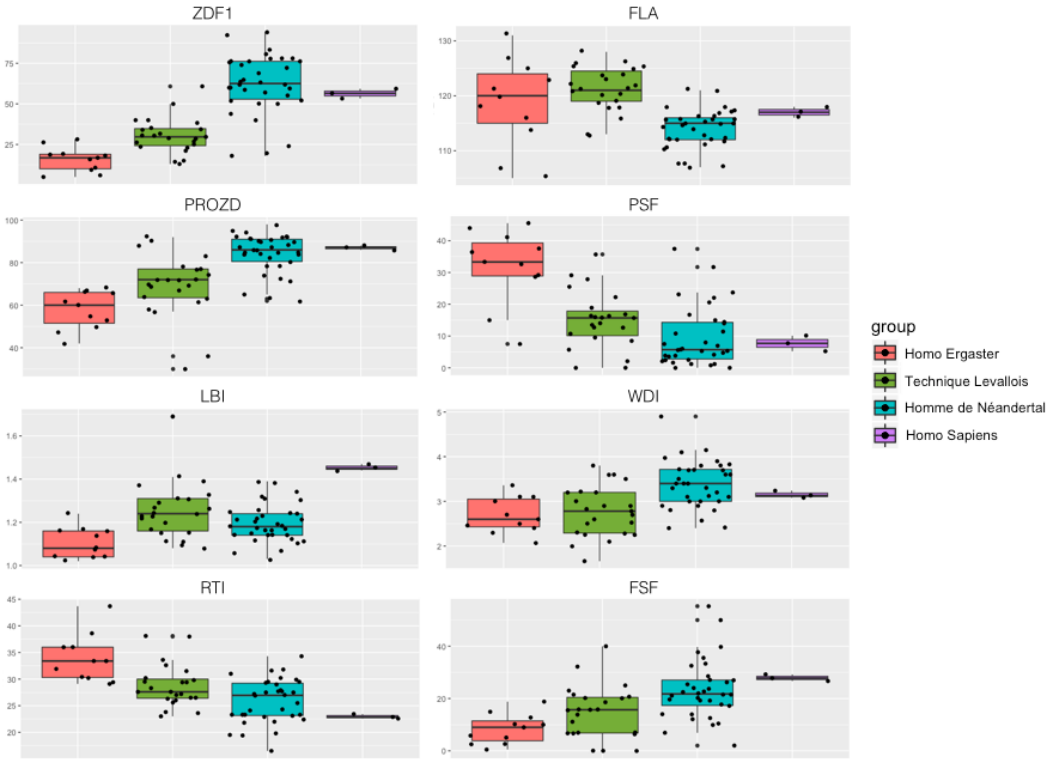


FIGURE 5 – Boxplot des différents attributs numériques en fonction du groupe

Étant donné que la variable age est sensiblement équivalente à la variable groupe et que la description des données la présente comme très vague, on se contente d'afficher l'évolution des variables numériques en fonction des groupes.

- **LBI :** Légère augmentation avec le temps entre le paléolithique inférieur et moyen, forte augmentation pour homo sapiens par rapport au reste.
- **RTI :** Décroissance au cours du temps, cela témoigne d'une taille plus fine de la roche, et donc plus précise.
- **WDI :** Pas de changements significatifs.
- **FLA :** Pas de changements significatifs.
- **PSF :** Le plan de frappe est bien plus souvent rudimentaire pour le paléolithique inférieur que pour les autres époques.
- **FSF :** À l'inverse, le plan de frappe est de plus en plus souvent complexe pour les époques postérieures au paléonlitique inférieur, les pierres sont

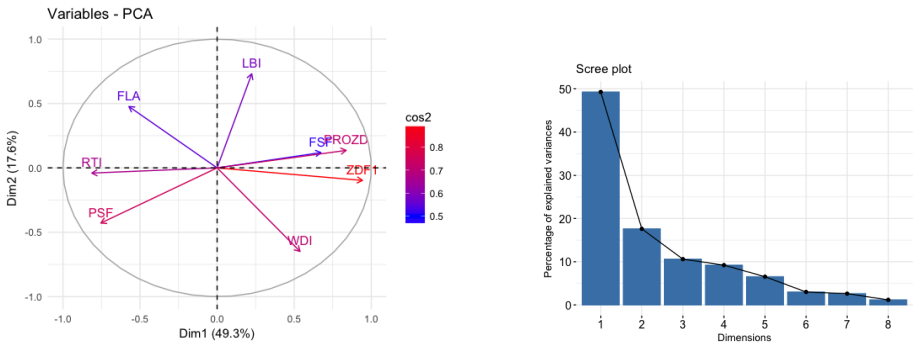
donc de mieux en mieux travaillées.

— **ZDF1 :** La fréquence de la surface plane totalement travaillée augmente sensiblement entre le paléolitique inférieur et les les époques postérieures, les homo sapiens et hommes de Néandertal ont une fréquence relative moyenne relativement similaire et très supérieure aux autres.

— **PROZD :** Le constat est le même que pour ZDF1.

Les données permettent bien de montrer un progrès technologique entre les différentes périodes de la préhistoire. Les Hommes ont progressivement affiné leurs techniques de production d'outils, ils ont travaillés plus en profondeur et ont utilisé de plus en plus de surfaces sans doute afin de les adapter au mieux à leur besoin. On peut aussi supposer que les Hommes utilisant la technique Levallois devaient se situer entre la période de l'Homo Ergaster (paléolithique inférieur) et l'Homme de Néandertal (paléolithique moyen). Les variables LBI, RTI et FSF permettent de faire la différence entre l'Homme de Néandertal et l'Homo Sapiens.

2.3 Relation entre les variables numériques



(a) Cercle de corrélation

(b) Variance expliquée

FIGURE 6 – Analyse en composante principale

Les deux premières composantes reconstituent pratiquement 70 % de la variance de départ. On remarque que sur le premier axe, les variables PROZD, FSF et ZDF1 sont positivement corrélées tandis que les variables RTI et PSF le sont négativement. Cela signifie que la proportion de surface plane travaillée, la surface plane totalement travaillée et la complexité du plan de frappe sont positivement corrélés tandis que l'épaisseur relative de la surface du plan de frappe et la rudimentarité du plan de frappe le sont négativement. C'est en effet cohérent puisque un outil plus fin nécessite un travail plus complexe. Sur le deuxième axe, la variable LBI est postivement corrélée tandis que la variables WDI l'est négativement. Cela signifie que l'indice de largeur/longueur augmente lorsque que l'indice de largeur/profondeur diminue.

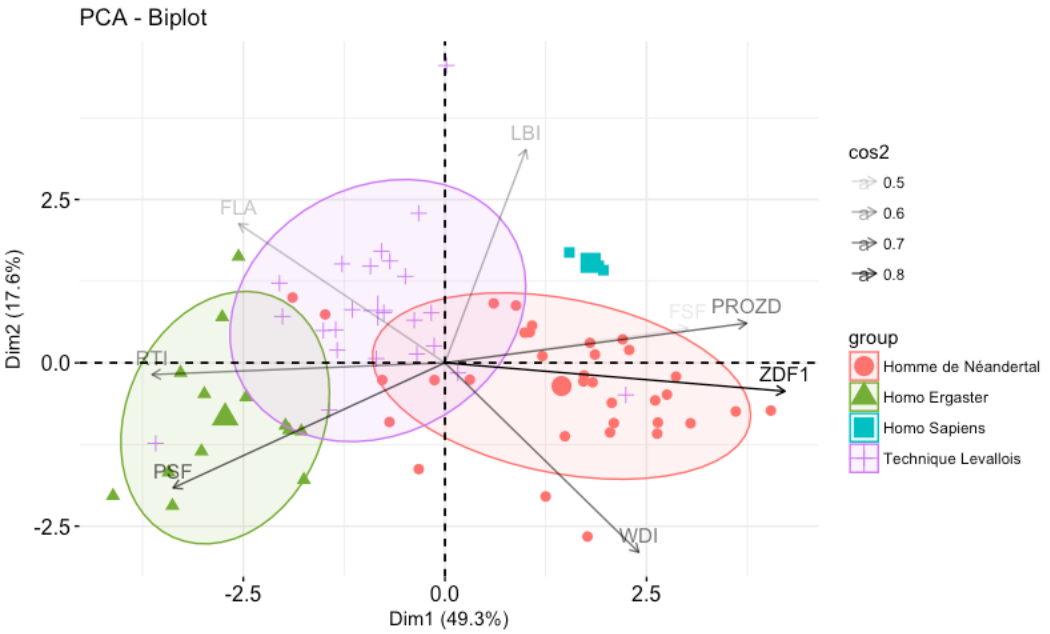
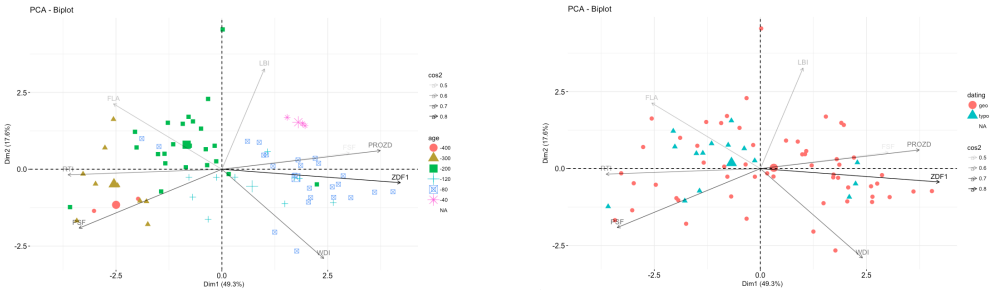


FIGURE 7 – ACP avec groupe d'éclats et coloration par groupe

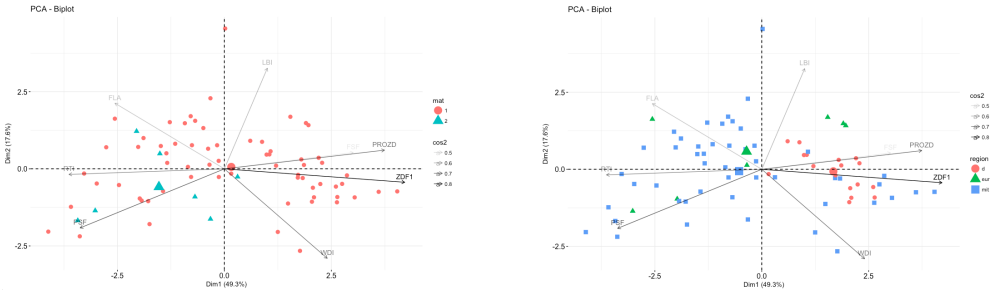
Ce graphe confirme bien nos suppositions du départ, les groupe d'éclats du paléolithique correspondent aux caractéristiques rudimentaires et peu complexes et les groupes se décalent peu à peu vers plus de finesse et plus de complexité dans la taille des roches. Cette figure confirme bien l'amélioration des techniques de taille de pierre par les Hommes au cours de la préhistoire.



(a) age (b) dating

FIGURE 8 – Analyse en composante principale pour les variables qualitatives

On voit que l'age suit bien la variable groupe. L'acp ne permet pas de conclure quant à la méthode de datation. Pour le type de pierre taillée, on



(a) mat (b) region

FIGURE 9 – Analyse en composante principale pour les variables qualitatives

voit que le silex est rapidement devenu la pierre de référence. Les individus de germanie devait plutôt être des hommes de Néandertal et ceux du reste de l'europe devaient être soit de Néandertal soit Ergaster.

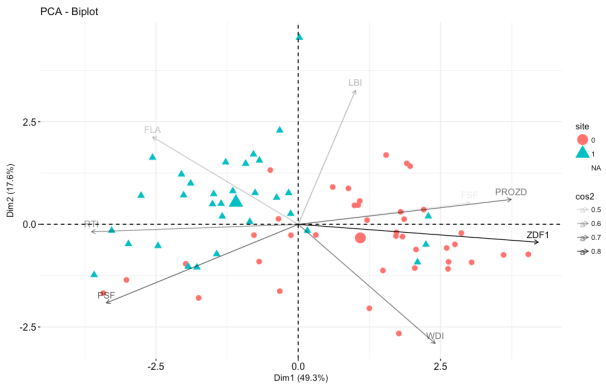


FIGURE 10 – Boxplot des différents attributs numériques en fonction du groupe

On voit que les copeaux plus jeunes proviennent plutôt de carrière et les plus anciens non. Cela montre que l'Homme a progressivement pensé à s'abriter dans des grottes sans doute.

3 Parkinson

3.1 Présentation des données

Les données ont été recueillies lors d'une étude à l'université d'Oxford en collaboration avec 10 centres médicaux américains. Les données sont composées d'entier et de réels. Il y a 5875 observations et 26 attributs.

Les données contiennet un ensemble de mesures vocales (16 mesures) biomédicales de 42 personnes (Unified Parkinson Disease Rating Scale). Il y a en moyenne 200 enregistrements par sujets.

3.2 Variables

subject# : Identifiant unique à chaque sujet.
age : Age du sujet.
sex : Sexe du sujet '0' = homme, '1' = femme.
test_time : Temps total depuis le début des prises.
motor_UPDRS : Score UPDRS moteur clinique , linéairement interpolé.
total_UPDRS :Score UPDRS total clinique , linéairement interpolé.
Jitter(%), Jitter(Abs), Jitter :RAP, Jitter :PPQ5, Jitter :DDP : Plusieurs mesures de variation en fréquence fondamentale.
Shimmer, Shimmer(dB), Shimmer :APQ3, Shimmer :APQ5, Shimmer :APQ11, Shimmer :DDA : Plusieurs mesures de variations en amplitude
NHR, HNR : Deux mesures de ratio du bruit sur la tonalité de la voix
RPDE : Mesure non linéaire de complexité dynamique
DFA : Exposant de mise à l'échelle de la fractal du signal
PPE : Mesure non linéaire de la variation fondamentale de la fréquence

3.3 ACC

Procédure CANCORR
Analyse canonique de redondance

Variance brute de la(du) Variables VAR expliqué(e) par					
	Leurs propres variables canoniques			Les variables canoniques inverses	
Nombre de variables canoniques	Proportion	Proportion cumulée	R carré canonique	Proportion	Proportion cumulée
1	0.9566	0.9566	0.1656	0.1585	0.1585
2	0.0434	1.0000	0.0630	0.0027	0.1612

FIGURE 11 – Analyse canonique de redondance

On peut expliquer 95% de la variance des variables avec le premier axe.

Procédure CANCORR
Structure canonique

Corrélations entre le(la) Variables VAR et leurs variables canoniques		
	V1	V2
motor_UPDRS	0.9393	-0.3431
total_UPDRS	0.9997	-0.0239

Corrélations entre le(la) Variables WITH et leurs variables canoniques		
	W1	W2
age	0.7659	0.2222
test_time	0.1855	0.0348
Jitter___	0.1797	-0.1871
Jitter_Abs_	0.1667	0.1489
Jitter_RAP	0.1550	-0.1558
Jitter_PPQ5	0.1526	-0.2084
Jitter_DDP	0.1551	-0.1558
Shimmer	0.2236	-0.1960
Shimmer_dB_	0.2396	-0.2144
Shimmer_APQ3	0.1933	-0.1205
Shimmer_APQ5	0.2026	-0.1700
Shimmer_APQ11	0.2928	-0.2861
Shimmer_DDA	0.1933	-0.1204
NHR	0.1466	-0.2199
HNR	-0.3976	0.0585
RPDE	0.3890	0.2337
DFA	-0.2771	0.1196
PPE	0.3810	-0.1948

FIGURE 12 – Corrélations entre les variables

On peut voir que l'age influence fortement le développement de la maladie de Parkinson. Plus l'age est important plus le développement de la maladie peut être important. Les variables RPDE et PPE influence aussi positivement le développement de la maladie. La variable HNR influence négativement le développement de la maladie. Plutôt que d'utiliser ces variables, on utilise la première variable de l'ACC. Étant donné qu'elle représente 90% de la variance des autres variables, on se fie à celle ci pour expliquer les résultats des sujets. Plus cette variable sera positive, plus la maladie de Parkinson aura des chances d'être accentuée.

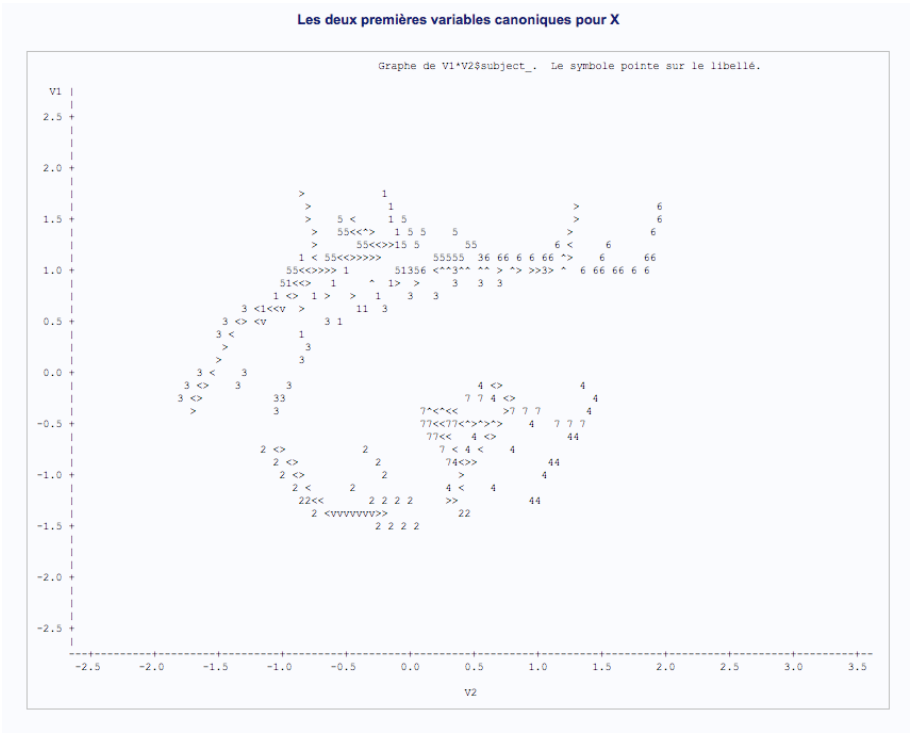


FIGURE 13 – Axes de principaux de l'ACC en fonction du sujet

4 Wine

4.1 Jeux de données

Les données sont issues d'analyses chimiques de différents vins (178) d'une même région d'Italie et de différents producteurs, trois au total. Les analyses permettent de déterminer 13 valeurs quantitatives représentant les caractéristiques chimiques de chaque vin.

Les différentes variables sont :
- Alcool
- Acides maliques
- Cendres
- Alcalinité
- Magnésium
- Total Phénols
- Flavanoides
- Non flavanoides phénols
- Proanthocyanidines
- Intensité de la couleur
- Teinte
- OD
- Proline

Le jeu de données ne comporte pas de données manquantes.

Il y a de plus une dernière variable de type catégorielle qui indique quel est le producteur du vin.

4.2 Quelles sont les relations entre les caractéristiques chimiques des vins ?

On commence par réaliser une matrice de corrélation entre les différentes variables hormis la variable de réponse.

FIGURE 14 – Représentation de la matrice de corrélation entre les variables

On peut voir des corrélations positives entre les Phenols et les Flavanoides, mais aussi entre ces deux molécules et l'OD, enfin entre l'Alcool et la Proline.

Les corrélations négatives sont plus rares, on en voit entre la teinte et la couleur, entre les Flavanoides et les non Flavanoides, enfin entre la couleur et l'acide malique.

4.3 Les vins des trois vignerons sont-ils différents selon les caractéristiques chimiques ?

Pour étudier les similarités des des vins entre les trois vignerons on réalise une ACP. Celle-ci va permettre la mise en évidence des différences entre les vins des trois producteurs.

La représentation suivante est constituée de trois éléments importants. On a d'abord la représentation des différents vins dans un plan maximisant l'inertie. On a ensuite la visualisation de l'influence de chaque variable dans cet espace avec les flèches. Enfin on colorie les différents vins en fonction du producteur

pour faire ressortir les clusters.

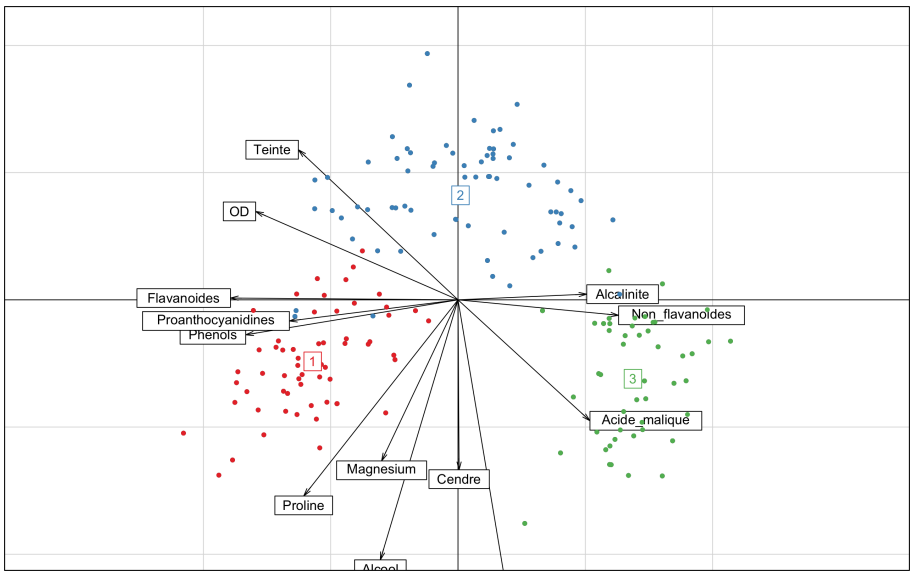


FIGURE 15 – Représentation des variables dans l'espace projeté classées par vigneron (ACP)

On voit donc que les vins se distinguent selon le producteur en fonction de leurs caractéristiques chimiques.

Les vins du producteur 1 seront plus concentrés en Flavanoides, en Proanthocyanidines, en Phenols, en Proline, en OD, en Teinte, en Magnesium et en Alcool.

Les vins du producteur 2 seront moins concentrés en Proline, en Magnesium, en Alcool, en Cendre et en Couleur.

Les vins du producteur 3 seront plus concentrés en Alcalinite, en Non flavanoides, et en Acides maliques.

4.4 Caractérisation des vins des trois vignerons

Pour obtenir plus de précision sur la caractérisation des vins de chaque vignerons on réalise une AFD. Cette analyse va permettre de quantifier les caractéristiques chimiques des vins des différents producteurs. A la différence de l'ACP ce modèle est prédictif et pourra éventuellement servir à prédire quel est le producteur d'un nouveau vin. Le premier et le second axe se caractérisent comme une

combinaison linéaire des variables explicatives. La variable de réponse dans ce modèle est le producteur.

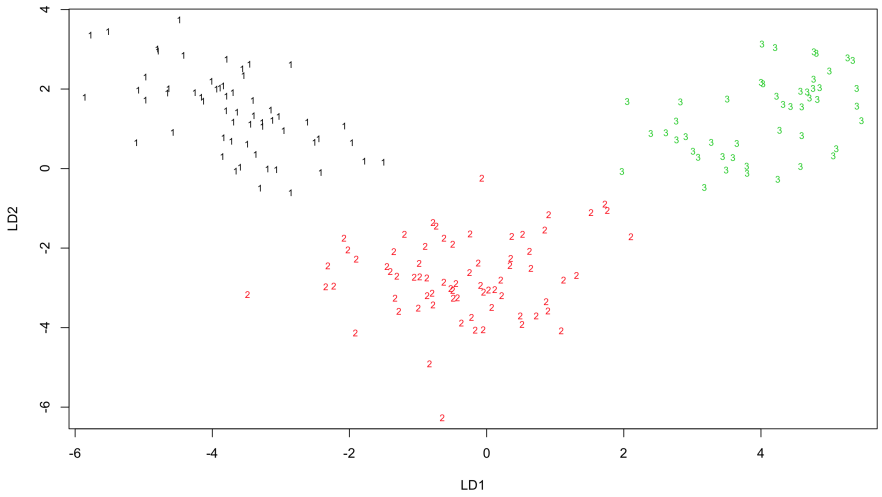


FIGURE 16 – Représentation des vins dans l'espace projeté classées par vigneron (AFD) et coefficients des axes

	LD1	LD2
Alcool	-0.403399781	0.8717930699
Acide_malique	0.165254596	0.3053797325
Cendre	-0.369075256	2.3458497486
Alcalinite	0.154797889	-0.1463807654
Magnesium	-0.002163496	-0.0004627565
Phenols	0.618052068	-0.0322128171
Flavanoides	-1.661191235	-0.4919980543
Non_flavanoides	-1.495818440	-1.6309537953
Proanthocyanidines	0.134092628	-0.3070875776
Couleur	0.355055710	0.2532306865
Teinte	-0.818036073	-1.5156344987
OD	-1.157559376	0.0511839665
Proline	-0.002691206	0.0028529846

Cette représentation graphique permet de confirmer la réponse à la question précédente concernant le fait que les vins des trois vignerons se distinguent selon leurs caractéristiques chimiques.

On réalise ensuite un barplot des moyennes de chaque caractéristiques chimiques selon chaque vignerons. Cette représentation apporte des précisions numériques sur les caractéristiques chimiques des vins des trois vignerons.

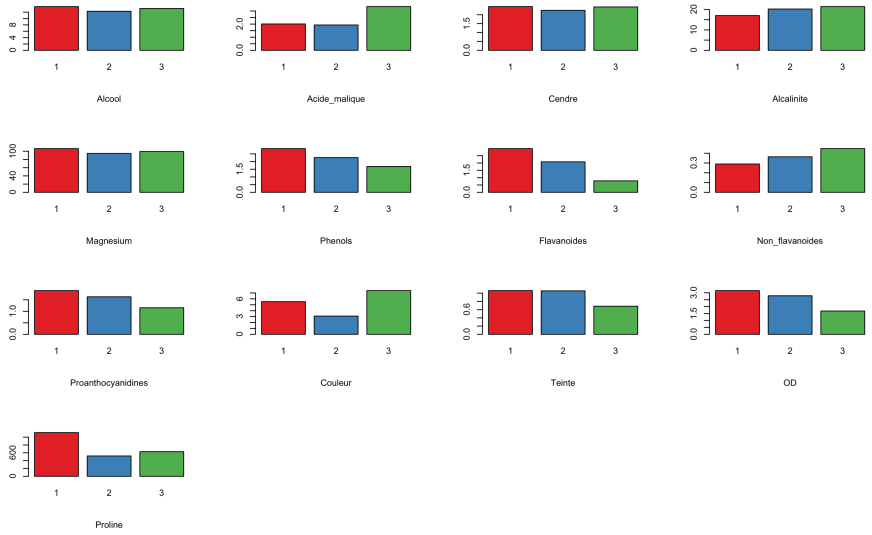


FIGURE 17 – Barplot des moyennes selon chaque vignerons pour chaque caractéristiques

On retrouve les éléments mis en évidence par l'ACP mais de manière quantifiée.

Les vins du producteur 1 sont plus concentrés en OD, en Proline, en Alcool, en Magnesium, en Phenols, en Flavanoides, en Proline et en Proanthocyanidines.

Les vins du producteur 2 sont moins concentrés en Couleur et en Proline.

Les vins du producteur 3 sont plus concentrés en Acides maliques, en non Flavanoides et en Couleur. Ils seront par contre moins concentrés en Phenols, en Flavanoides, en Proanthocyanidines, en Teinte et en OD.

Code - Exercice 1

```r
#Packages#
library(MASS)
library(ade4)
library(RColorBrewer)

####################DONNEES#########################################

#Importation du jeu de donnees#

table=read.table('~/Desktop/Mamo_Table.data', header=FALSE,
 ↪  sep=",")
colnames(table)=c("Anomalie", "Age", "Forme", "Marge",
 ↪  "Densite", "Severite")
table$Severite=as.factor(table$Severite)

#Discretisation de l'attribut age#

table <- subset(table, table$Age != "?")
table$Age=cut(as.numeric(table$Age),
 ↪  breaks=c(as.numeric(quantile(as.numeric(table$Age))[1]),
 ↪  as.numeric(quantile(as.numeric(table$Age))[2]),
 ↪  as.numeric(quantile(as.numeric(table$Age))[3]),
 ↪  as.numeric(quantile(as.numeric(table$Age))[4]),
 ↪  as.numeric(quantile(as.numeric(table$Age))[5])))
levels(table$Age)=c("1","2","3","4")

#Eviction des donnees manquantes#

table=replace(table,table=="?",NA)
table=na.omit(table)

#Eviction de donnees aberantes#

table$Anomalie=replace(table$Anomalie,table$Anomalie==55,5)
table=droplevels(table)
attach(table)

#Visulaisation peu performante des corrélations#

pairs(~Anomalie+Age+Forme+Marge+Densite, data=table,
 ↪  main="Simple Scatterplot Matrix", pch=20, col=Severite)


####################AFCM#############################################
```

```r

#Les variables sont qualitatives on pense à l'analyse factorielle de correspondances multiples

#On réalise l'AFCM en choisissant 2 axes#

acmsev <- dudi.acm(table,scannf=FALSE,nf=2)

#Il est possible de visualiser les valeurs propres#
barplot(acmsev$eig)

#On ecrit une fonction keepvar pour determiner le nombre d'axes a conserver pour un certain seuil d'information du modèle (ici 70%)#

var=acmsev$eig/sum(acmsev$eig)
keepvar= function (seuil){
        res=0
        for(i in 2:(length(var)+1)){
                res=res+var[i-1]
                if(res>seuil){
                        print(i)
                        print(res)
                        break
                }
        }
}

keepvar(0.7)

#On obtient 12 pour conserver 73% de l'information du modèle#

#On retient donc les 12 premiers axes#

acmsev <- dudi.acm(table,scannf=FALSE,nf=12)

#Il est possible d'obtenir la répartition dans les différentes classes de chaque attributs sous forme nominal#

matX <- as.matrix(acm.disjonctif(table))
matD <- diag(1/dim(acmsev$tab)[1], dim(acmsev$tab)[1], dim(acmsev$tab)[1])
mat1n <- rep(1,dim(acmsev$tab)[1])
t(matX)%*%mat1n

```

```r
#Il est possible d'obtenir la répartition dans les différentes classes de chaque attributs sous forme de pourcentages#

t(matX)%*%matD%*%mat1n

#Equivalent aux fréquences relatives associés à chaque variable#

frequences <- t(matX)%*%matD%*%mat1n
matDm <- diag(as.numeric(frequences), dim(t(matX)%*%mat1n)[1], dim(t(matX)%*%mat1n)[1])
round(matDm,4)

#Il est aussi possible de l'avoir sous forme plus exploitable#

sapply(1:6, function(x) round(summary(table[,x])/dim(acmsev$tab)[1],4))

#Il est possible de retrouver tabletabavec#

matDmm1 <- diag(1/as.numeric(frequences), dim(t(matX)%*%mat1n)[1], dim(t(matX)%*%mat1n)[1])
mat1nm <- matrix(1, nrow=dim(acmsev$tab)[1], ncol=dim(t(matX)%*%mat1n)[1])
matY <- matX%*%matDmm1-mat1nm
round(matY,4)


#Cercle de correlation des variables dans l'espace projeté#

s.arrow(acmsev$co,clabel = 0.5)

#Boxplot des différentes valeures projetées sur le premier axe#

boxplot(acmsev,col=brewer.pal(6, "Set1"))

#Distribution des variables sur le premier axe#

score(acmsev,xax=1)

#Représentation par classes des données projetées#

scatter(acmsev, col = brewer.pal(6, "Set1"))


#Représentation par classes de la variable de réponse et des résultats de la biopsie#
```

```
117
118 par(mfrow=c(1,2))
119 s.class(acmsev$li, Severite, col = brewer.pal(6, "Set1") )
120 s.class(acmsev$li, Anomalie, col = brewer.pal(6, "Set1") )
121
122 ##########################BONUS#####################################
123
124 #Essai sur d'autre types de visualisations#
125
126 s.potatoe <- function(dfxy, fac, xax = 1, yax = 2, col.border =
    ↪  rep(1, length(levels(fac))), col.fill = rep(1,
    ↪  length(levels(fac))), shape = -0.5, open = "FALSE", ...) {
127         dfxy <- data.frame(dfxy)
128         opar <- par(mar = par("mar"))
129         par(mar = c(0.1, 0.1, 0.1, 0.1))
130         on.exit(par(opar))
131         x <- dfxy[, xax]
132         y <- dfxy[, yax]
133         for (f in levels(fac)) {
134                 xx <- x[fac == f]
135                 yy <- y[fac == f]
136                 xc <- chull(xx, yy)
137                 qui <- which(levels(fac) == f)
138                 border <- col.border[qui]
139                 col <- col.fill[qui]
140                 xspline(xx[xc], yy[xc], shape = shape, open =
                    ↪  open, border = border, col = col, ...)
141         }
142 }
143
144 shape=1
145
146 col.border <- c("red", "blue")
147 col.fill <- c(rgb(1, 0, 0, 0.3), rgb(0, 0, 1, 0.3))
148
149
150 bkg <- function(...) s.class(acmsev$li, Severite, inc = FALSE,
    ↪  cell = 0, adda = FALSE, cst = 0, col = col.border, clab =
    ↪  1.5, ...)
151
152
153 bkg(sub = paste("shape =", round(1, 2)), csub = 2.5)
154 s.potatoe(acmsev$li, Severite, col.border = col.border, shape =
    ↪  shape, col.fill = col.fill)
155
156 ##############################################################
```

```r

shape=1

col.border <- c("red", "green", "blue","orange")
col.fill <- c(rgb(1, 0, 0, 0.3), rgb(0, 1, 0, 0.3), rgb(0, 0, 1, 0.3),rgb(1,0.7,0.3 ,0.3))


bkg <- function(...) s.class(acmsev$li, Anomalie, inc = FALSE, cell = 0, adda = FALSE, cst = 0, col = col.border, clab = 1.5, ...)


bkg(sub = paste("shape =", round(1, 2)), csub = 2.5)
s.potatoe(acmsev$li, Anomalie, col.border = col.border, shape = shape, col.fill = col.fill)
```

Code - Exercie 2

```r
 ################ Exercice 2 #################
library("ade4")
library("factoextra")
library("FactoMineR")
library(MASS)
library(mvtnorm)
library(mda)
library(ggplot2)

setwd("~/Documents/Master_Data_Science/Analyse Factorielle/projet")
table=read.csv('Stone_flakes.txt',header=TRUE,sep="",na.strings='?')

groups = c('Homo Ergaster','Technique Levallois','Homme de Néandertal','Homo Sapiens')

table[,10][table[,10]==1] = groups[1]
table[,10][table[,10]==2] = groups[2]
table[,10][table[,10]==3] = groups[3]
table[,10][table[,10]==4] = groups[4]

for (i in 1:16){
  table[,i] = as.factor(table[,i])
}
```

```
table = table[rowSums(is.na(table[,2:9])) == 0, ]
table = na.omit(table)

# Decision Tree
library(rpart)
tree <- rpart(group ~ LBI + RTI + WDI + FLA + PSF + FSF + ZDF1 +
↪  PROZD,
              data=table,minsplit=10)
plot(tree,margin=.1)
text(tree, use.n = TRUE,cex=.8)

# Random Forest
library(randomForest)
forest <- randomForest(group ~ LBI + RTI + WDI + FLA + PSF + FSF
↪  + ZDF1 + PROZD,
                       data=table,
                       importance=TRUE,
                       nodesize=10)
varImpPlot(forest)

# Boxplot for each numeric variable
library("gridExtra")
theme_array =  theme(axis.title.x=element_blank(),
                axis.ticks.x=element_blank(),
                axis.text.x=element_blank(),
                legend.position = 'None')
p_ZDF1 <- qplot(group, ZDF1, data=table, geom=c("boxplot",
↪  "jitter"), fill=group) + theme_array
p_FLA <- qplot(group, FLA, data=table, geom=c("boxplot",
↪  "jitter"), fill=group) + theme_array
p_PROZD <- qplot(group, PROZD, data=table, geom=c("boxplot",
↪  "jitter"), fill=group) + theme_array
p_PSF <- qplot(group, PSF, data=table, geom=c("boxplot",
↪  "jitter"), fill=group) + theme_array
p_LBI <- qplot(group, LBI, data=table, geom=c("boxplot",
↪  "jitter"), fill=group) + theme_array
p_WDI <- qplot(group, WDI, data=table, geom=c("boxplot",
↪  "jitter"), fill=group) + theme_array
p_RTI <- qplot(group, RTI, data=table, geom=c("boxplot",
↪  "jitter"), fill=group) + theme_array
p_FSF <- qplot(group, FSF, data=table, geom=c("boxplot",
↪  "jitter"), fill=group) + theme_array

vplayout <- function(x, y) viewport(layout.pos.row = x,
↪  layout.pos.col = y)

```

```
# Plot
grid.newpage()
pushViewport(viewport(layout = grid.layout(4, 2)))
print(p_ZDF1, vp = vplayout(1, 1))
print(p_FLA, vp = vplayout(1, 2))
print(p_PROZD, vp = vplayout(2, 1))
print(p_PSF, vp = vplayout(2, 2))
print(p_LBI, vp = vplayout(3, 1))
print(p_WDI, vp = vplayout(3, 2))
print(p_RTI, vp = vplayout(4, 1))
print(p_FSF, vp = vplayout(4, 2))

qplot(age, ZDF1, data=table, colour=group) + theme_array

# PCA
res.pca = PCA(table, scale.unit=TRUE, quali.sup=c(1,10:16))

# Explained Variance
fviz_screeplot(res.pca)

# Coordinates of variables
names_quant_var = c('Length-breadth index'
                    ,'Relative-thickness index'
                    ,'Width-depth index'
                    ,'Flaking angle'
                    ,'platform primery'
                    ,'Platform facetted'
                    ,'Dorsal surface totally worked'
                    ,'Proportion of worked dorsal surface')

res.pca$var$coord

fviz_pca_var(res.pca, col.var="cos2") +
  scale_color_gradient2(low="white", mid="blue",
                        high="red", midpoint=0.5) +
                         ↪  theme_minimal()
plot(res.pca,  choix="ind", habillage=9)

var_type = c('group', 'age', 'dating', 'mat', 'region', 'site')
for (i in c(9:15)){
  i = 10
  fviz_pca_biplot(res.pca,
                  habillage=i, addEllipses = TRUE,
                   ↪  ellipse.level=0.75,
                  col.var = "black", alpha.var ="cos2",
```

```
                  label = "var", pointsize = 2.5, ggtheme =
                   ↪  theme_minimal())
}
# Contributions of variables on PC1
fviz_contrib(res.pca, choice = "var", axes = 4)
fviz_pca_ind(res.pca,col.ind="cos2")
fviz_pca_ind(res.pca, habillage=10)
fviz_pca_biplot(res.pca,
                habillage=i+6, addEllipses = FALSE,
                 ↪  ellipse.level=0.75,
                col.var = "black", alpha.var ="cos2",
                label = "var", pointsize = 3, ggtheme =
                 ↪  theme_minimal(), axes=c(1,2))

qplot(age, mat, data=table, colour=mat, geom="histogram") +
 ↪  theme_array
hist_cut <- ggplot(table, aes(x=age, fill=mat))

hist_cut <- ggplot(table, aes(x=age, fill=mat)) + geom_bar()
hist_cut + geom_bar(position="fill")

qplot(age, mat, data=table, geom=c("boxplot", "jitter"),
 ↪  fill=group) + theme_array


########
# Does the data reflect the technological progress during
 ↪  several hundred thousand years?
table_mda = na.omit(table[,2:10])
mda_out <- mda(group ~ ., data = table_mda)
lda_out <- lda(group ~ ., data = table_mda)

# contours
mda_predict <- predict(mda_out, table_mda)
lda_predict <- predict(lda_out, table_mda)

table(lda_predict$class, table_mda$group)

rownames(table) = table[,1]

## multiplot
# Multiple plot function
#
# ggplot objects can be passed in ..., or to plotlist (as a list
 ↪  of ggplot objects)
# - cols:   Number of columns in layout
```

```
# - layout: A matrix specifying the layout. If present, 'cols' is ignored.
#
# If the layout is something like matrix(c(1,2,3,3), nrow=2, byrow=TRUE),
# then plot 1 will go in the upper left, 2 will go in the upper right, and
# 3 will go all the way across the bottom.
#
multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
  require(grid)

  # Make a list from the ... arguments and plotlist
  plots <- c(list(...), plotlist)

  numPlots = length(plots)

  # If layout is NULL, then use 'cols' to determine layout
  if (is.null(layout)) {
    # Make the panel
    # ncol: Number of columns of plots
    # nrow: Number of rows needed, calculated from # of cols
    layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                     ncol = cols, nrow = ceiling(numPlots/cols))
  }

  if (numPlots==1) {
    print(plots[[1]])

  } else {
    # Set up the page
    grid.newpage()
    pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))

    # Make each plot, in the correct location
    for (i in 1:numPlots) {
      # Get the i,j matrix positions of the regions that contain this subplot
      matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

      print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
```

```
                                        layout.pos.col =
                                        ↪  matchidx$col))
    }
  }
}
```

Code - Exercie 3 - SAS

```
proc import datafile="/folders/myshortcuts/SAS/Analyse
↪  Factorielle/TP3/parkinsons_updrs.data" out = parkinson_1
DBMS=CSV;
delimiter=",";
run;

proc cancorr data=parkinson_1 out=acc red;
        var motor_UPDRS total_UPDRS;
        with age sex test_time Jitter___ Jitter_Abs_ Jitter_RAP
        ↪  Jitter_PPQ5 Jitter_DDP Shimmer Shimmer_dB_
        ↪  Shimmer_APQ3 Shimmer_APQ5 Shimmer_APQ11 Shimmer_DDA
        ↪  NHR HNR RPDE DFA PPE;
run;

proc print data=acc_1; run;

proc plot data=acc_1; run;
proc v1*v2$subject#;
title "Les deux premières variables canoniques pour X"; run;

proc plot data=sortie1; run;
proc w1*w2$subject#;
title "Les deux premières variables canoniques pour Y"; run;
```

Code - Exercie 4

```
 #Packages#
library(MASS)
library(ade4)
library(RColorBrewer)
library(corrplot)


#Importation du jeu de donnees#
```

```
table=read.table('~/Desktop/wine.data',header=FALSE,sep=",")
colnames(table)=c("Identifiant", "Alcool", "Acide_malique",
"Cendre", "Alcalinite", "Magnesium", "Phenols",
"Flavanoides", "Non_flavanoides", "Proanthocyanidines",
"Couleur", "Teinte", "OD", "Proline")
table$Identifiant=as.factor(table$Identifiant)

############################

#On représente la matrice des corrélations#
corr=cor(table[,-1])
corrplot.mixed(corr)

#On réalise une ACP pour étudier les corélation entre variables#
acp=dudi.pca(table[,-1],scannf=FALSE,nf=2)
res <- scatter(acp, clab.row = 0, posieig = "none")
s.class(acp$li, fac = table$Identifiant, col = brewer.pal(3,
"Set1"), add.plot = TRUE, cstar = 0, cellipse = 0)

############################

#Réalisation d'un AFD sur les variables avec comme réponse le
vigneron#
afd=lda(table$Identifiant ~ .,data=table)
plot(afd,col=as.integer(table$Identifiant),pch=20)

#Réalisation de piechart pour les moyennes de chaque vignerons
selon chaque attributs#
par(mfrow=c(4,4))
apply(afd$means, 2,  function(x) pie(x,col=brewer.pal(3,
"Set1")))

#Réalisation de barchat pour les moyennes de chaque vignerons
selon chaque attributs#
mat=as.matrix(afd$means)
par(mfrow=c(4,4))
for(i in 1:13){
        barplot(mat[,i],col=brewer.pal(3,
        "Set1"),sub=colnames(mat)[i])
}
```