Compte-rendu de projet en Machine Learning  
Parkinsons Data Set



# Résumé

1. Motivation et positionnement du projet
   1. Objectif de notre étude
   2. Pourquoi c’est important ?
2. Analyse descriptive
3. Classification non-supervisée
4. Classification supervisée
5. une autre piste : : ACP et régression logistique
6. Epilogue

# Citations obligatoires

Lien dataset : <https://archive.ics.uci.edu/ml/datasets/Parkinsons>

Lien Github : <https://github.com/Lizotope/Parkinson_R.git>

Ce dataset a été créé par Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig (2008), pour les ouvrages :

* 'Suitability of dysphonia measurements for telemonitoring of Parkinson's disease', IEEE Transactions on Biomedical Engineering : "<http://www.maxlittle.net/publications/TBME-00342-HTML.html>"
* 'Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection', Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. BioMedical Engineering OnLine 2007, 6:23 (26 June 2007)

# Motivation et positionnement du projet

## Objectif de notre étude

Repérer les dysphonies qui caractérisent la maladie de Parkinson, via une mesure de la dysphonie introduite par les auteurs même -> la PPE (Pitch Period Entropy). Évaluer si la PPE pourrait constituer un nouveau facteur prédictif de la maladie.

Cette mesure PPE serait peu sensible aux variations de phonies issues des perturbations extérieures et des variations classiques de la voix.

Le chercheur s'est concentré sur des enregistrements de phonation soutenue seulement. Les patients utilisent un micro standard. Un certain nombre de mesures permettent de relever les propriétés des signaux, et parmi celles qu’il est d’usage de retenir, nous pouvons mentionner fréquence fondamentale, shimmer, jitter, bruit. L’auteur a réalisé d'autres mesures moins anodines telles que la DFA, et RPDE (voir tableau récapitulatif du paragraphe 2).

## Pourquoi c’est important ?

L'auteur explique qu'il existe plus d'un million de cas rien que dans l'Amérique du Nord, que la prévalence de la maladie augmente significativement à partir de l'âge de 60 ans et qu'aucun traitement curatif n'existe pour le moment.

Néanmoins il explique que l'efficacité des traitements visant le soulagement des symptôme est d'autant plus efficace dans les stades précoces de la maladie [1] ([1] : [4]. Singh N, Pillay V, Choonara YE. Advances in the treatment of Parkinson's disease. Progr Neurobiol 2007;81:29–44.).

Pour en savoir plus : <https://www.ffn-neurologie.fr/grand-public/maladies/maladie-de-parkinson>.

Une détection précoce permettrait :

* une optimisation de la stratégie thérapeutique (traitant la symptomatologie et au mieux pouvant ralentir l'évolution irrémédiable de la maladie).
* une adaptation anticipée des choix de vie (travail, vies familiale et sociale) du patient.

Peu de facteurs prédictifs sont aujourd'hui connus, malgré le grand nombre d'études à ce propos.

Ex de facteurs éventuellement prédictifs en cours d'étude : marqueurs biologiques, imageries...

L'essor de la télémédecine et le prix des visites médicales pour le patient, outre-atlantique, a motivé le chercheur à se focaliser sur le développement des outils de détection à distance.

La mesure des dysphonies des patients pourrait être l'un d'eux. Elle présenterait notamment l'avantage d'être un outil non invasif, simple à administrer.

----------------------------------------------

Issu de l'abstract de la publication dans le magazine Nature : *"in this paper, we present an assessment of the practical value of existing traditional and nonstandard measures for discriminating healthy people from people with Parkinson's disease (PD) by detecting dysphonia. We introduce a new measure of dysphonia, pitch period entropy (PPE), which is robust to many uncontrollable confounding effects including noisy acoustic environments and normal, healthy variations in voice frequency. We collected sustained phonations from 31 people, 23 with PD. We then selected ten highly uncorrelated measures, and an exhaustive search of all possible combinations of these measures finds four that in combination lead to overall correct classification performance of 91.4%, using a kernel support vector machine. In conclusion, we find that nonstandard methods in combination with traditional harmonics-to-noise ratios are best able to separate healthy from PD subjects. The selected nonstandard methods are robust to many uncontrollable variations in acoustic environment and individual subjects, and are thus well suited to telemonitoring applications."*

# Analyse descriptive

* 1. **Initialisation**

Avant de nous intéresser à aux analyses descriptives et exploratoires, nous initialisons notre étude via le script ***initialisation.R***

* chargement du jeu de données brutes sous une dataframe et des librairies relatives aux fonctions qui seront appelées,
* identification des individus et des attributs, identification de l’attribut qui identifie le signal, renommage des attributs.
* vérification des types de valeurs, de la dimension du jeu de données, visualisation
* stockage du dataframe dans un objet R de type matrix pour faciliter quelques manipulations
* Séparation du dataframe en deux dataframes : un ne contenant que les signaux de patients sains, l’autre de patients malades. Stockage, description.
* Standardisation des données des 3 dataframes, stockage. Ceci nous servira pour une partie de notre analyse.
  1. **Taille du jeu de données**

Le jeu de données porte sur des enregistrements phoniques de 32 patients, dont 3 (n°35, 21, et 27) avec 6 enregistrements, et 29 avec 7 enregistrements. Nous avons donc un jeu de données de 195 “individus” (195 signaux).

* 1. **Choix des attributs par le créateur du jeu de données**

La fréquence fondamentale (F0, mesure globale de la hauteur de la voix aiguë grave), la gigue (jitter, différence de F0 entre deux cycles de vibrations consécutifs), le shimmer (rapports d’amplitudes entre deux cycles de vibration consécutifs), le rapport signal-bruit (rapport entre l’énergie du bruit et l’énergie harmonique du signal) sont classiquement les mesures retenues pour analyser un son vocal.

Mais l’auteur du jeu de données a choisi d’aller plus loin, d’optimiser robustesse des mesures dépendantes de l'environnement, du contexte de l'expérience, du genre du sujet. Il a choisi entre autres d’ajouter le RPDE (recurrence period density entropy) et DFA (detrended fluctuation analysis).

Il a choisi également d’introduire une nouvelle mesure de la dysphonie, Pitch Period Entropy (PPE) qui est peu sensible aux nombreux effets de confusion incontrôlables, tels que les environnements acoustiques bruyants et telles que les variations normales et saines de la fréquence vocale.

Nous comptons un total de 23 attributs.

* 1. **Préparation des attributs**

Nous avons recueilli les descriptions des différents attributs en partie dans le document <https://pdfs.semanticscholar.org/4461/eff4577b9a378d5cbd2df59840066e356e92.pdf>

Nous les avons également regroupées en 8 catégories afin d’en fournir une description brève et accessible.

|  |  |  |
| --- | --- | --- |
| **Vocal Feature** | **Short description** | **Description** |
| Fo (Hz) | Fundamental frequency Fo | “MDVP\_Fo” : Average fundamental frequency of a periodic waveform |
| Fhi (Hz) | “MDVP\_Fhi” : Highest fundamental frequency of a periodic waveform |
| Flow (Hz) | “MDVP\_Flow” : Lowest fundamental frequency of a periodic waveform |
| Jitter (%) | Several measures of variation in fundamental frequency (**Frequency parameters**) | “MDVP\_JitterRel” : Average absolute difference between consecutive periods, divided by the  average period. |
| Jitter  (Abs) | “MDVP\_JitterAbs” : Average absolute difference between consecutive periods which gives information about the cycle-to-cycle variation of fundamental frequency given in seconds. |
| Jitter:  RAP | “MDVP\_Rap” : Relative Average Perturbation (RAP), which is the average absolute difference between a period and the average of it and its two neighbours, divided by the average period. |
| Jitter:  PPQ5 | “MDVP\_PPQ” : Five-point Period Perturbation Quotient, computed as the average absolute difference between a period and the average of it and its four closest neighbours, divided by the average period. |
| Jitter:  DDP | “Jitter\_DDP” : Average absolute difference between consecutive differences between consecutive periods, divided by the average period |
| Shimmer | Several measures of variation in amplitude (**Amplitude parameters**) | “MDVP\_Shimmer” : Average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude. |
| Shimmer  (dB) | MDVP\_ShimmerDB” : Average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20. It gives information about the variability of the peak-to-peak amplitude in decibels. |
| Shimmer:  APQ3 | “Shimmer\_APQ3” : 3-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbours, divided by the average amplitude. |
| Shimmer:  APQ5 | “Shimmer\_APQ5” : 5-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its four closest neighbours, divided by the average amplitude. |
| Shimmer:  APQ11 | “MDVP\_APQ” : 11-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its ten closest neighbours, divided by the average amplitude. |
| Shimmer:  DDA | “Shimmer\_DDA” : Average absolute difference between consecutive differences between the amplitudes of consecutive periods. |
| Noise to Harmonics Ratio (NHR) | 2 measures of ratio of noise to tonal components in the voice | “NHR”: Amplitude of noise relative to tonal components. It quantifies the noise which occurs due to turbulent airflow, resulting from incomplete vocal fold closure in speech pathologies. |
| Harmonics to Noise Ratio (HNR) | “HNR”: Amplitude of tonal relative to noise components. It has the same aim as NHR |
| Recurrence period density entropy  (RPDE) | 2 Nonlinear dynamical complexity measure | “RPDE” and “D2” : Addresses the ability of the vocal folds to sustain stable vocal fold vibrations,quantifying the deviations from exact periodicity |
| D2 |
| Detrended fluctuation analysis  (DFA) | Signal fractal scaling exponent | “DFA”: Quantifies the self-similarity of the noise present in the speech caused by the turbulent air flow |
| Pitch period entropy  (PPE) | 3 Nonlinear measures of fundamental frequency variation | “PPE”, “spread1”, “spread2” : Measures the impaired control of stable pitch during sustained phonations |
| spread1 |
| spread2 |
| status |  | “status”: Health status of the subject (one) - Parkinson's, (zero) - healthy |

Un attribut (celui que l’on cherche à prédire) est binaire : le statut. Il vaut 1 pour les signaux correspondants à des patients malades, 0 sinon. Les 22 autres attributs sont numériques.

* 1. **Valeurs manquantes**

La fonction *describe* nous permet de constater qu’aucune donnée ne manque.

* 1. **Caractéristiques de position et valeurs remarquables**

La fonction *describe* indiquent 10-quantiles, moyenne et nombre de valeurs manquantes (cf 2.5), mais la fonction *summary* permet d’être plus succinct tout en donnant les valeurs maximale et minimale de chaque attributs.

* + 1. **Etendue et valeurs minimales/maximales**

Tous les attributs sont globalement étalés entre -1 et 1, avec des valeurs proches à zéro, à l’exception de spread1 et D2 qui sont d’un ordre de grandeur 10 fois plus grand (entre -10 et 10), ainsi que les mesures de fréquences fondamentales.

Les fréquences fondamentales des voix humaines oscillent généralement entre 75 à 140 hertz (Hz) pour les hommes, 170 à 250 Hz pour les femmes. En observant le minimum de F0 min (la plus petite valeur observée parmi tous les signaux, attribut “MDVP\_Flow”) à savoir 65Hz, le maximum de F0 max (la plus grande valeur observée parmi tous les signaux) à savoir 592Hz, les moyennes de F0 min (116Hz) et F0 max(197Hz), les valeurs relevées à l’expérience semblent cohérentes.

Enfin, l’attribut DFA mesure l’auto-similarité des signaux (voir “analyse des fluctuations redressées”). Les valeurs sont comprises entre 0,5 et 1, ce qui témoignerait de “corrélations persistantes à long terme”, ce qui, de notre compréhension, signifie que l’évolution du signal tend à suivre des tendances (donc que le signal tend à avoir un comportement fractal).

* + 1. **Distribution des 23 attributs**
* **Attribut statut :**

Nous comptons 147 signaux “malades”, et 48 “sains”. Voir figure 262-1

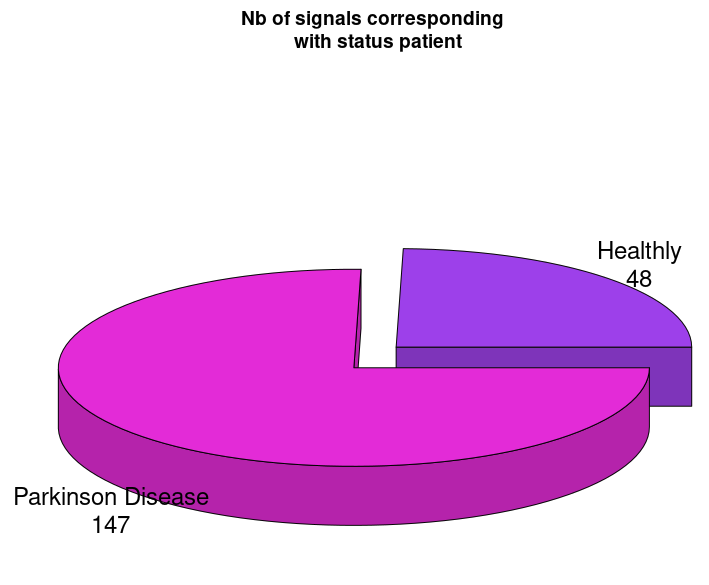


Figure 262-1

Pour tous les 22 autres attributs, nous avons affiché la boîte de Turkey pour repérer les valeurs “exceptionnelles”. Nous avons complété par un affichage “en ligne de crêtes” (violplot) pour mieux visualiser la distribution.

* **Attributs relatifs à la fréquence fondamentale :**

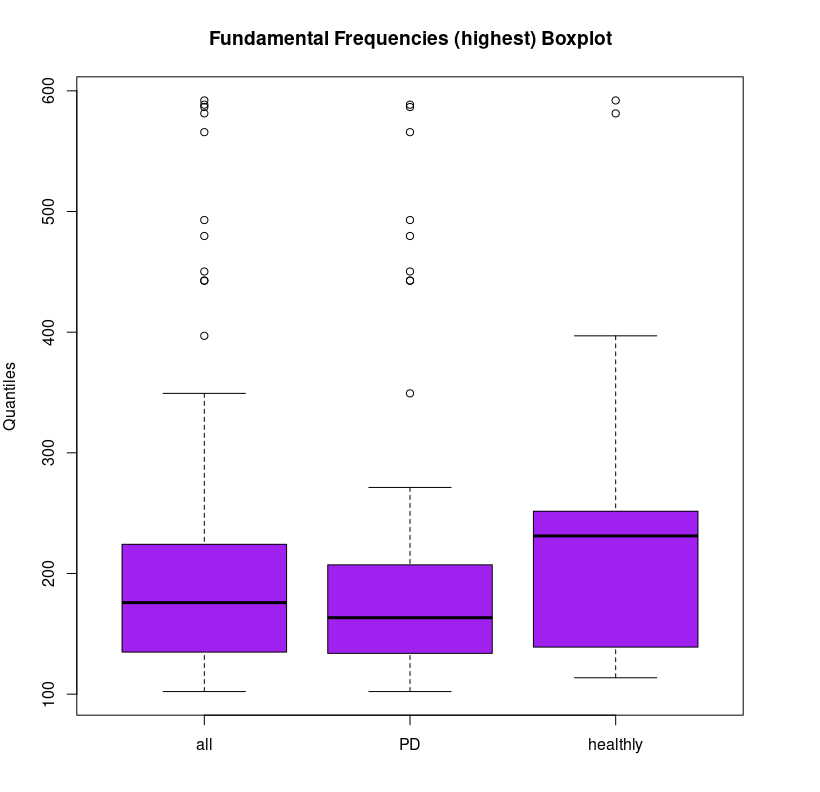
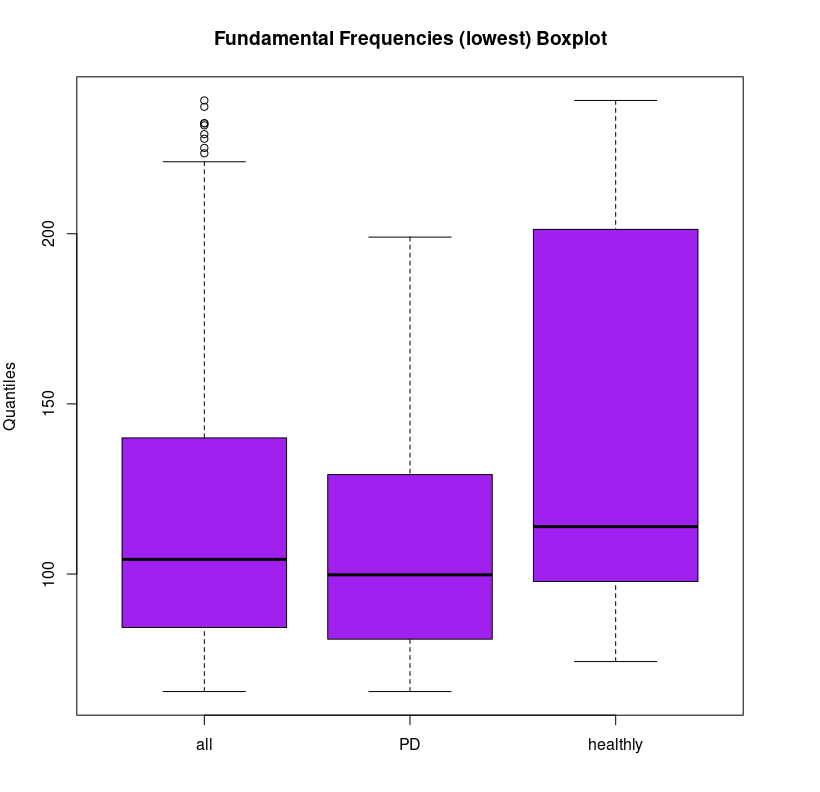


Figure 262-4 Figure 262-3

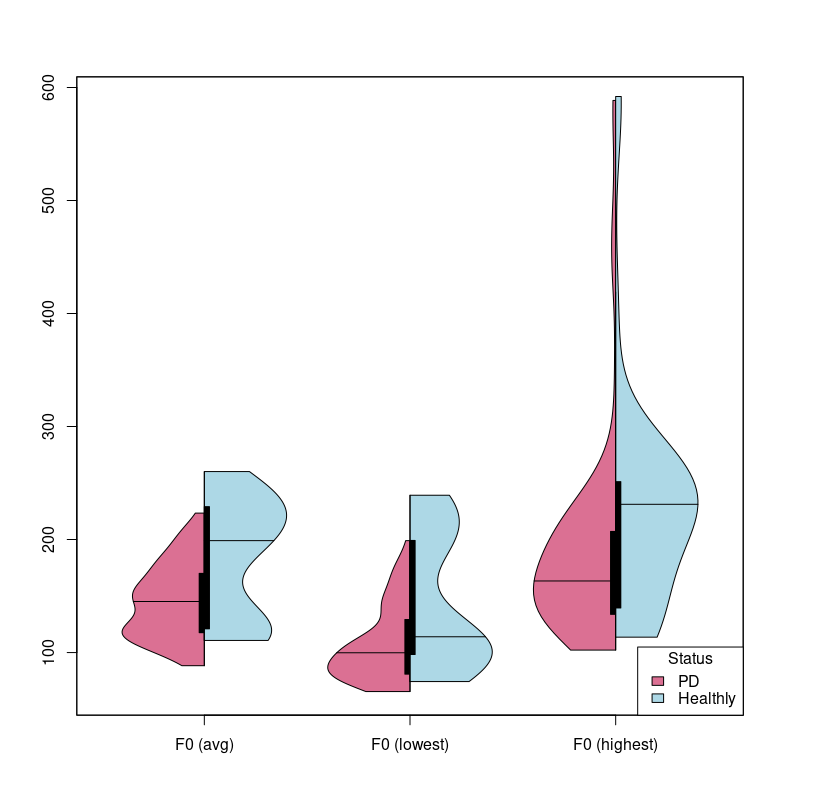
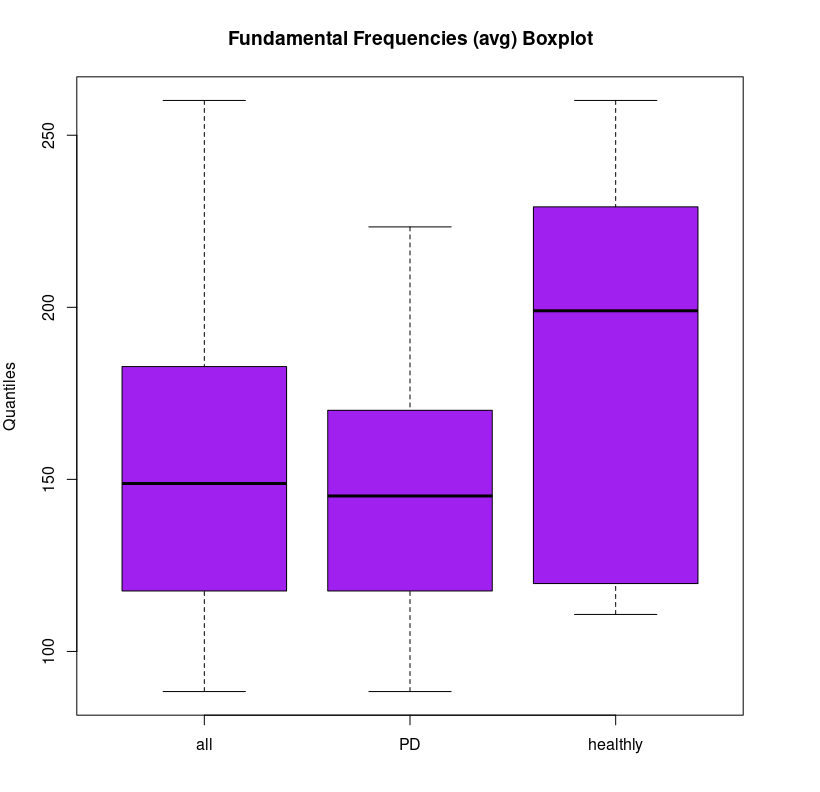


Figure 262-2 Figure 262-5

On constate nettement que les signaux “sains” montent jusqu’à des fréquences plus hautes que celles des signaux “malades”, les figures 262-2 et 262-5 nous confirment qu’il ne s’agit pas de cas “sains” isolés, bien que ces derniers ne représentent qu’environ un quart des signaux. Les autres caractéristiques (distribution des fréquences basses/moyennes/hautes, médiane des fréquences les plus basses) ne démontrent pas de franche distinction entre le caractère malade/pas malade.

Questionnement : une incapacité à atteindre une fréquence fondamentale haute signe-t-elle la maladie ?

* **Attributs relatifs à la gigue (“jitter”) :**

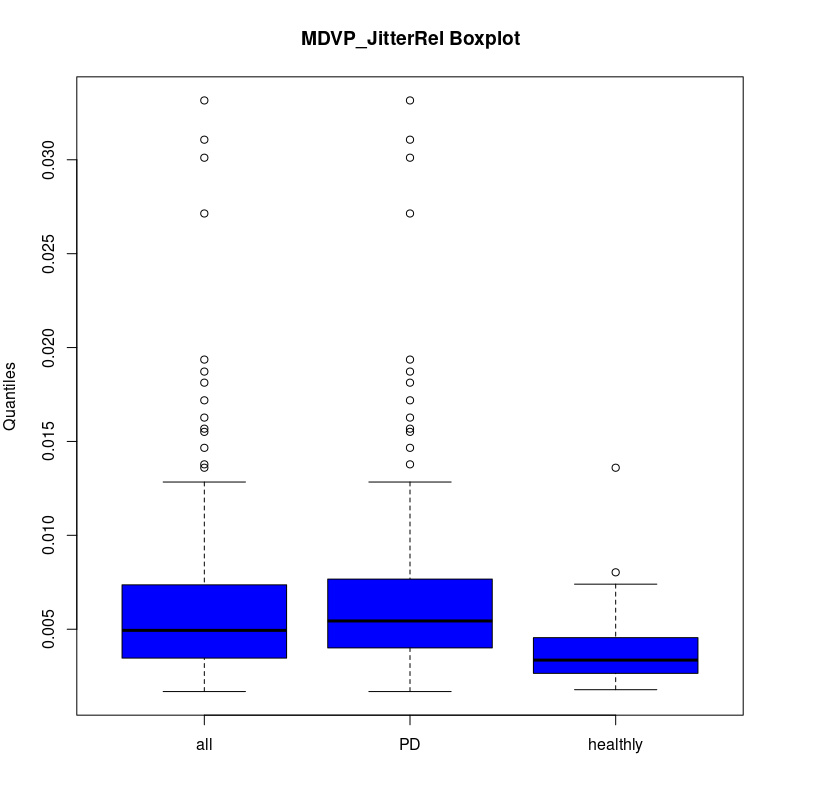
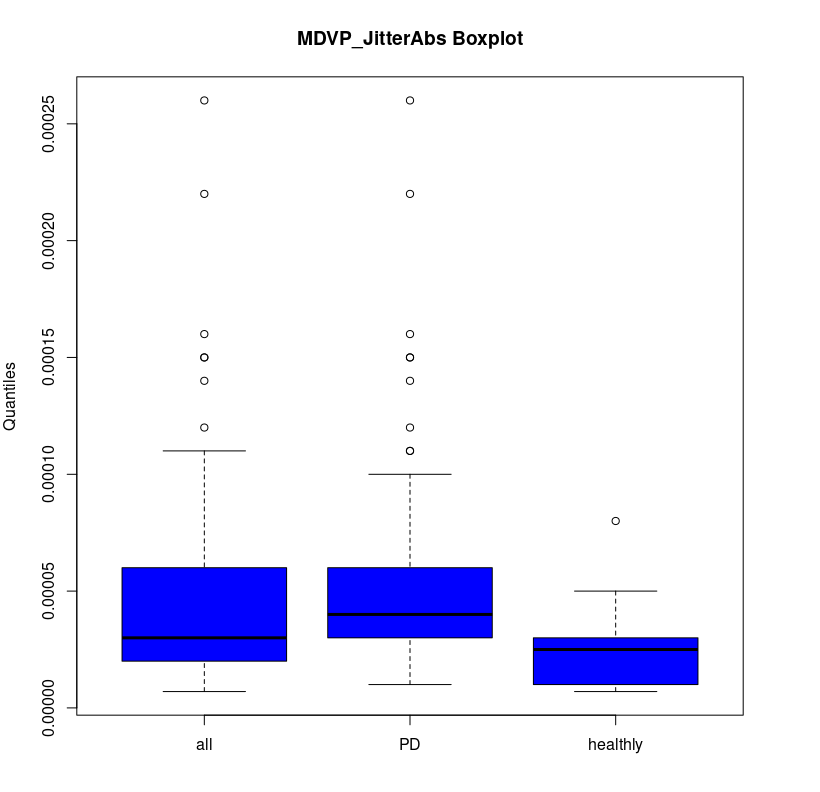


Figure 262-6 Figure 262-7

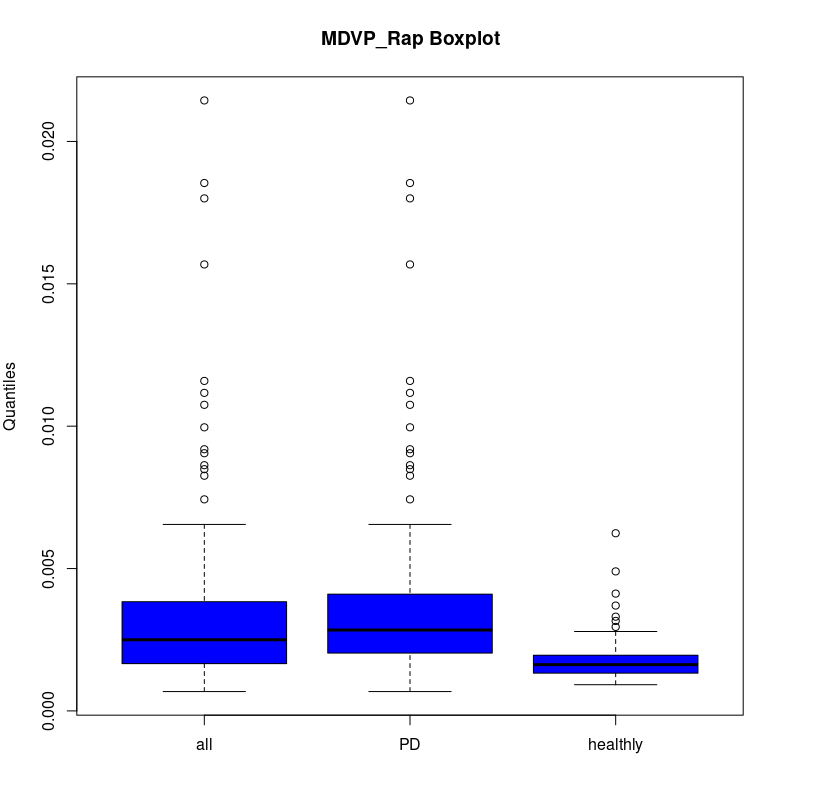
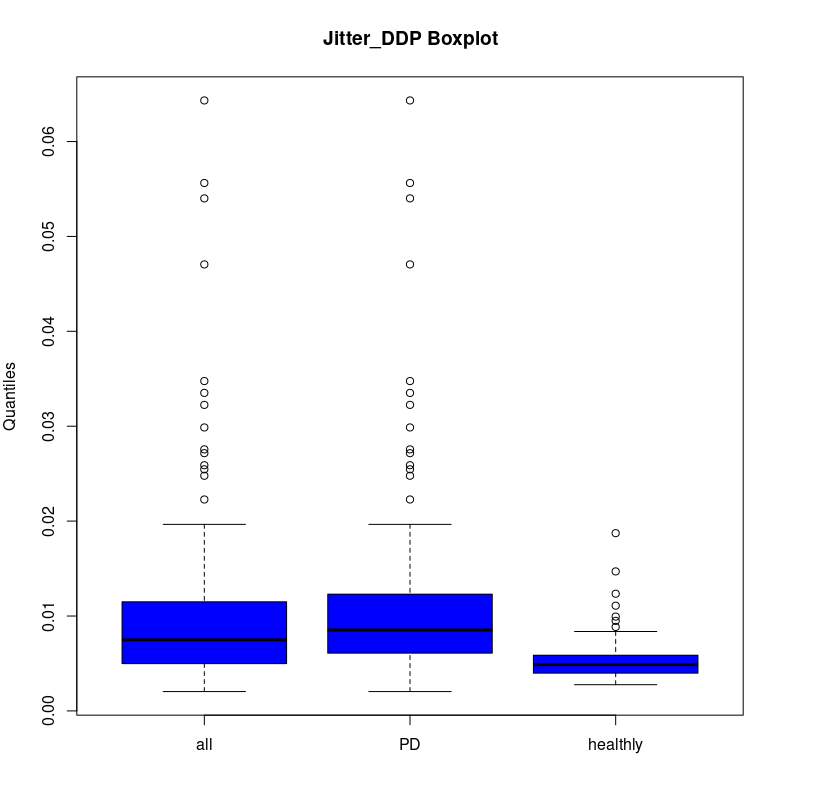


Figure 262-8 Figure 262-9

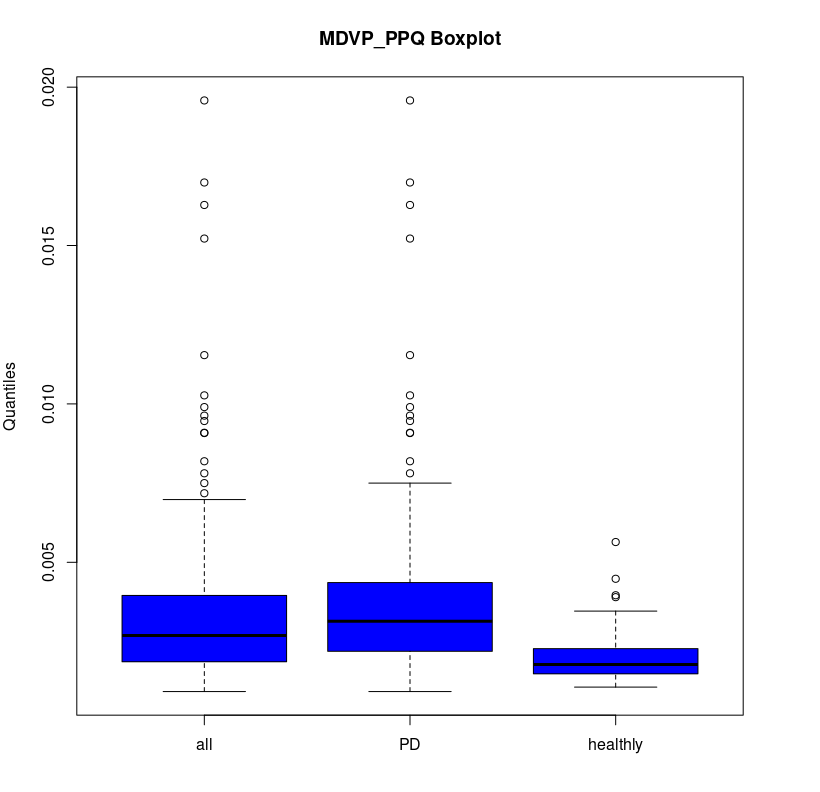


Figure 262-10

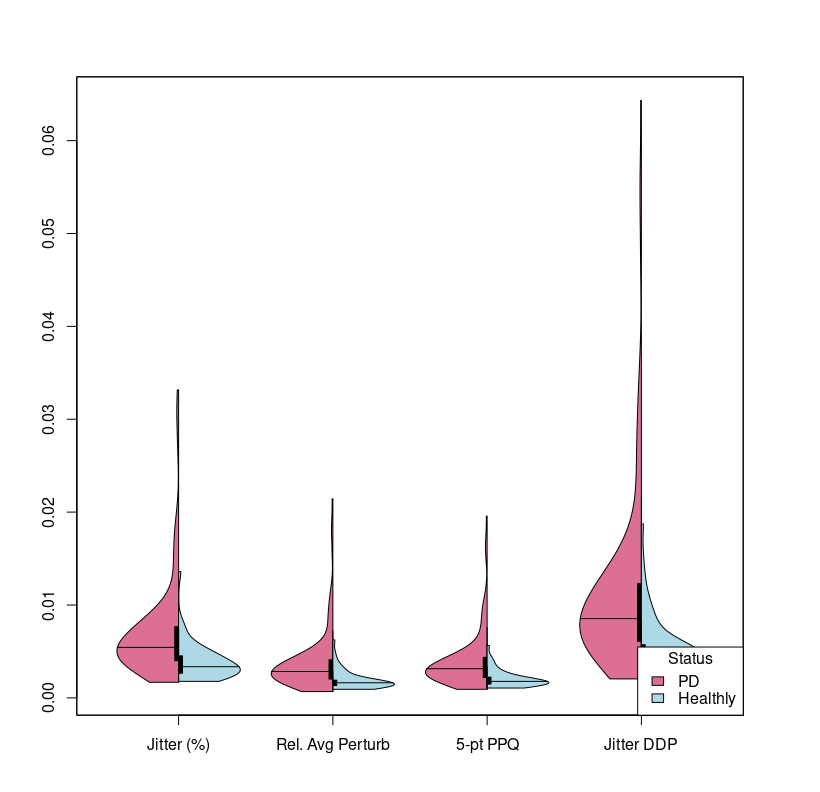
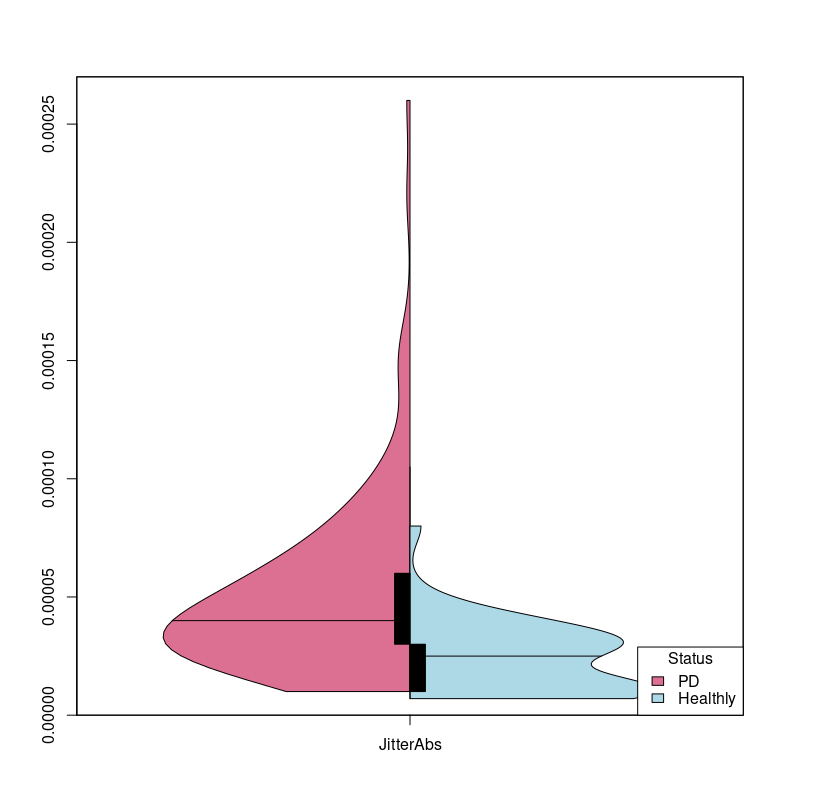
 

Figure 262-11 Figure 262-12

La gigue est une mesure des perturbations de la fréquence fondamentale du signal.

Il ne semblerait pas qu’il y ait davantage de valeurs exceptionnelles (en dehors des 1ers et 4ème quartiles) chez les malades que les non malades compte tenu de la proportion des malades parmi tous les patients. Nous pouvons aussi constater que toutes les valeurs de la gigue se situe très proche de zéro, et tendent même à se tasser vers zéro. Néanmoins il semblerait que les gigues soient plus basses et plus denses autour d’une médiane très faible pour les signaux “sains”.

Questionnement : une gigue plus élevée signe-t-elle la maladie ?

* **Attributs relatifs au shimmer :**

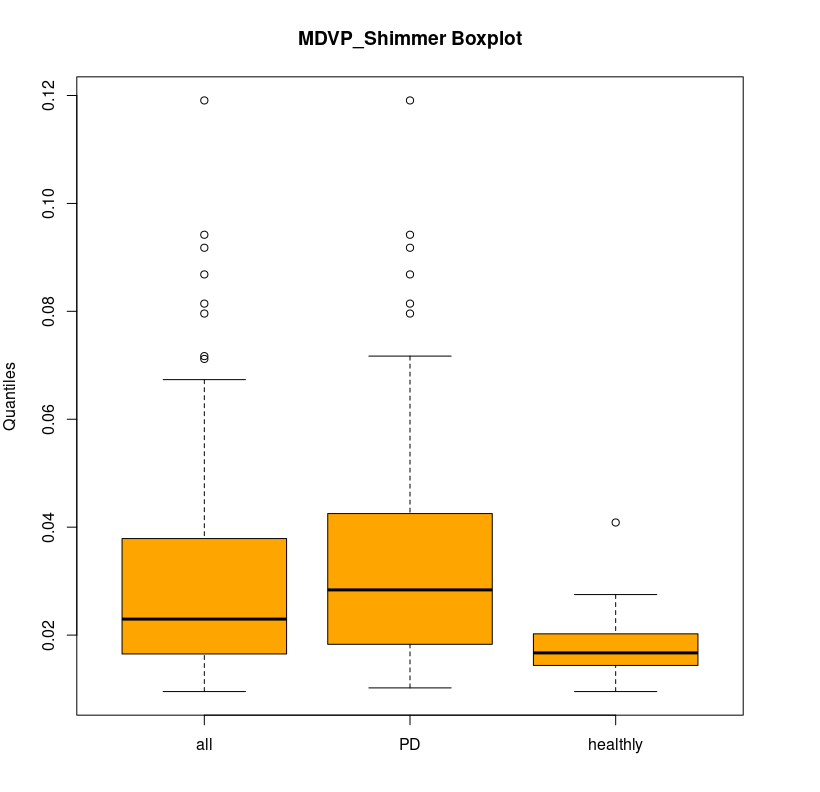
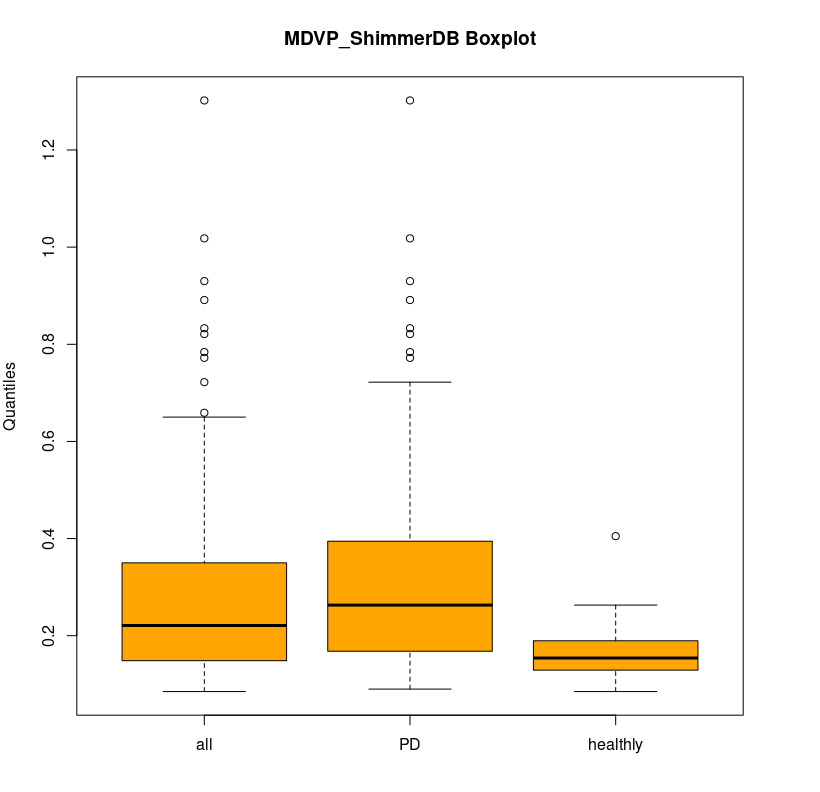


Figure 262-13 Figure 262-14

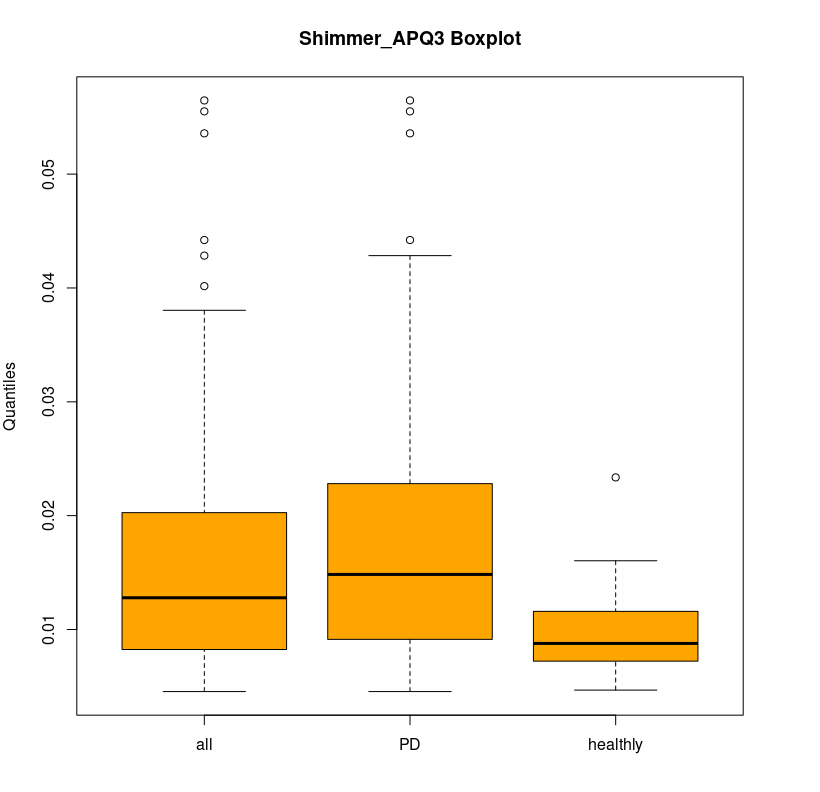
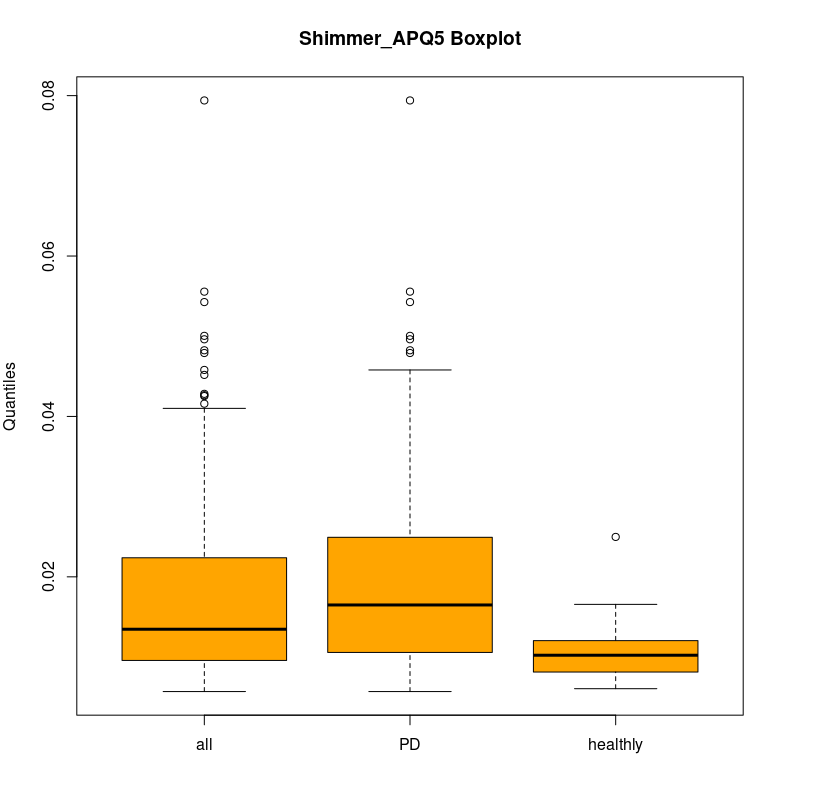
 

Figure 262-15 Figure 262-16

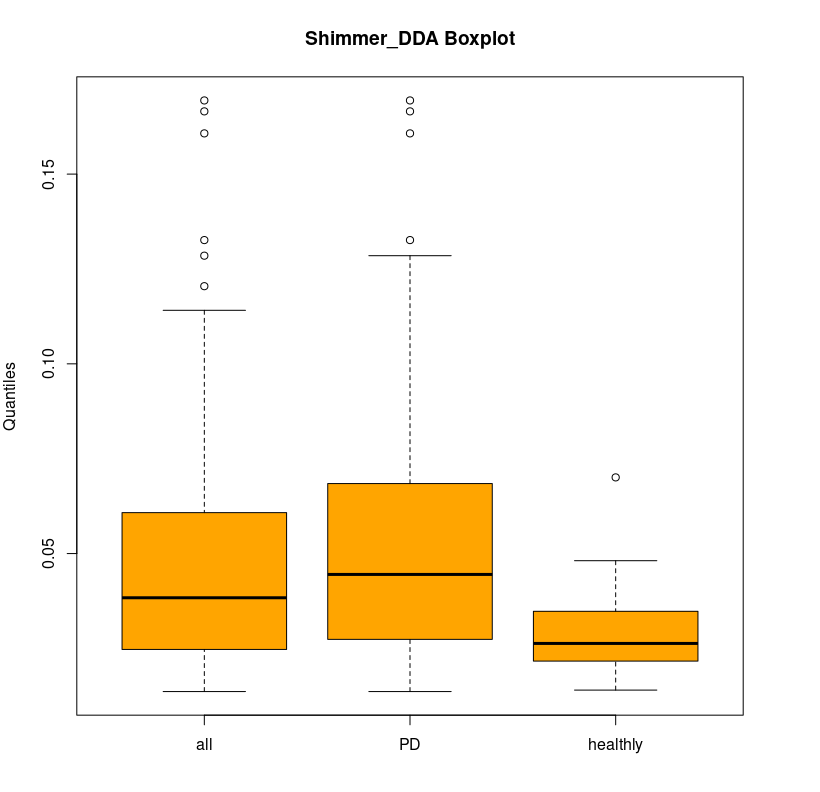
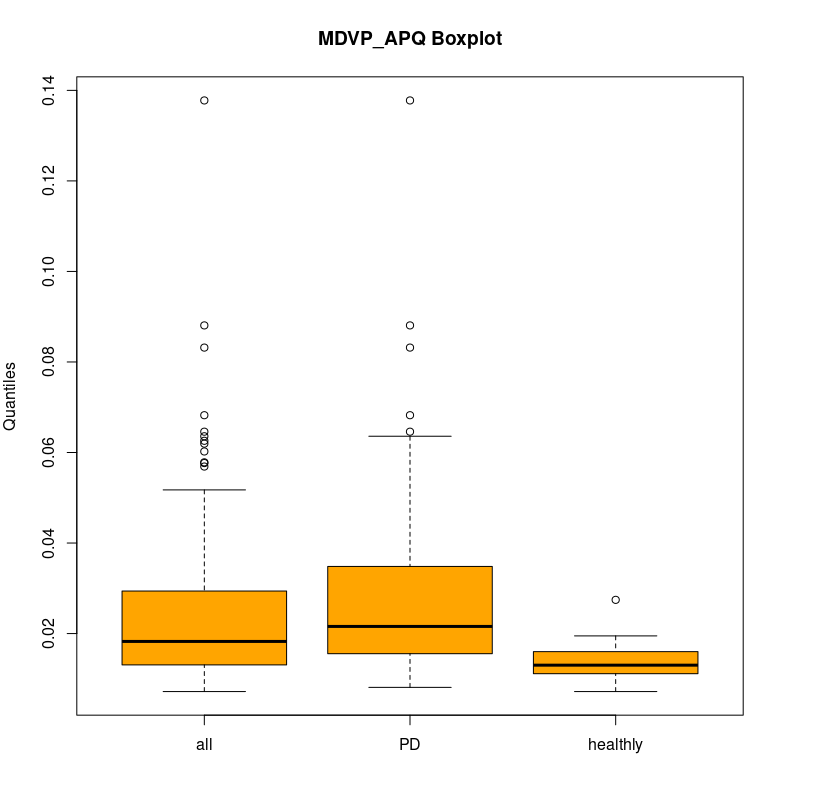


Figure 262-17 Figure 262-18

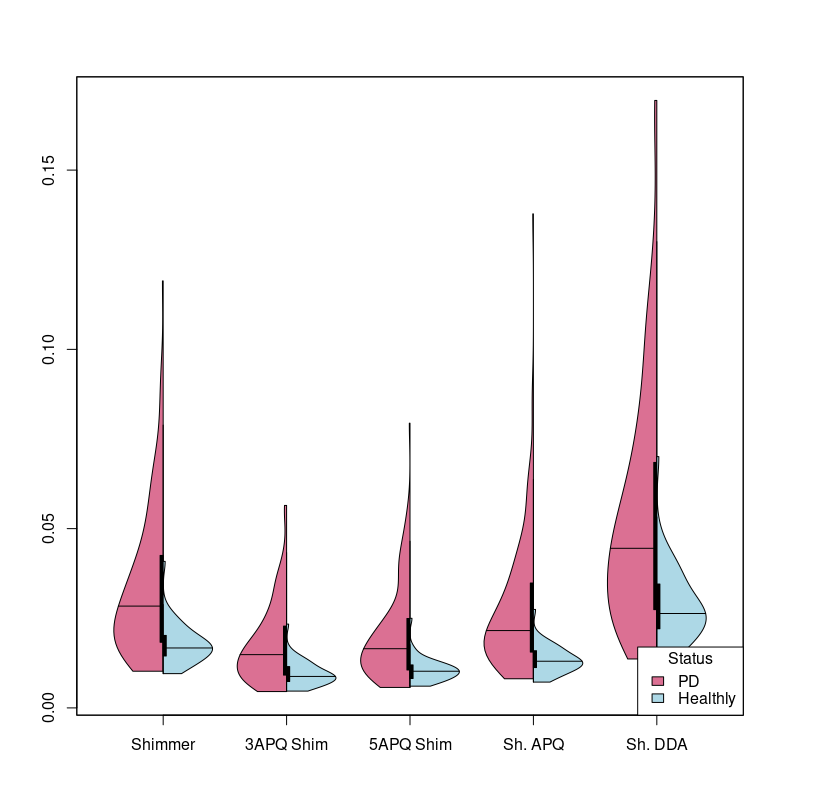
 Le shimmer correspond à une variation d’amplitude entre périodes successives quand une personne tente de tenir la phonation à une fréquence et une intensité constante. Les valeurs obtenues sont très faibles à nouveau, ce dans les 2 camps. Néanmoins les distributions semblent légèrement plus resserrées pour les signaux sains, et d’un shimmer plus faible.

Figure 262-19

Questionnement : les patients malades ont-ils plus de difficulté à maintenir une stabilité dans l’amplitude de leur voix ?

* **Attributs relatifs au bruit :**

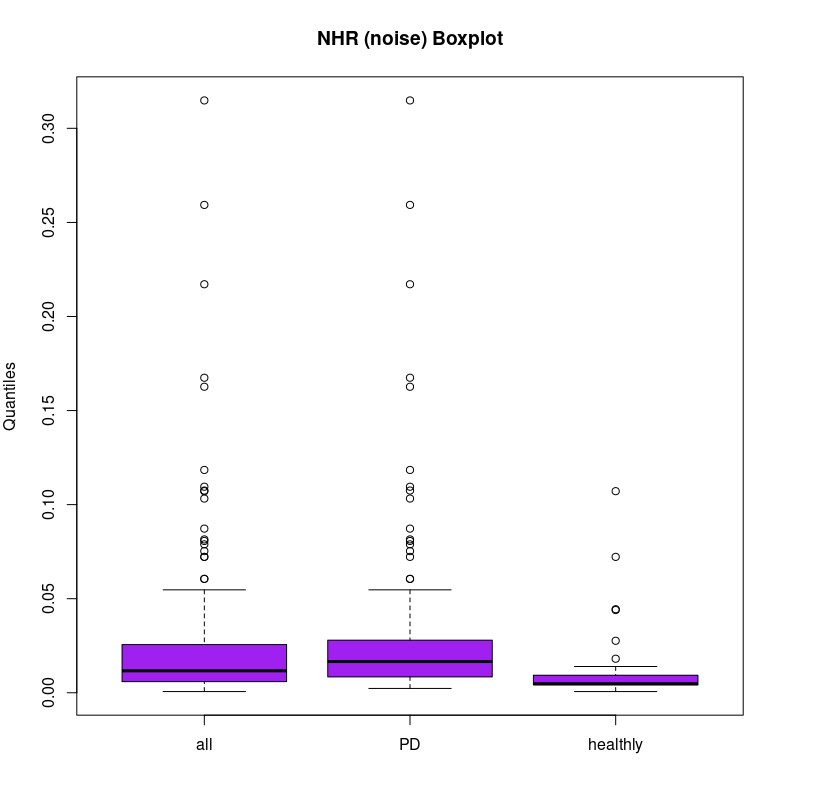
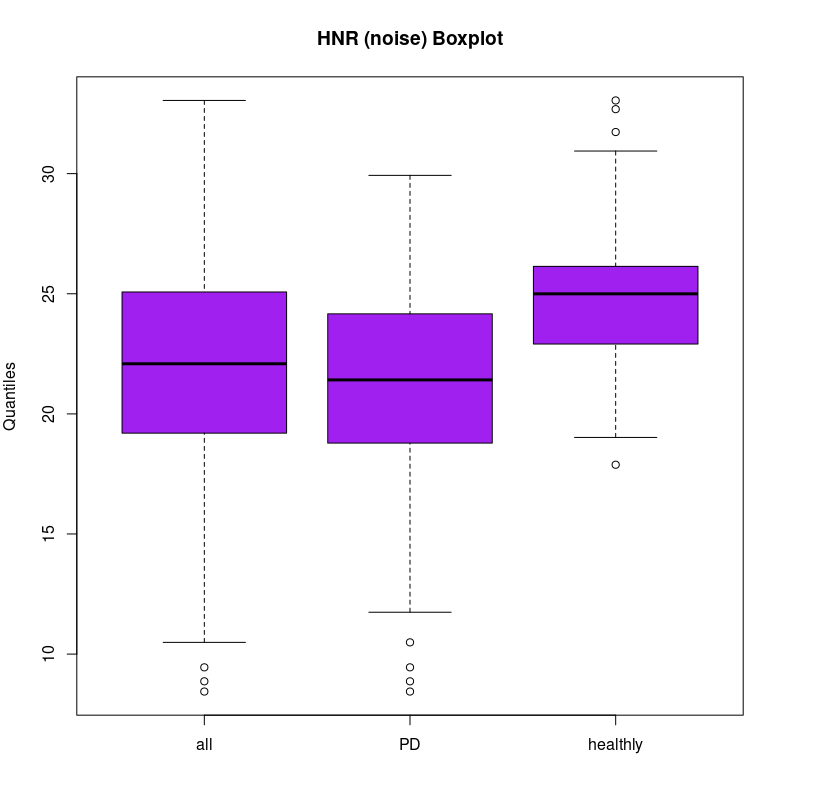
 

Figure 262-20 Figure 262-21

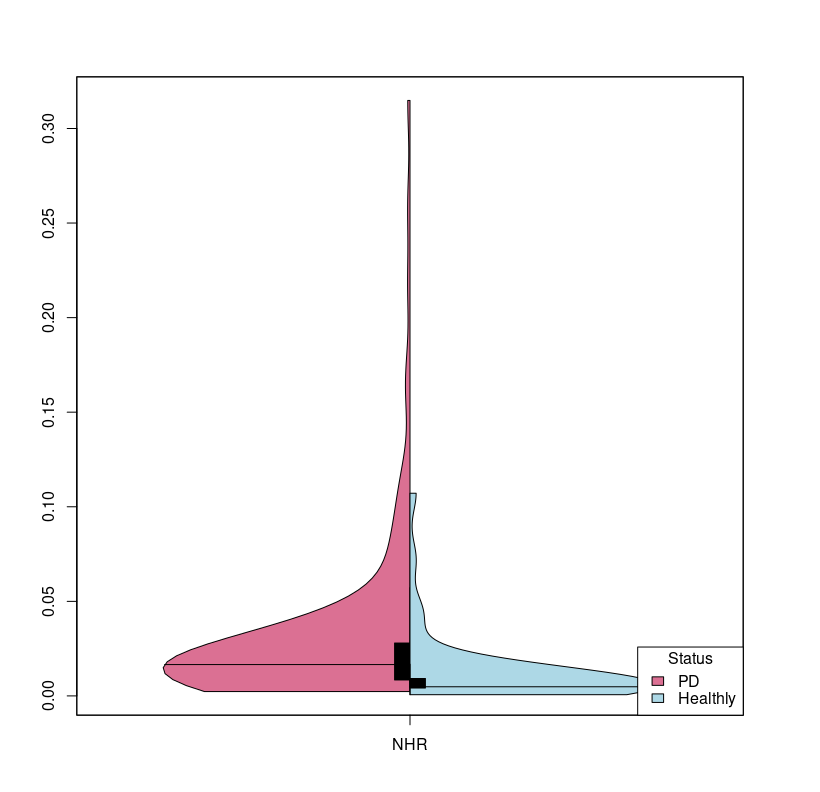
 Il ne semble pas y avoir de différence qui pourrait être franchement imputable au statut de la maladie. Au regard de l’effectif des deux populations, il ne semble pas que les valeurs sortant des 1ers et 3è quantiles soient particulier l’affaire d’une population ou d’une autre. Néanmoins il pourrait y avoir moins une étendue de bruit très légèrement plus faible pour les patients sains. La littérature dit que la fermeture incomplète des cordes vocales dans les pathologies de la parole serait responsable de bruit dans le signal de la voix.

Figure 262-22

Questionnement : la maladie de Parkinson impacte-t-elle suffisamment la fermeture des cordes vocales pour qu’un bruit puisse y être perceptible lors de l’enregistrement de la voix ?

* **Attributs quantifiant les déviations de la périodicité exacte: “RPDE” et “D2” (“mesures de complexité dynamique non linéaires”) :**

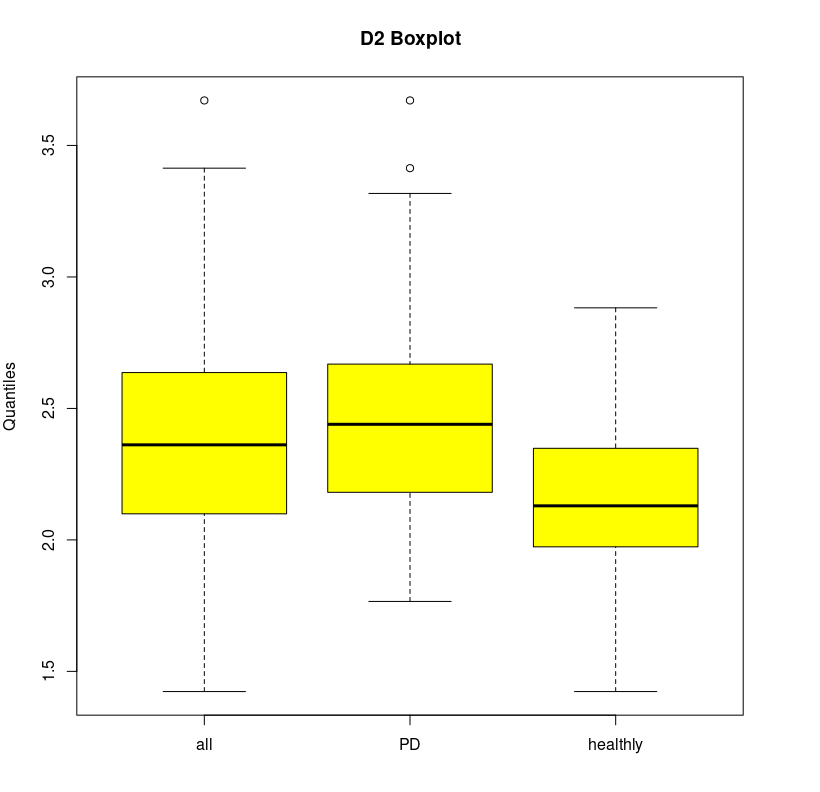
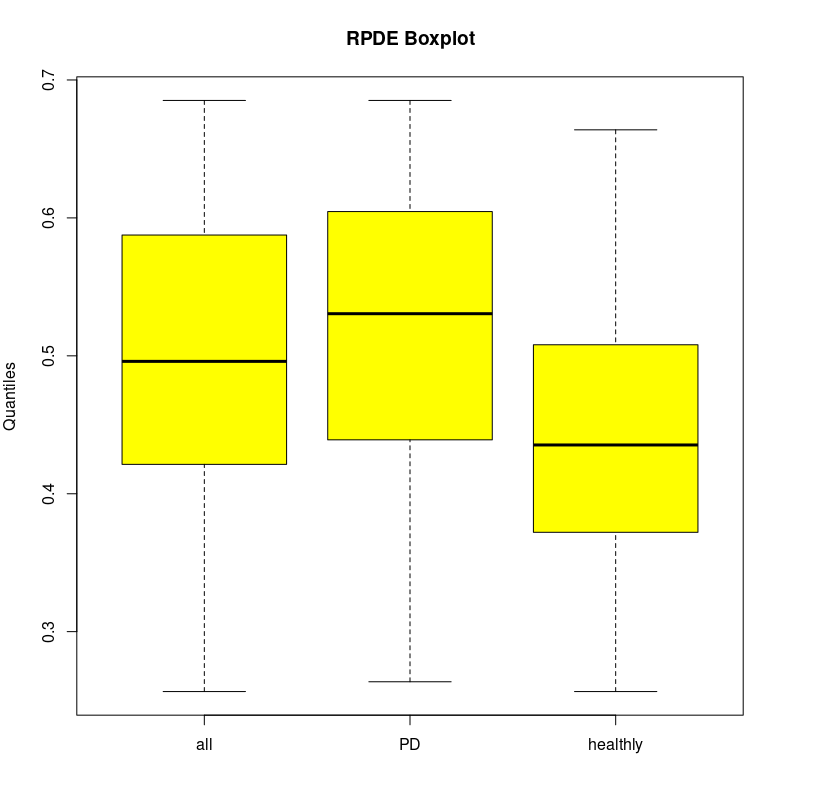
 

Figure 262-23 Figure 262-24

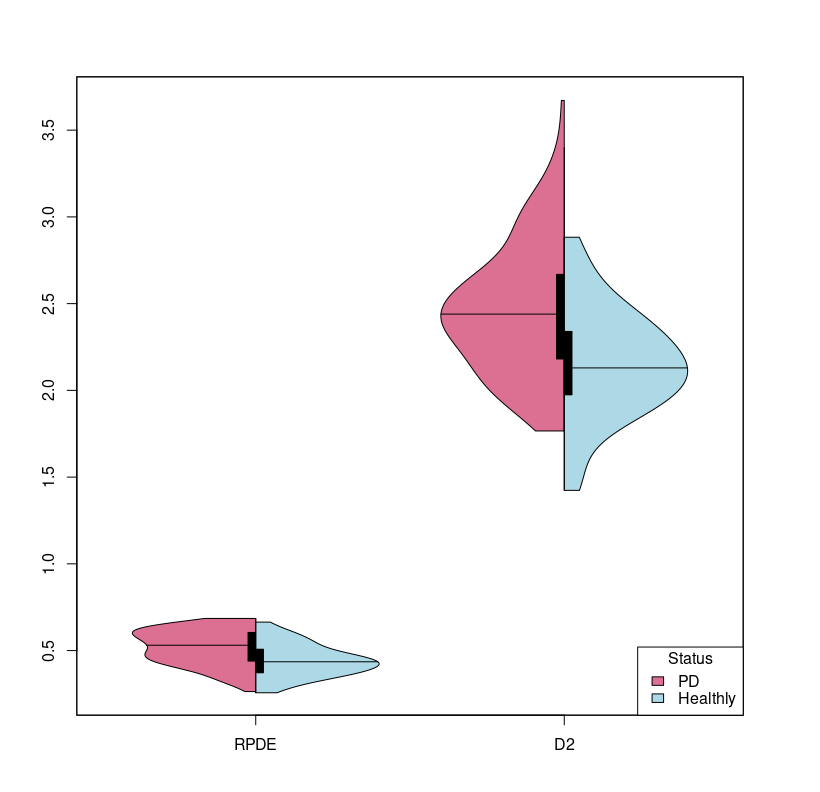
Ces deux mesures s’intéressent également à la capacité des cordes vocales à maintenir des vibrations stables des cordes vocales en quantifiant les écarts par rapport à la périodicité exacte. La différence entre les deux populations n’est ici pas notable. Nous pouvons percevoir l’attribut D2 légèrement abaissée, ce qui amène à un questionnement similaire à l’étude du shimmer.

Figure 262-25

Questionnement : les malades ont-ils une plus faible capacité à maintenir la stabilité de période ?

* **Attributs mesurant l'altération du contrôle de la hauteur stable lors de phonations soutenues (“mesures non linéaires de variation de fréquence fondamentale”) :**

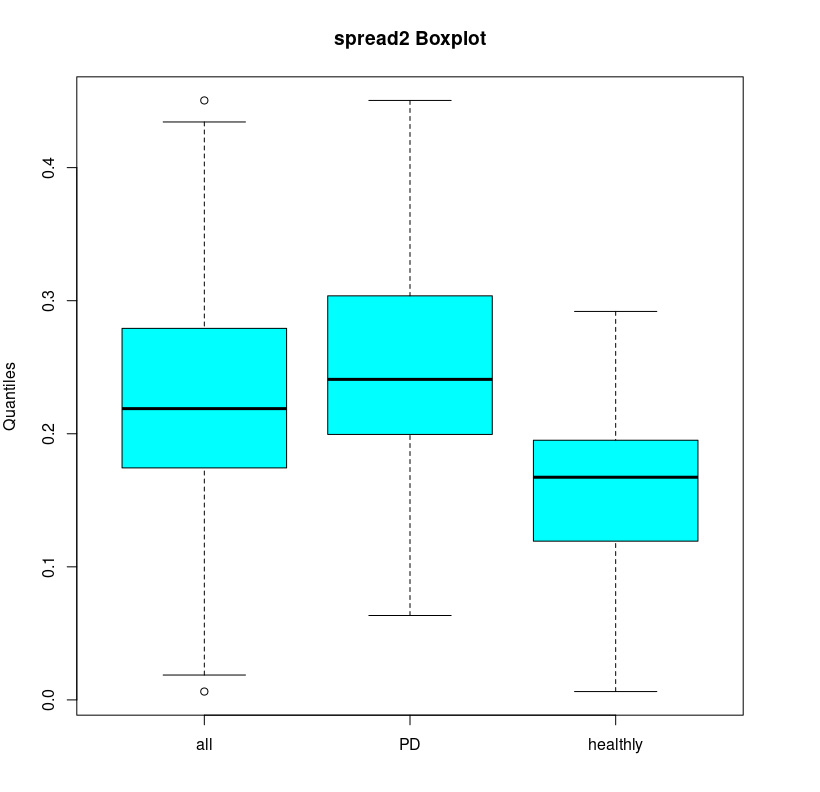
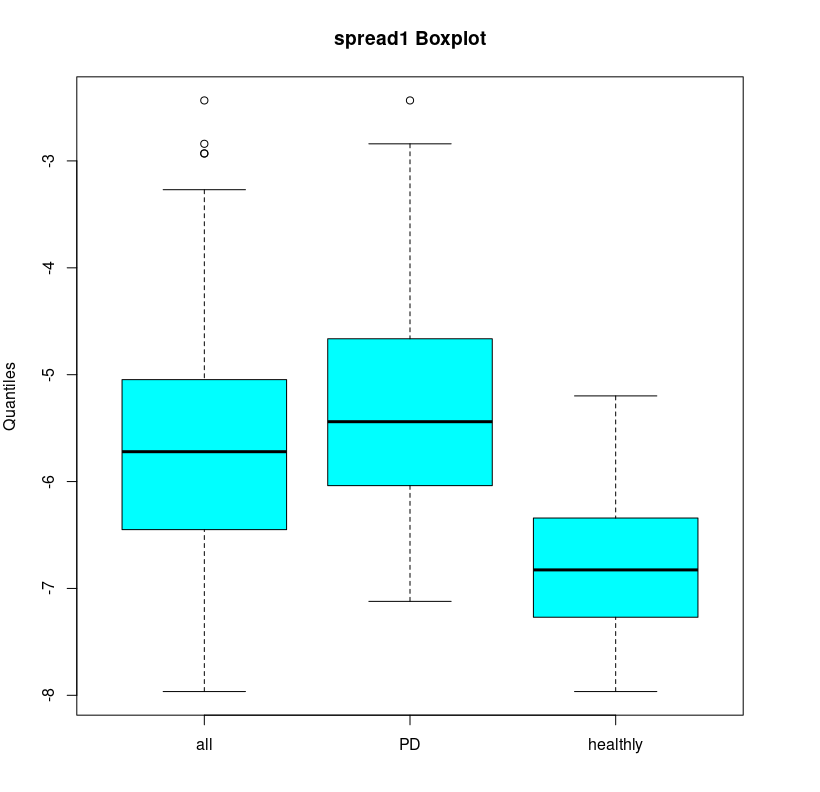


Figure 262-26 Figure 262-27

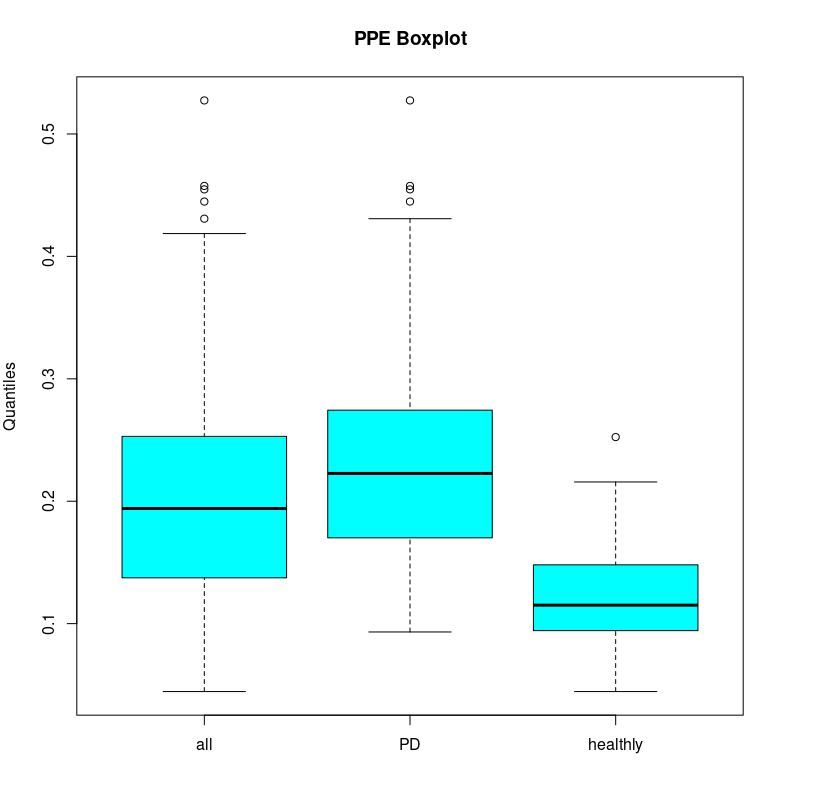
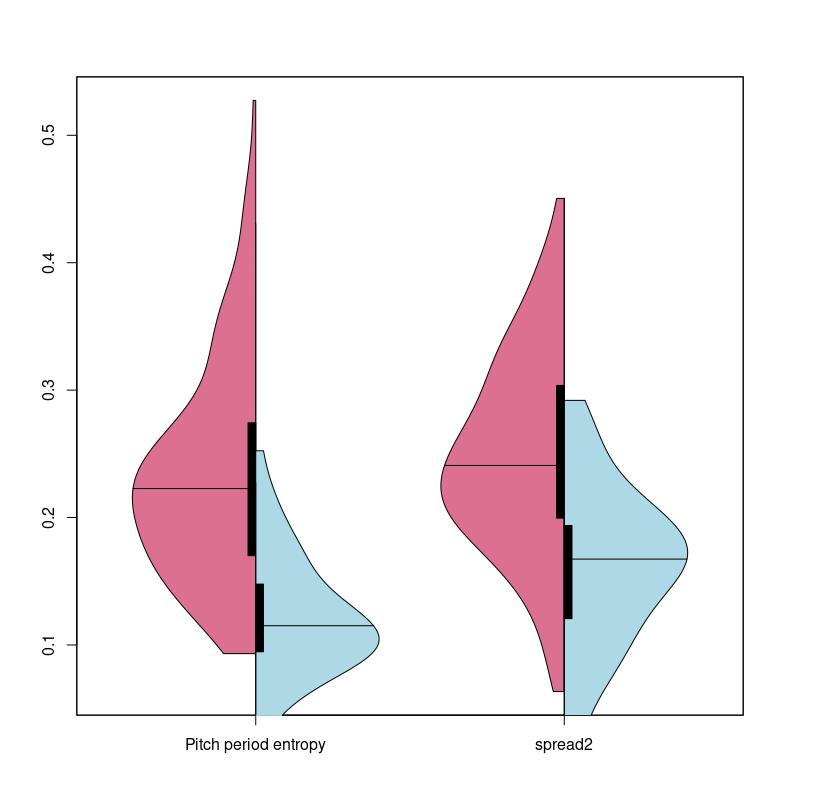
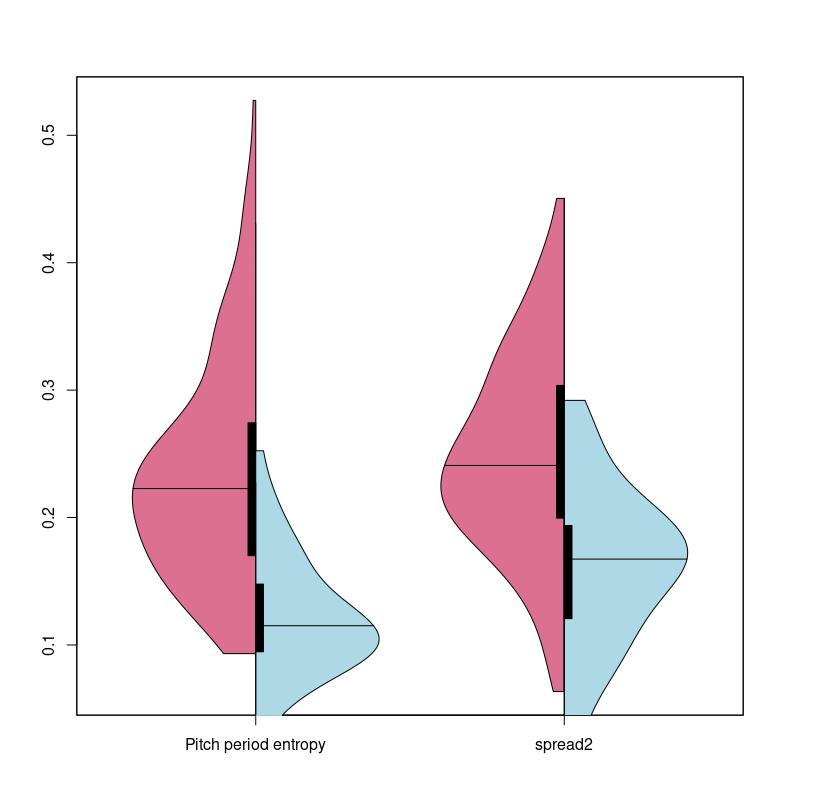
 

Figure 262-28 Figure 262-29



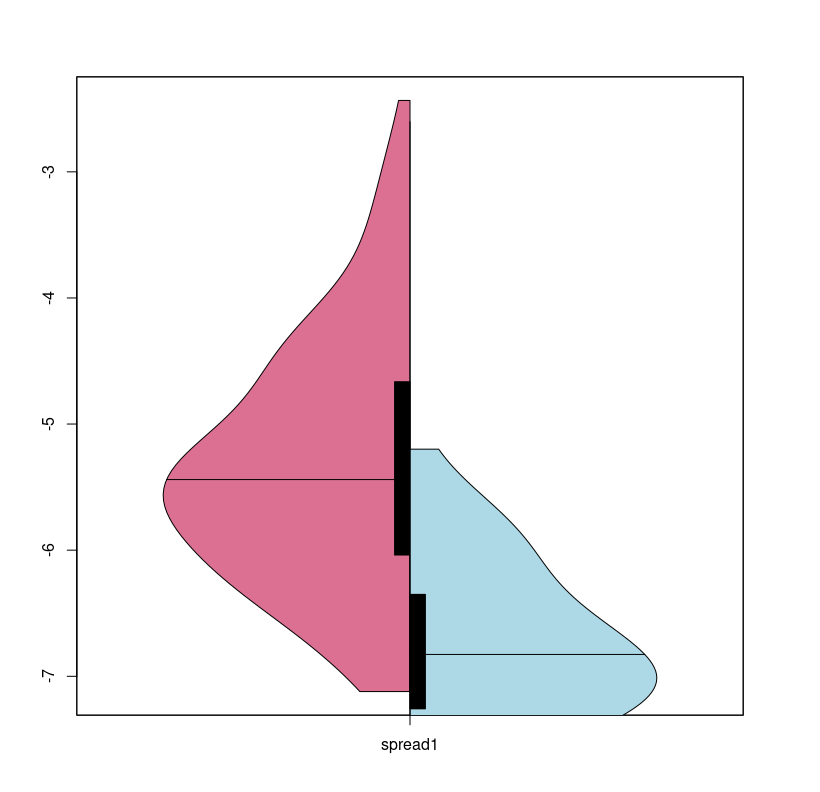
PPE (Pitch period Entropy ), spread1 et spread2 mesurent l'altération du contrôle de la hauteur stable lors de phonations soutenues. Chacune de ces mesures que l’on ne retrouve communément pas lors des analyses vocales semblent pourtant porter des différences significatives entre les deux populations. Ces 3 mesures portant sur le même phénomène, il semble que cohérent que pour chacune de ces 3 mesures, la population des malades et celle des non malades se distingue nettement. 

Figure 262-30

(la mise à l’échelle automatique du repère semble dysfonctionner)

Questionnement : la Pitch Period Entropy, introduite par l’auteur du jeu de données, sera-t-elle un attribut caractéristique de la maladie ?

* **Analyse des fluctuations redressées (DFA)**

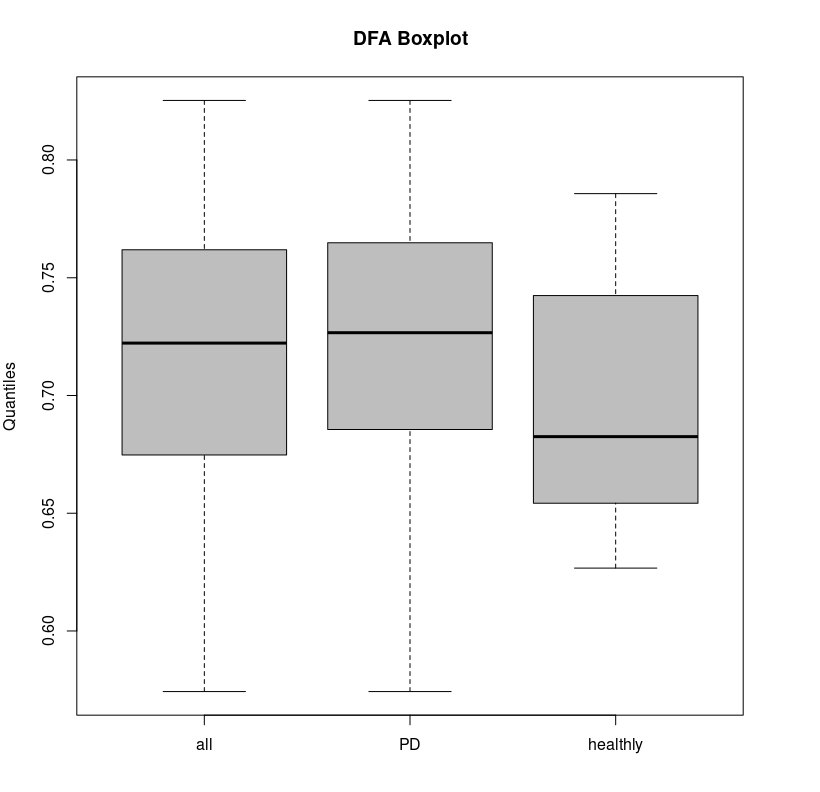
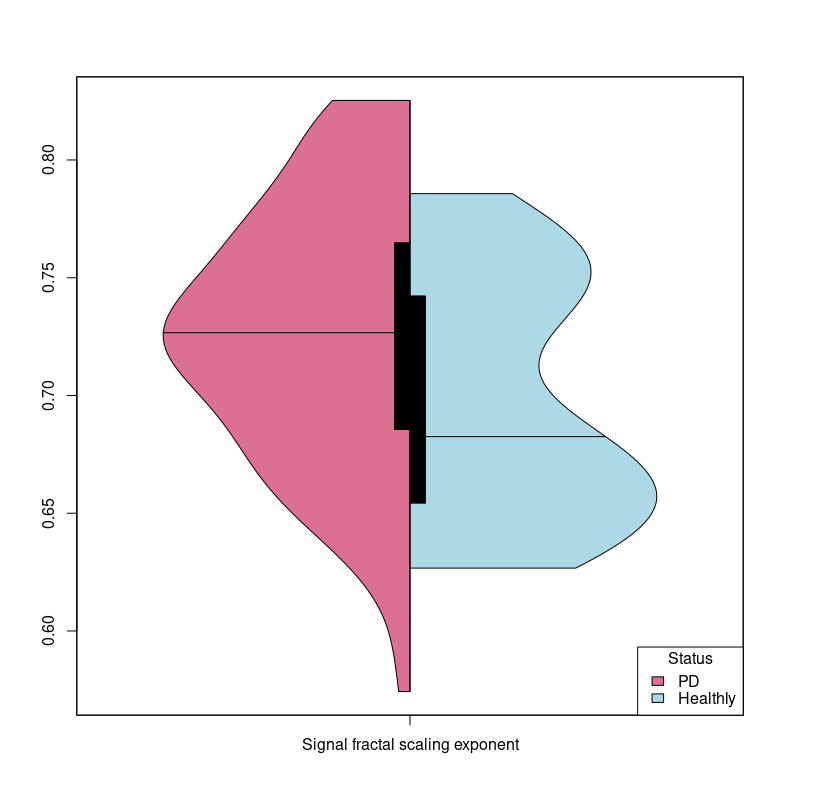
 

Figure 262-31 Figure 262-32 (la mise à l’échelle automatique du repère semble dysfonctionner)

Cette mesure quantifie l’auto-similarité éventuelle du signal. Elle évite de considérer à tort des signaux comme non stationnaires alors qu’ils peuvent suivre une régularité (voir 2.6.1). Il n’y a pas de valeur exceptionnelle pour aucune des deux populations. Il semble y avoir une tendance à l’autosimilarité pour la totalité des signaux, et légèrement moins marquée chez les signaux sains.

Questionnement : les différences n’étant pas réellement marquées, sont-elles vraiment caractéristiques de la maladie ?

Les mesures traditionnelles de la voix telles que la gigue ou le shimmer produisent des valeurs proches de zéro et de manière peu différenciée entre les deux populations.

Néanmoins des mesures plus originales, telles que le bruit ou les mesures non linéaires semblent (DFA, PPE, spread1, spread2, RPDE , capables de détecter finement d’éventuelles “déviations”) porter davantage d’information.

Parmi les 23 attributs, certains semblent porter la même information (distribution similaire). Elles pourraient mesurer le même phénomène physique, ou ce phénomène physique pourrait induire un autre, qui ferait également l’objet de mesures présentes dans notre jeu de données.

Ainsi, étudions les éventuelles corrélations entre les 23 attributs ci-dessus.

* 1. **Etudes des corrélations**

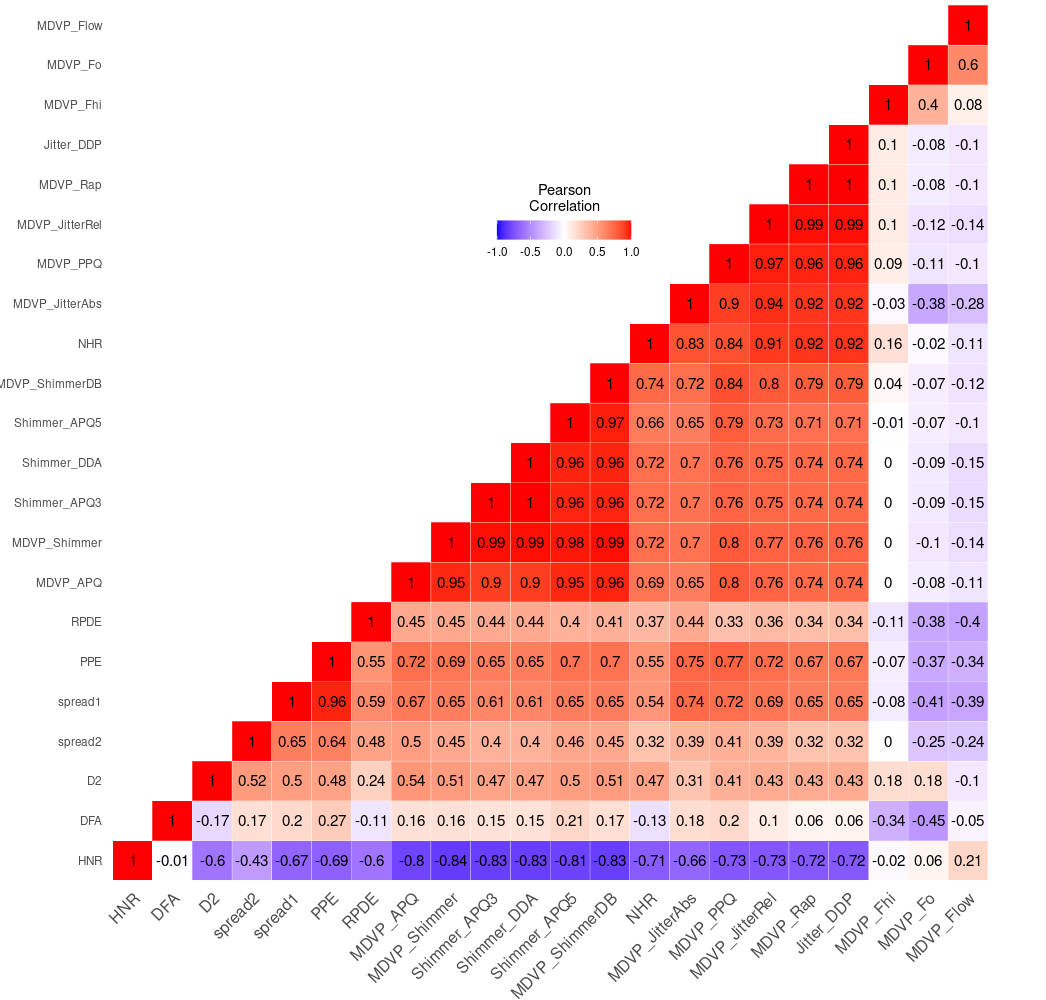


Figure 27-1

Le coefficient de corrélation utilisé ci-dessus est celui de Pearson, davantage capable de discriminer les relations entre variables qui n’augmente pas de manière constante.

----> réduction de dim ??

A prévoir de grosses corrélations entre certains paramètres, car certains paramètres mesurent des aspects très voisins (Jitter vs Shimmer) voire le même phénomène physique. Ex : Jitter(%) et Jitter(Abs). Si la corrélation entre deux paramètres dépassent 0.95, alors pour la suite de l'étude, seul un des deux paramètres n'est utilisé.

-------------------------------------------------------------------

MDVP.RAP et Jitter.DDP sont linéairement dépendants : cf plot

# Classification non supervisée

D’après la problématique qui est la nôtre, on devrait à priori avoir 2 groupes : un pour les personnes atteintes de la maladie de Parkinson et un autre pour les personnes en bonne santé.

Sur cette base nous implémenterons tour à tour les méthodes de clustering : kmeans, pam et CAH.

**3.1 KMEANS**

**3.1.1 Estimation du nombre de profils type**

Pour déterminer le nombre optimal de classes pour le kmeans, nous utilisons l’indice « silhouette moyenne » selon deux approches.

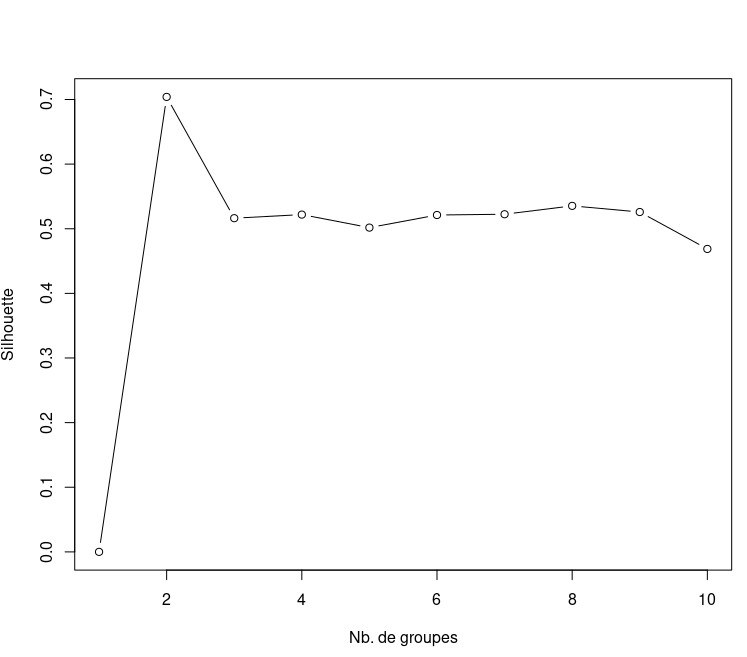
Les résultats obtenus sont les suivants

Figure 311-1

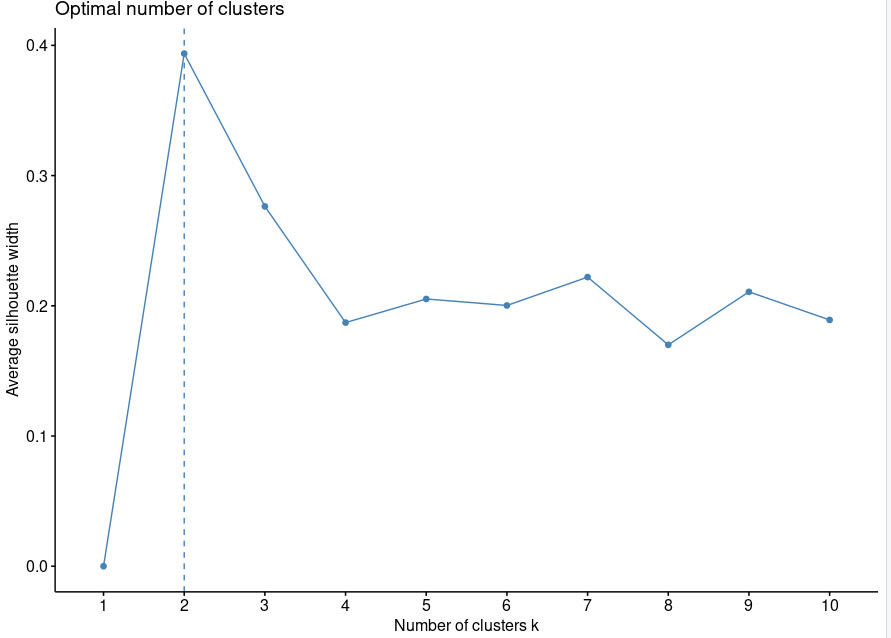


Figure 311-2

On peut donc soupçonner que «2 est le bon nombre de profils type».

**3.1.2 Profils types et interprétation**

L’application du kmeans avec 2 groupes donne :

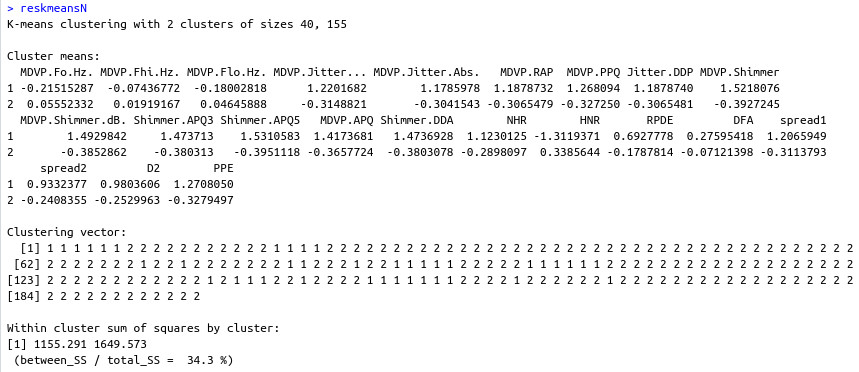
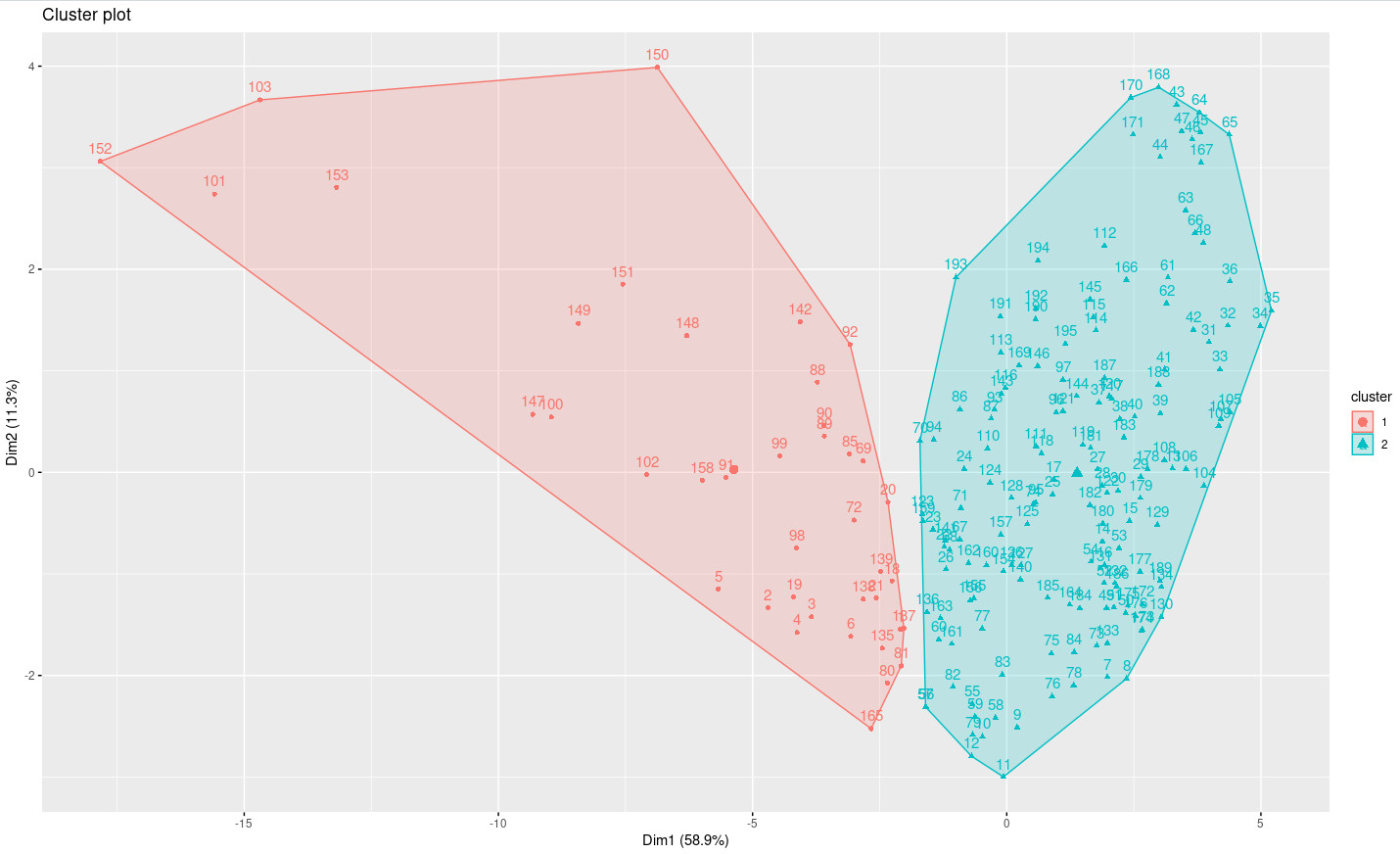


Figure 312-1

On est en présence de deux groupes dont les effectifs semblent se rapprocher de ceux du jeu de données de départ. Seulement, l’inertie est mauvaise; et d’après le classement proposé, les deux groupes ne correspondent pas à ceux qu’on a dans la réalité. Par exemple, les 31 premiers individus sont malades -donc de classe 2 d’après la classification proposée par kmeans- mais il se trouve qu’on a plusieurs individus de la classe 1 entre le 1er et le 31e.

La représentation graphique nous permet de confirmer cela malgré l’obtention de deux groupes distincts et bien séparés

Figure 312-2

En définitive on a deux groupes séparés mais leur structure est différente de celles du jeu de données initial ; la méthode kmeans ne nous donne donc pas un résultat intéressant.

**3.2 PAM.**

**3.2.1 Estimation du nombre de profils type**

L’estimation du nombre idéal de classes à partir de la valeur silhouette moyenne nous donne le graphique suivant:

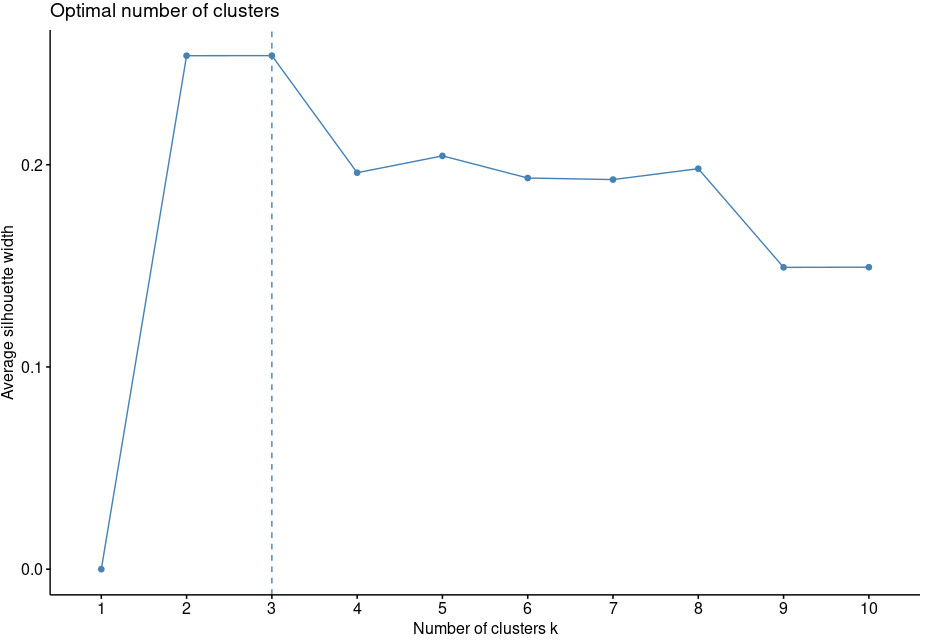


Figure 321-1

On aurait donc 3 profils type

**3.2.2 Profils types et interprétation**

Avant de regarder le cas d’un clustering à trois groupes comme suggéré par le point précédent, nous nous pencherons sur deux groupes. En effet, il s’agit de l'intuition énoncée à l’entame de ce travail et s’appuyant sur la structure du jeu de données initial.

L’application de la méthode PAM à 2 classes donne :

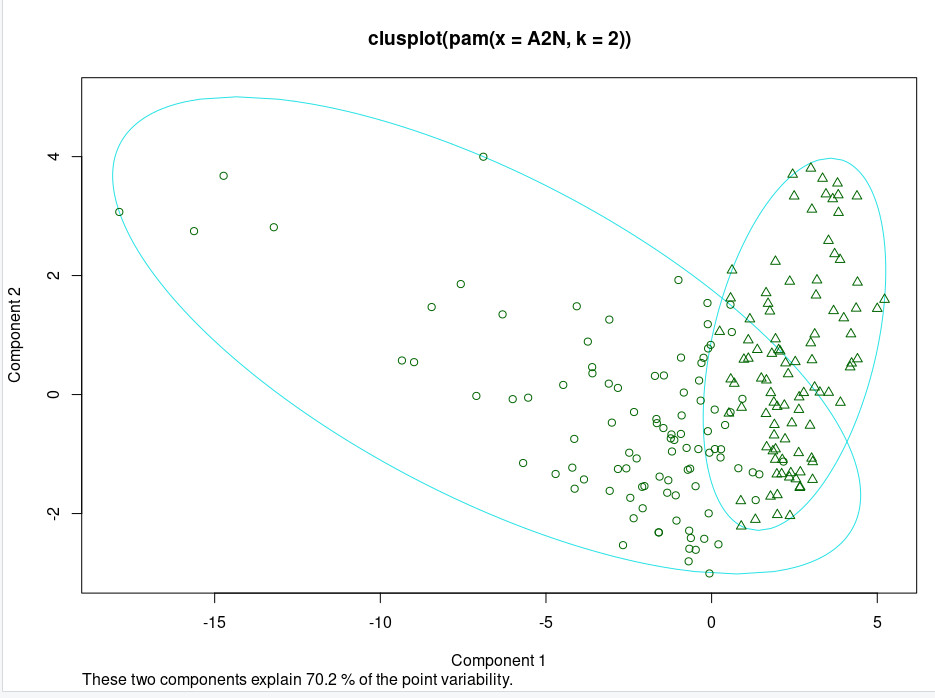


Figure 322-1

Ce graphique nous montre qu’on a clairement 2 groupes qui ne sont pas isolés malgré un fort taux d’explicabilité.

La répartition des individus dans les classes vient illustrer doublement cela comme on peut le voir :

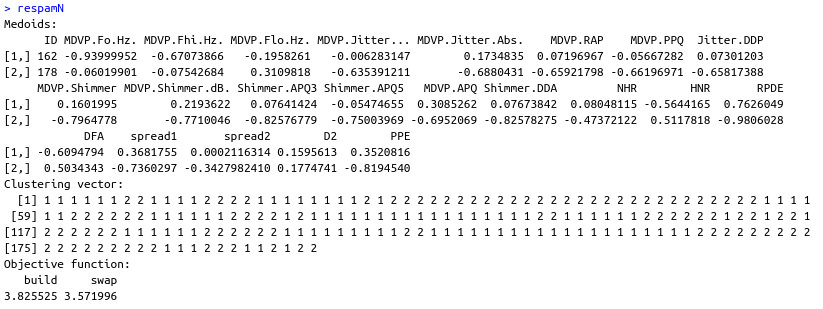


Figure 322-2

On peut se risquer à dire que “2 n’est pas le bon nombre de profils type” ; Ceci est confirmé par la visualisation des valeurs silhouettes sur la figure suivante.

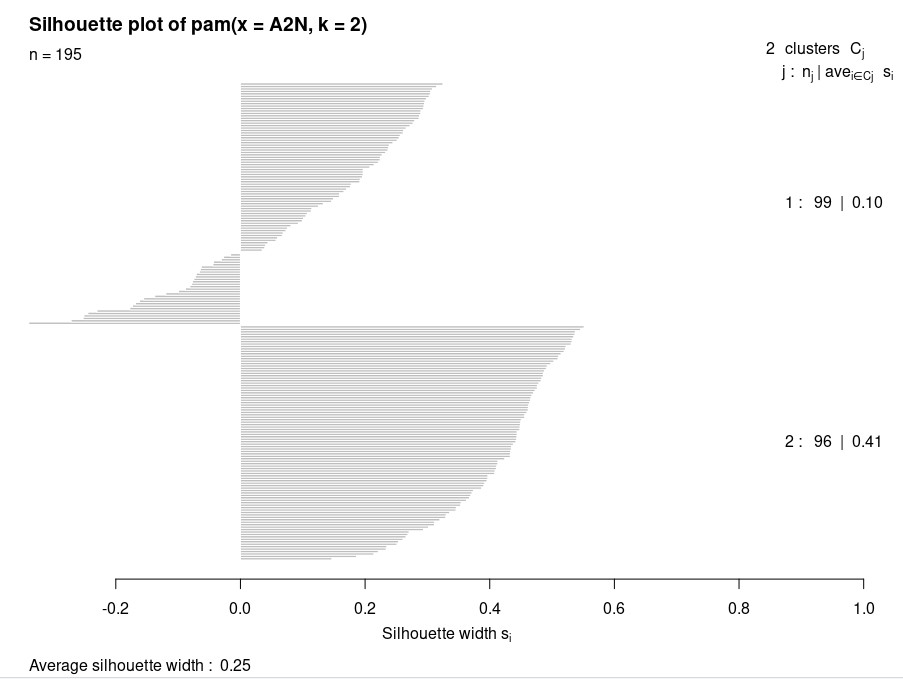


Figure 322-3

En effet nous avons :

- des valeurs silhouettes très faibles

- des points (de la classe 1) avec des valeurs silhouettes négatives (donc les points sont mal classés)

- les effectifs des 2 classes sont sensiblement les mêmes, ils ne reflètent en aucune manière les proportions du jeu de données initial.

Une analyse similaire menée avec trois clusters nous donne des résultats qui ne sont guère meilleurs que les précédents. L’illustration est faite sur les figures ci-après:

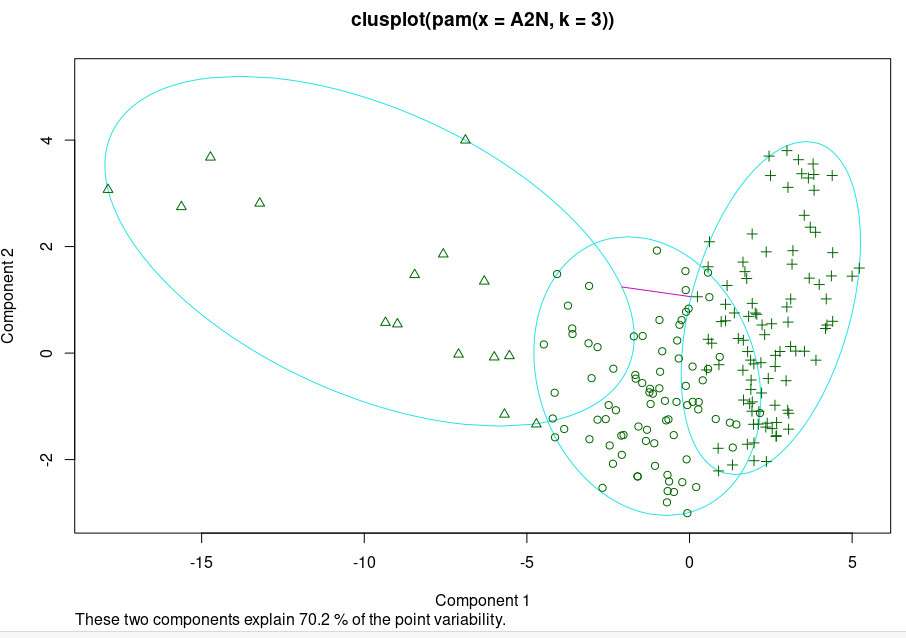


Figure 322-4

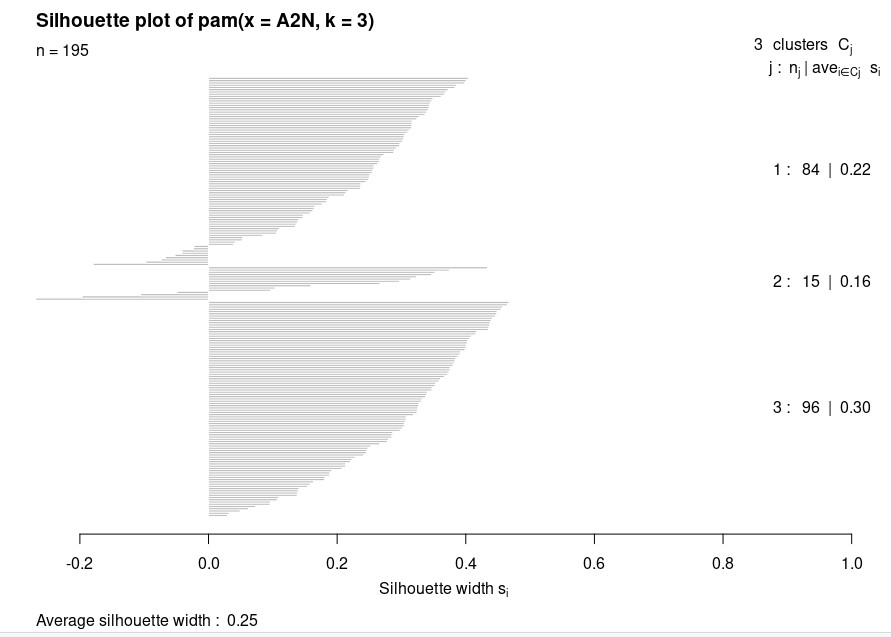


Figure 322-5

Au vu de ce qui précède, on peut conclure que la méthode PAM ne permet pas d’obtenir un bon clustering de notre jeu de données.

**3.3 CAH**

**3.3.1 Estimation du nombre de profils type**

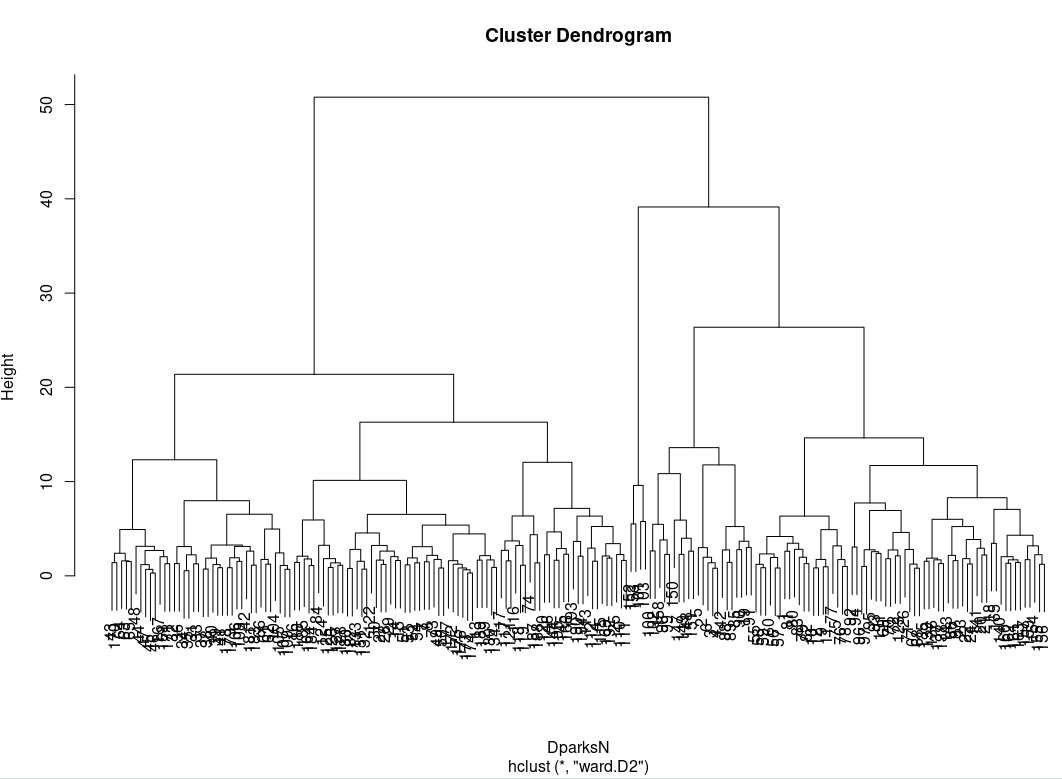
Le calcul du dendrogramme donne la figure suivante

Figure 331-1

A la vue de cette figure on peut envisager deux classes. Pour le confirmer, regardons le graphique des pertes relatives d’inertie.

La meilleure partition selon ce critère est représentée par un point noir et la seconde par un point gris.

Un découpage en deux classes minimise ce critère.

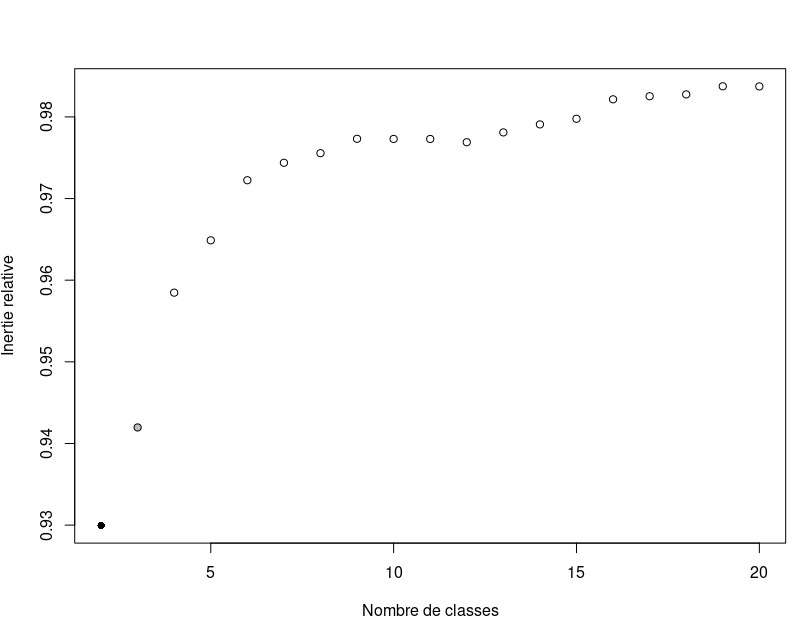


Figure 331-2

Nous pouvons procéder à la coupe comme suit :

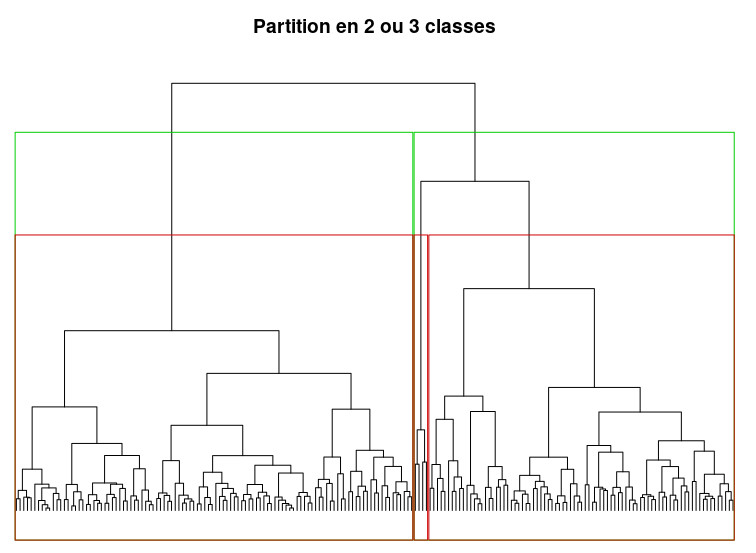


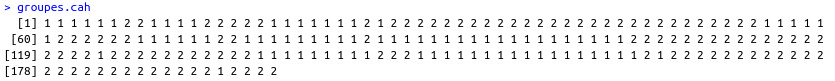
Figure 331-3

Au vu de cette figure, une analyse à 3 classes ne semble pas pertinente car il n’y aurait pas un supplément conséquent d’informations

Nous allons donc retenir un découpage en deux classes.

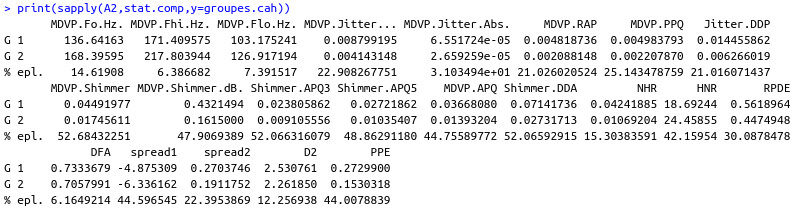
**3.3.2 Profils types et interprétation**

Les deux groupes du dendrogramme sont visualisés sur la figure ci-après

Figure 332-1

Les individus dans ces groupes ne sont pas classés comme dans le jeu de données de départ.

Pour pousser l’analyse nous décidons de comparer les moyennes des variables actives conditionnellement aux groupes. Il est possible de quantifier globalement l’amplitude des écarts avec la proportion de variance expliquée. C’est ce qui est fait dans le graphique qui suit

Figure 332-2

Nous constatons que la définition des groupes est dominée par les “grands groupes de variables” MDVP.Shimmer , MDVP.APQ, Shimmer, spread1, PPE; même si nous notons que la proportion de variance expliquée demeure plutôt faible (<=50%).

Nous concluons par une ACP pour une analyse tenant compte des liaisons entre variables, en espérant avoir des éléments nouveaux

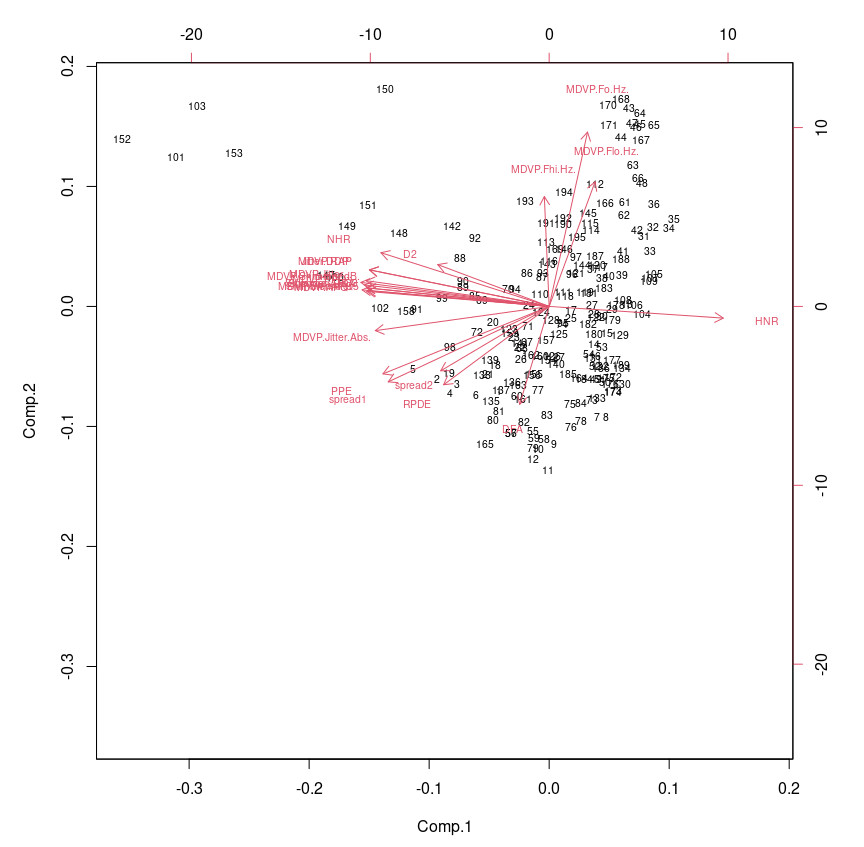


Figure 332-3

Malheureusement l’ACP ne montre pas de séparation nette entre les individus. Nous concluons que la CAH non plus ne nous donne pas un résultat intéressant.

Nous espérons obtenir de meilleures conclusions avec une approche de type analyse supervisée.

# Classification supervisée

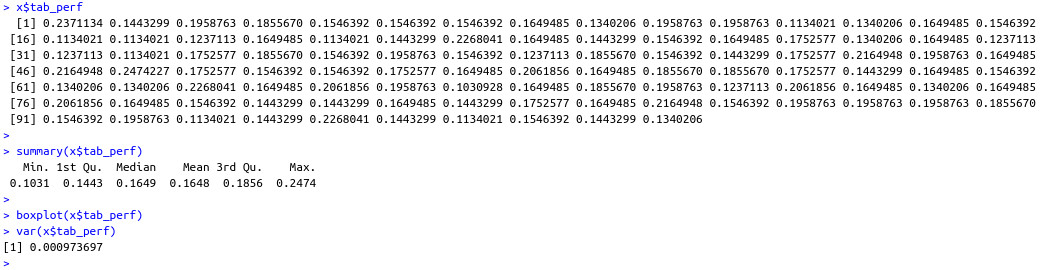
**4.1 Le protocole utilisé**

Pour ce qui est de la classification supervisée, nous utiliserons un protocole d'apprentissage et de tests par Bootstrap avec 50% de la population totale pour l’apprentissage et 50% pour les tests. En effet, la population est peu nombreuse (195 individus) et suivant le critère classe, 50% d’individus sains correspond à 24, ce qui est faible mais acceptable - contrairement à 20% habituellement pris et correspondant à 9 -.

**4.2 Les performances de l’arbre appris**

Nous voulons noter que sur un test itératif effectué manuellement par nos soins, nous avons une prédominance statistique (17 sur 20 soit 85%) de la variable PPE comme étant la plus prédictive. Pour plus d’efficacité, nous écrivons un script qui va automatiser l’expérience aléatoire et itérer un plus grand nombre de fois (ici 100 fois).

Analysons la distribution des erreurs obtenue :

Figure 42-1

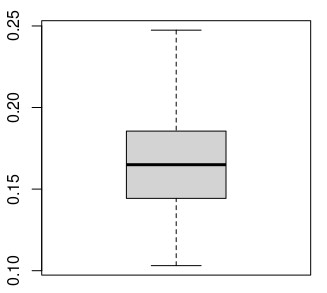
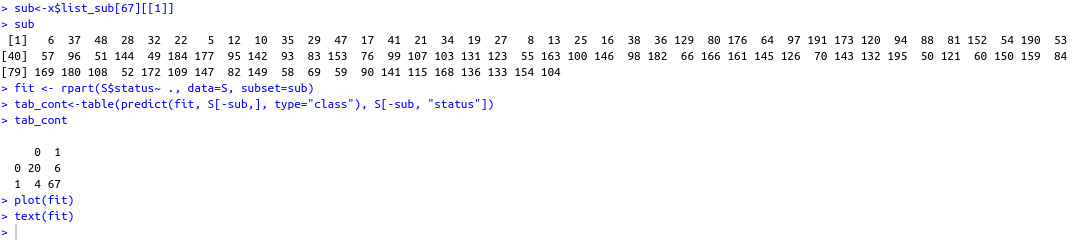


Figure 42-2

Ainsi le meilleur arbre aurait une erreur de 10,3%; pour une moyenne et une variance de distribution respectivement égales à 16,48% et 0.00097. Ce qui est déjà intéressant comme résultat. Mais dans le souci d’être le plus précis possible, nous choisissons d’enlever cette valeur (10,3%) pour voir comment se comporte la variance. Nous obtenons une erreur de 11,34% pour une moyenne des erreurs égale à 16,55% et une variance de 0.00094. Par conséquent la différence n’est pas significative. Nous restons donc sur les premières valeurs, soit une performance de 10,3% d’erreur, avec une moyenne de 16,48% et une variance de 0.00097.

A partir de l’échantillon associé, on calcule la table de contingence et l’arbre construit :

Figure 42-3

**4.3 Arbre induit et règles de décision.**

Le schéma de l’arbre annoncé est le suivant

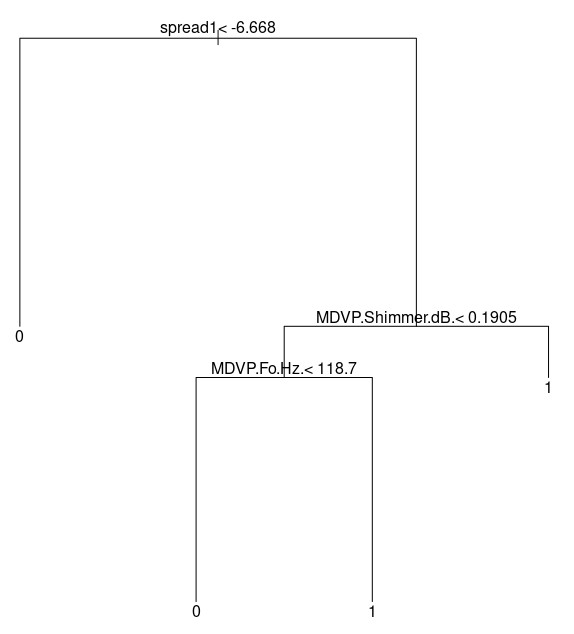


Figure 43-1

Par ailleurs les règles de décision sont résumées comme suit

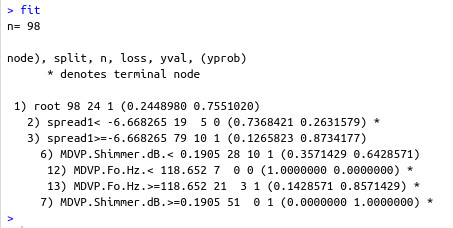


Figure 43-2

Il apparaît ici que la variable la plus prédictive est ‘spread1’

Sur 98 individus prélevés la majorité (environ 75,51%) est malade (yval=1) et le reste (24 équivalent à 24,48% environ) est en bonne santé.

Sur les 98 on en a 19 qui vérifient la condition ‘spread1<-6.668’ et la majorité (73,68% environ) est en bonne santé (yval=0)

On peut donc conclure qu’on a une confiance à 73,68% que si un individu a une valeur de ‘spread<-6.668’ alors il est en bonne santé.

De la même manière si spread1 >=-6.668’ et ‘MDVP.Shimmer.dB>=0.1905’ on prédit (avec une confiance de 100%) que la personne est malade.

Enfin si ‘spread >=-6.668’ et ‘MDVP.Shimmer.dB<0.1905’ et

* si ‘MDVP.Fo.Hz<118.652’ on prédit que la personne est en bonne santé (confiance de 100%)
* si ‘MDVP.Fo.Hz>=118.652’ on prédit que la personne est en malade avec une confiance de 85,71%

# Une autre piste : ACP et Régression logistique

## ACP

* + 1. Justification

Les algorithmes précédents nous ont conféré des clusters peu éloignés, qui ne nous permettent pas une classification satisfaisante.

En cause, des données qui semblent “déséquilibrées” : :

* 1 : soit, nous avons un nombre d’individus plutôt faible (195) au regard du nombre d’attributs analysés (23) : manque d’entraînement possible.
* 2 : soit, nous avons un nombre d’attributs trop élevé (23) au regard du nombre d'individus étudiés. Certaines caractéristiques d’individus bruitent potentiellement l’émergence d’une tendance :
  + cause 2.1 : parce qu’une/des caractéristique(s), potentiellement fortement corrélée(s) à une/des autre(s), donnerai(en)t une redondance d’information, et la prépondérance de cette dernière pourrait venir occulter l’information que pourrait apporter une/d’ autre(s) caractéristique(s),
  + cause 2.2 : ou parce qu’une/des caractéristique(s) n’a aucun effet sur le statut que l’on cherche à prédire.

Lors de l’analyse exploratoire, nous avons relevé de fortes corrélations (cf Image XXX) entre certains des 23 attributs. Ceci suffit à suspecter la cause 2.1 posée ci-dessus.

L’Analyse en Composantes Principales semble adaptée à notre situation d’un faible nombre d’individus pour un fort nombre d’attributs quantitatifs. Les points de notre jeu de données s’exprimeront sous de nouveaux attributs - nos composantes principales - qui seront les plus décorrélées possibles, et donc les plus informatives possibles.

A l’issue de l’ACP, nous chercherons à prédire l’attribut “statut”. Celui-ci ne doit donc pas entrer dans la décomposition linéaire en vecteur propres, il est exclu de facto de l’ACP.

Les points de notre jeu de données s’exprimeront toujours dans un repère à 22 dimensions, mais dans un repère où les dimensions seront choisies de manière à porter le plus l’information (de manière à ce que la variance soit maximisée avec un minimum de dimension). Ces dimensions choisies sont les vecteurs propres, parmi lesquels nous devrons élire celles qui maximisent la variance.

* + 1. Mise en œuvre

**Script de référence : Ana\_explo\_PCA.R**

L’ACP est à réaliser sur notre jeu de données de 22 attributs, normalisé, car les variables mesurées pour nos individus sont exprimées dans différentes unités.

Le package FactoMineR sera fort utile pour la mise en oeuvre de l’ACP (fonction PCA()), et la disponibilité des vecteurs/valeurs propres, de la contribution des attributs initiaux dans la composition des nouvelles composantes, et de l’inertie portée par chacun des axes associés aux vecteurs propres.

Le package factoextra (basé sur ggplot2) sera utilisé pour visualisation des caractéristiques de résultats de l’ACP et interprétation (get\_eigenvalue()), fviz\_eig()). L’extraction par get\_eigenvalue()) permet une visualisation de la quantité de variance expliquée par chaque axe principal, de manière ordonnée : les premiers axes sont associés aux plus grandes quantités de variation du jeu de données. La production du graphique des valeurs propres (“scree plot”, Figure 512-1 se fait par fviz\_eig()).

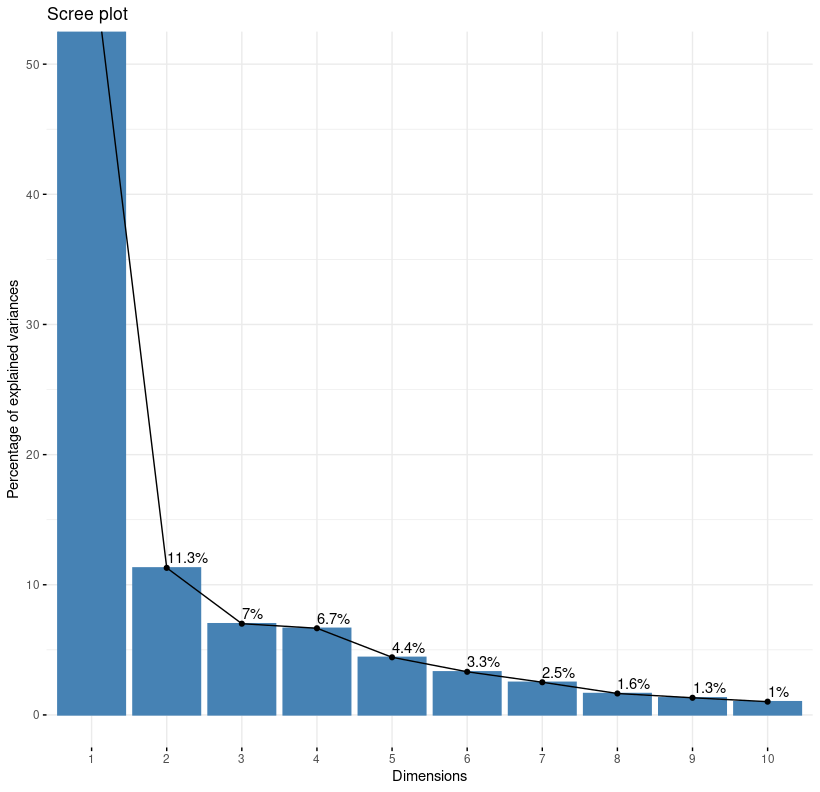


Figure 512-1

La fonction get\_pca\_var() du package factoextra extrait les résultats de l’ACP.

La lecture des “cosinus carré” permettra de visualiser la corrélation des attributs initiaux aux composantes principales, sur le cercle de “corrélation”. Le cercle de corrélation est tracé par fviz\_pca\_var() (factoextra) (figure 512-2).

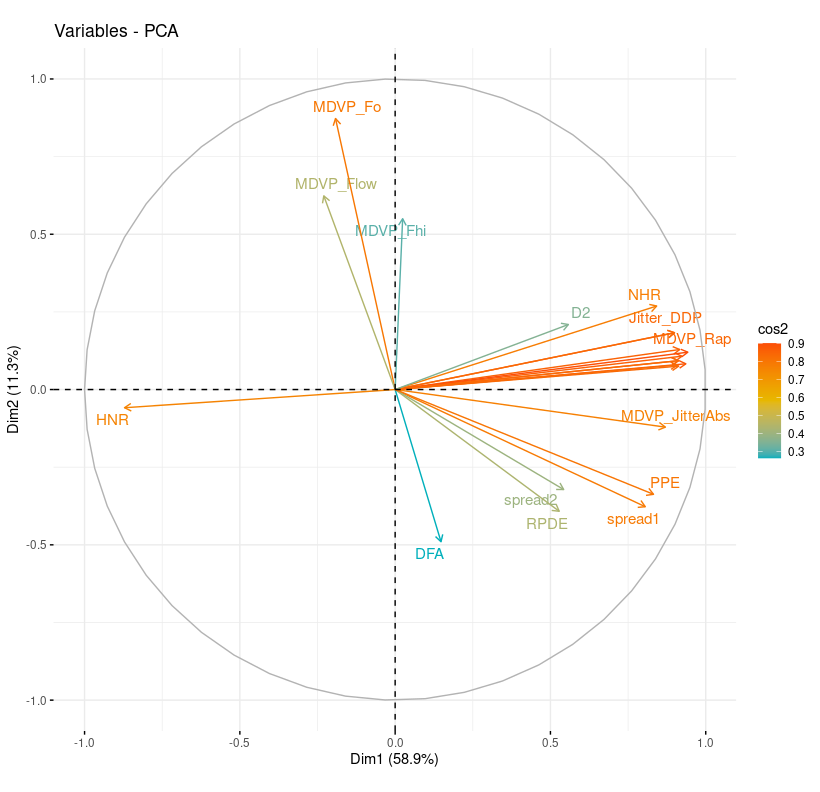


Figure 512-2

Celui-ci ne pouvant se tracer que sur le plan (de notre écran), nous n’y visualisons que la corrélation des attributs initiaux sur les 2 premières composantes principales. Pour y remédier, nous préférons le graphique tracé par corrplot() (factoextra)

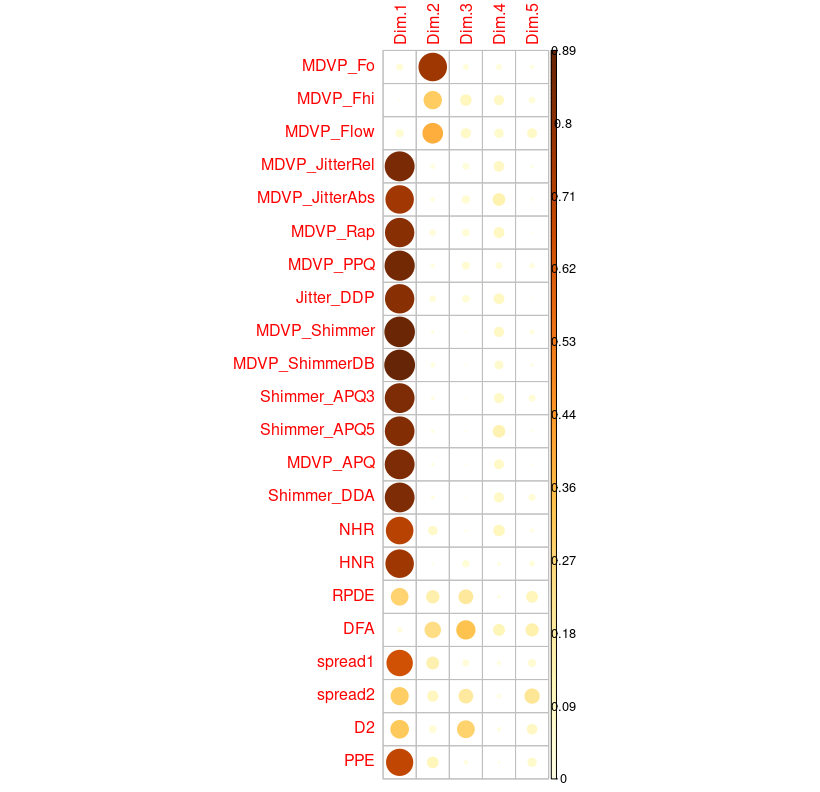


Figure 512-3

Nous utilisons les mêmes outils pour visualiser les contributions (cercle : figure 512-4; graphique pour les 5 dimensions : figure 512-5).

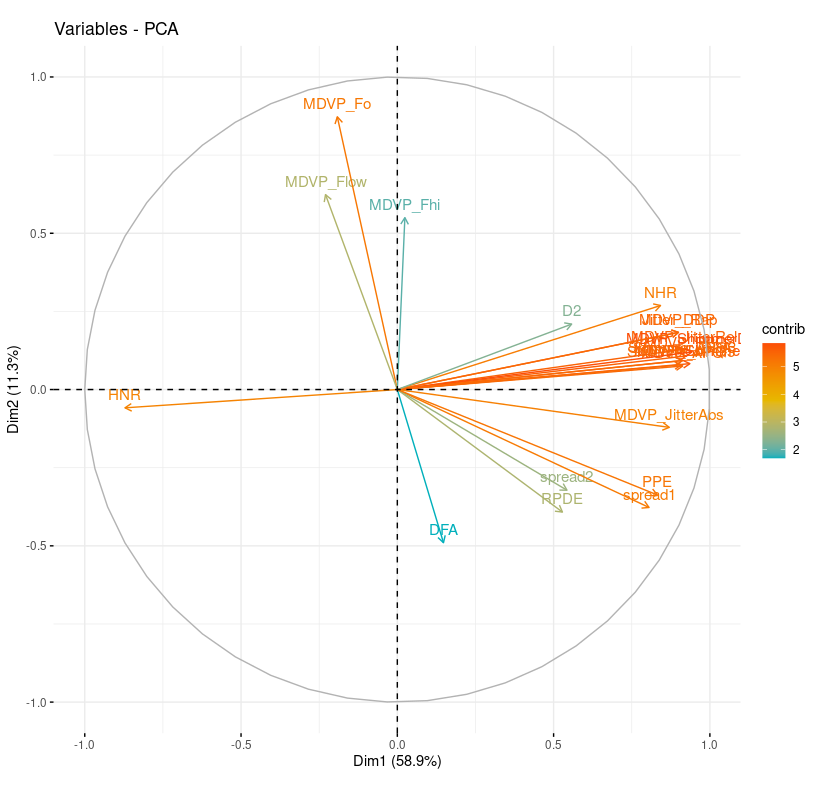


Figure 512-4

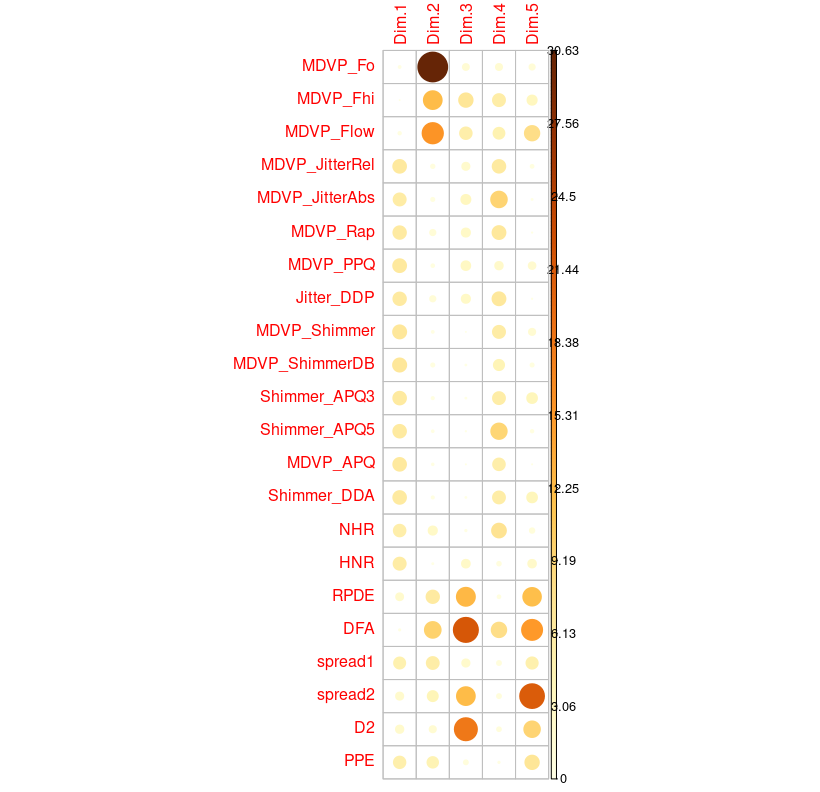


Figure 512-5

Nous ajoutons pour chacune des 5 premières composantes, un barplot des attributs qui contribuent le plus (en%) via la méthode fviz\_contrib() (factoextra) :

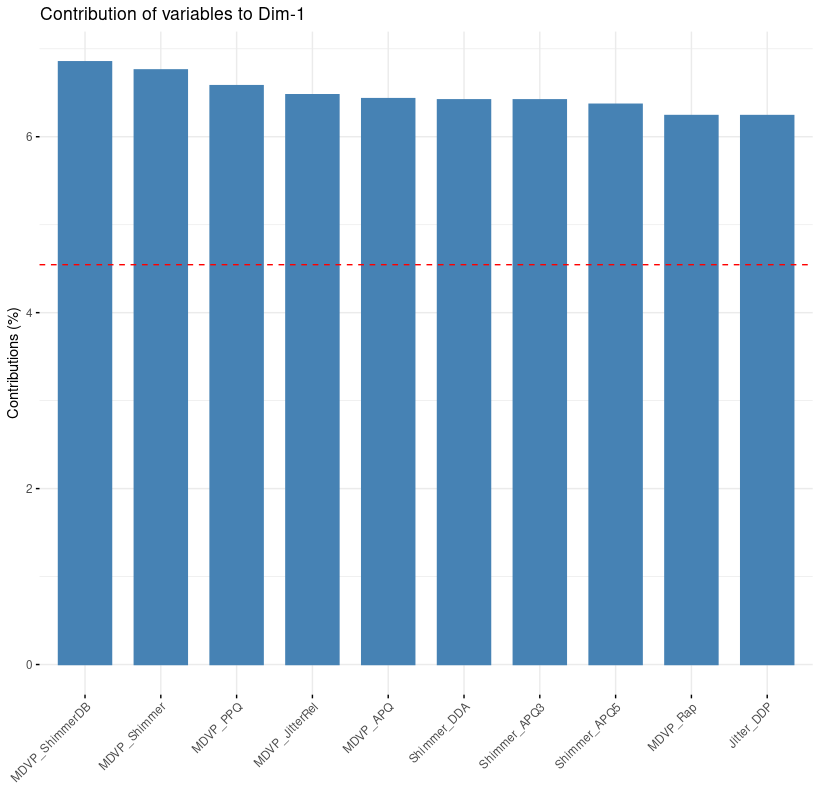


Figure 512-6

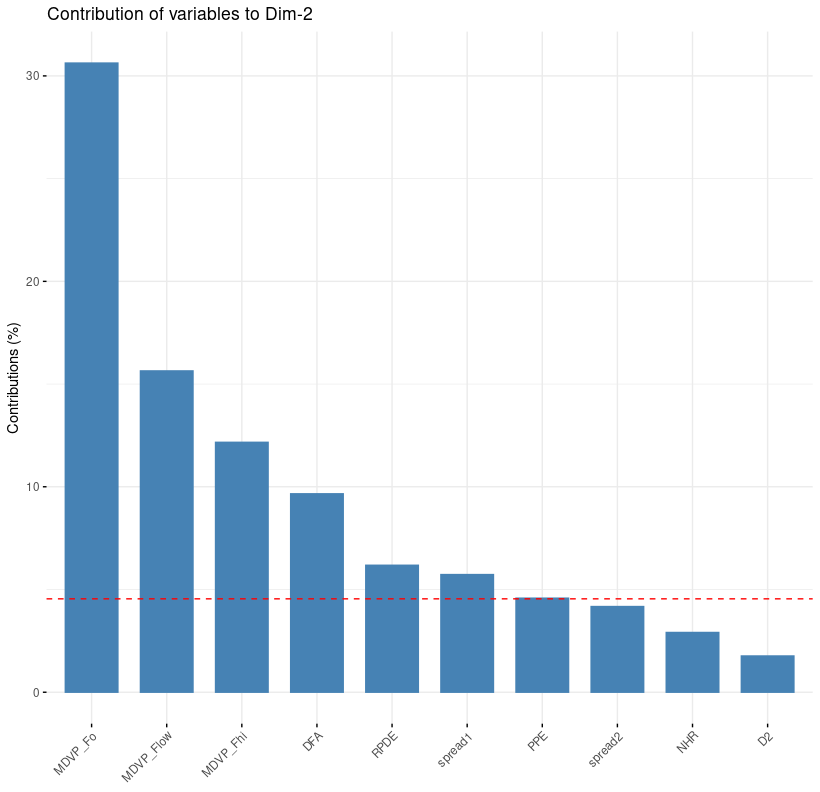


Figure 512-7

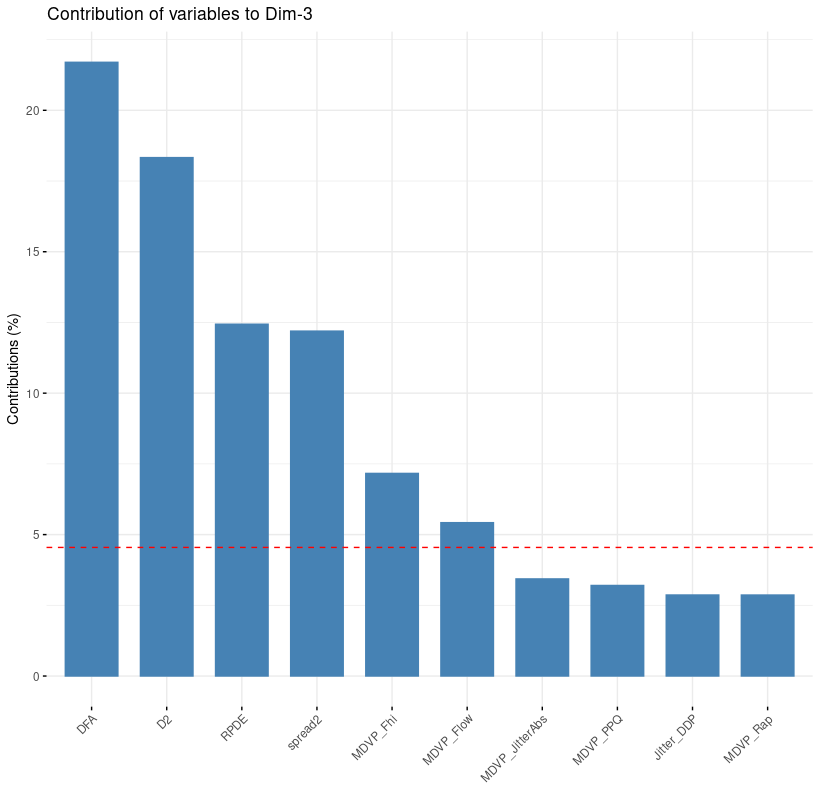


Figure 512-8

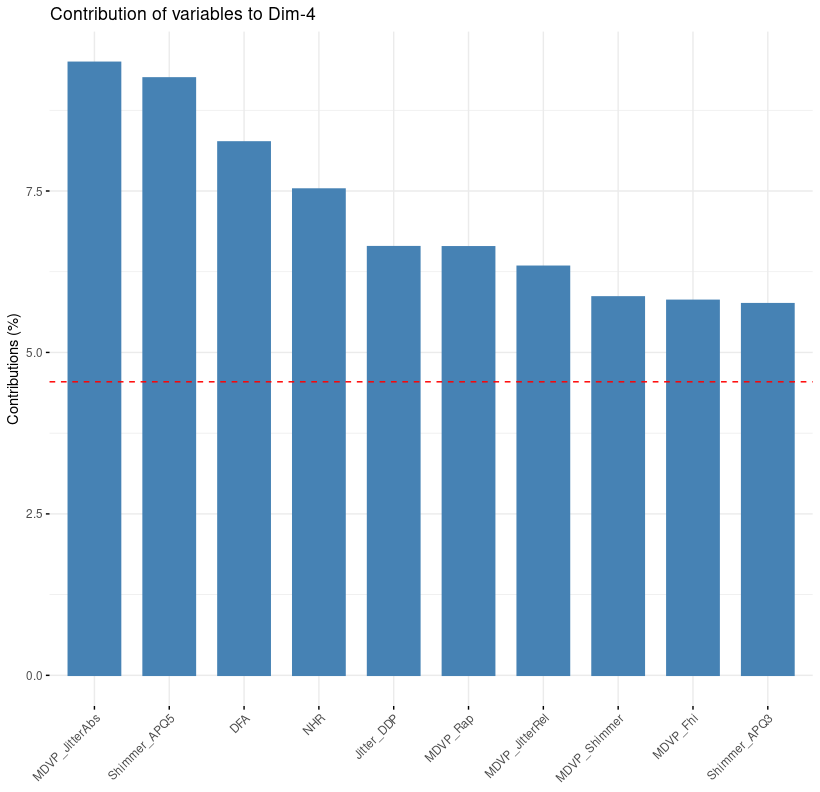


Figure 512-9

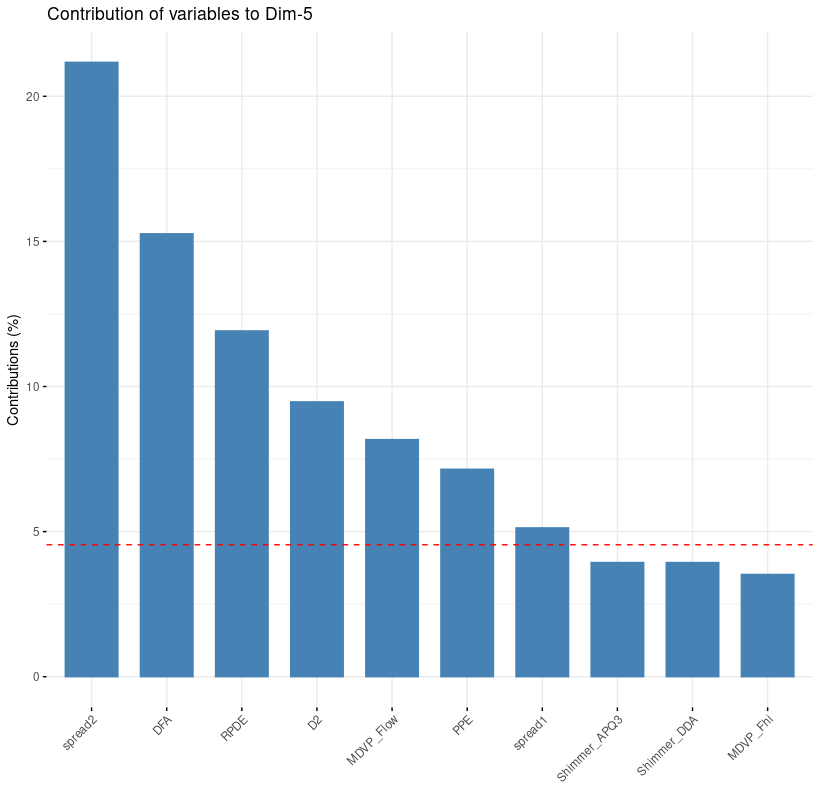


Figure 512-10

* + 1. Analyse des résultats

Analyse des vecteurs propres, choix des 5 premières composantes :

* + Les 5 premières composantes portent 86.5% de l’inertie de l’entier jeu de données (= somme des variances des caractéristiques étudiées).
  + Ceci est supérieure à la valeur référence de **76.25%**, la variabilité expliquée par ces 5 premières composantes est donc significative (cette inertie de référence est le quantile 0.95-quantile de la distribution des pourcentages d’inertie obtenue en simulant 2646 jeux de données aléatoires de dimensions comparables sur la base d’une distribution normale).
  + Choisir au delà de 5 nous expose à un effet de surapprentissage ( bas biais / haute variance)

Analyse des corrélations

* Plus une variable est proche du cercle de corrélation, plus elle est importante pour interpréter les composantes principales en considération.
* Ici, MDVP\_Fhi D2, et DFA sont peu corrélées aux composantes principales, ce qui n’est pas le cas du couple MDVP\_Flow / MDVP\_Fo. HNR est fortement corrélée à NHR négativement (ce qui est naturel puisque l’un est l’inverse de l’autre), cette dernière est fortement corrélée aux autres attributs de type Jitter et MDVP (sauf MDVP\_Flow et MDVP\_Fo).
* La visualisation des contributions aboutit aux conclusions similaires, les contributions n’étant pas plus que le reflet de la corrélation de chaque attribut avec les composantes principales, exprimées en “part” (%). Le graphique des contributions sur 5 dimensions permet clairement de voir la prépondérance de MDVP\_Fo dans la 2è composante.

## 5.2 Régression logistique binaire (GLM binaire)

## 5.2.1 Justification

La régression logistique est un modèle prédictif où l'étiquette de classe ou la cible est catégorielle. Cette variable possède deux catégories, oui / non. Pour cet ensemble de données sur la maladie de Parkinson, la catégorie sera de savoir si la personne est atteinte de la maladie de Parkinson ou si la personne n'est pas atteinte de la maladie de Parkinson. La régression logistique ne peut être utilisée qu'avec deux types de variables cibles. 1. Une variable cible catégorielle qui a exactement deux catégories (c'est-à-dire une variable binaire ou dichotomique). 2. Une variable cible continue dont les valeurs sont comprises entre 0,0 et 1,0 représentant des valeurs de probabilité ou des proportions

L'une des hypothèses critiques de la régression est que les variables indépendantes ne sont pas la combinaison linéaire les unes des autres.

Ici pas de pb car la glm s'applique les caractéristiques principales qui par def sont indépendantes les unes des autres

## 5.2.2 Mise en oeuvre du modèle

**Script de référence : Ana\_explo\_GLM.R**

Nous n’avons pas réussi à utiliser la fonction GLM sur les résultats de la méthode PCA(), mais sur les résultats d’une autre méthode de PCA nommée “princomp()”, et utilisant la méthode que la méthode PCA().

Etape 1 - définir les prédicteurs et l’attribut de réponse de la prédiction.

Les prédicteurs sont tous les attributs qui vont nous permettre de prédire si le signal correspond à un patient malade ou non. Ce sont donc tous les attributs du jeu de données à l’exception de celui pour lequel nous cherchons la réponse à la prédiction, à savoir le statut. Sur le script, leurs valeurs sont stockées dans *predictorX.*

Etape 2 - réaliser la PCA sur les prédicteurs

Nous stockons les résultats dans l’objet *pca.* Nous récupérons l’argument scores (*pca$scores)* que nous stockons dans *pc.comp*. Cet objet correspond à l’expression de notre jeu de données (sans le statut) dans une base de vecteurs propres dans laquelle la variance totale est maximale.

Etape 3 - Découpage “aléatoire” du nouveau jeu de données en données de tests et données d’entraînement du modèle GLM.

Nous paramétrons ici un découpage 80% (entrainement)-20%(test) du jeu de données de manière randomisée (en initialisant le seed et utilisant sanple()). La partie entraînement sera un dataframe nommé *train* et la partie test *test.*

Etape 4 – Application du modèle GLM sur le jeu d’entraînement *train\_df*

Nous déclarons la colonne *status* comme la réponse de notre modèle. Cette variable réponse s’appelle *responseYtrain.*

Nous avons vu dans le paragraphe précédent que la sélection des 5 premières dimensions de la PCA constituait un choix raisonnable. Nous appliquons donc le modèle sur les 5 premières colonnes du dataframe *train* en utilisant la mérhode glm().

La réponse n’étant qu’une probabilité pour que le signal corresponde à un patient malade, nous approchons par 1 toutes les réponses correspondant à une probabilité strictement supérieure 0.5. Les autres réponses sont approchées par 0.

Nous calculons la précision du modèle en comptant le nombre de fois où la réponse du modèle correspondait à la valeur à prédire.

Etape 5 – Application du modèle GLM sur le jeu de tests *test\_df*

Nous procédons exactement comme à l’étape précédente sur le dataframe *test\_df.*

Etape 6 - Répétition des étapes 4 et 5 sur un nombre *nboot.*

Nous répétons les étapes 4 et 5 dans une boucle un nombre *nboot* de fois, *nboot* est initialisé à 30.

Nous affichons les résultats de précision du modèle pour les *nboot* et le résultat moyen à la sortie de la boucle. Ce dernier tourne autour de 85%.

5.2.3 Analyse

Pour ajouter un nouvel individu, nous pouvons espérer dans 85% des cas avoir une bonne prédicition du modèle. L’auteur utilise pour sa part un svm.

# Epilogue

Peut-être que d’autres attributs auraient donné un meilleur résultat encore. Il existe depuis 2020 un logiciel libre d’analyse des sons vocaux, de nom de “Praat”. Celui-ci inclut par exemple des calculs caractérisant les arrêts de voix. Voir ici : <https://www.fon.hum.uva.nl/praat/>

Sur internet, on trouve des études qui conseille xgboost

Les ACP que nous avons utilisées aboutiraient peut-être à de meilleures prédictions si les coefficients permettant le calcul de la matrice de corrélation sur laquelle se base sur la PCA sont en mesure (!) de détecter les relations non linéaires entre attributs (ce qui n’est pas le cas du coefficient de Pearson).

