Projet LSTAT2110(A) – Analyse de données

Etude des mesures physiques des abalones par âge et sexe

Louis Descarpentries (3004 18 00 - BIRE2M) - Mattias VAN EETVELT (1660 18 00 - BIRE2M)

Contents

1 Introduction

Ce projet rentre dans le cadre du cours *LSTAT2110A – Analyse des données*, et a pour but de mettre en pratique deux méthodes vues en théorie. Celles-ci seront appliquées sur une base de données issue de Kaggle (détaillée dans la section 2).

La première méthode consiste en une Analyse en Composantes Principales (ACP). Il s'agit d'une méthode projetant les observations d'un espace à p dimensions avec p variables vers un espace à k dimensions où $k < p$. Ainsi, cette technique permet de simplifier la visualisation de la matrice de données, mais aussi de transformer les variables initiales en variables non corrélées (diminue la redondance) (XLSTAT,2022). Ces nouvelles variables sont appelées Composantes Principales, classées en fonction de leur degré de similarité par rapport à l'initial.

Ces étapes préalables sont nécessaires pour diverses représentations graphiques par la suite, comme le cercle des corrélations (corrélations entre les composantes) et la carte des individus (individus en fonction des composantes principales).

La seconde méthode correspond à une analyse de classification, permettant de grouper des objets dans des classes, en fonction de leur degré de similarité (les objets les plus similaires sont regroupés dans la même classe). Cette seconde méthode peut se diviser en plusieurs analyses de classification comme : classification hiérarchique, K-means clustering, analyse discriminante Linéaire (XLSTAT, 2022).

2 Présentation des données et analyse descriptives

2.1 Présentation des données

La base de données utilisée dans le cadre de ce projet est issue du site Kaggle, une plateforme web appartenant à Google qui rassemble des bases de données, du code utile au traitement de certaines problématiques du domaine de la Sciences des données, et plus largement, une plateforme qui permet des échanges entre passionnés, professionnels et novices de la Science des données (Kaggle, 2021).

Dans le cadre de ce projet, la base de données choisie a été publié en 2018 sur la plateforme par Rodolfo Mendes. Il est à noter que M. Mendes n'est pas le producteur des chiffres présentés car il se base sur les données issues d'une étude réalisée en 1994 par Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn et Wes B Ford dont les chiffres ont été publié par Sam Waugh du département d'informatique de l'Université de Tasmanie (UCI, 1995).

L'objet de la base de données concerne les ormeaux ou Abalone, et plus particulièrement les mesures physiques de ceux-ci. Le but original était de catégoriser les mollusques selon leur âge sur base de mesures physiques (détail dans la suite de la section), et ensuite pouvoir établir un modèle d'extrapolation à l'ensemble de la population pour connaître l'âge sans avoir à faire des mesures sur le terrain.

Avant tout développement, et de manière à faciliter la compréhension dans la suite du rapport, il semble important de décrire les caractéristiques de l'ormeau. *Haliotis tuberculata* ou ormeau est un mollusque mesurant 8 à 11 cm à maturité, possédant une coquille ovale, aplatie, en forme d'oreille et percée de trous alignés suivant sa croissance lui facilitant sa respiration. L'intérieur de sa coquille est nacré, ce qui le rend convoité dans la fabrication de bijoux par exemple. Cette espèce est présente en Méditerranée, en Atlantique nord, dans la Manche, en Mer du Nord et se développe dans des environnements rocailleux car elle se nourrit d'algue en rappant les roches. L'ormeau est très convoité en gastronomie, ce qui a mené à sa surpêche dans différents pays, lui conférant aujourd'hui un statut de protection et des normes de pêche très réglementées. Toutefois, l'élevage d'ormeau s'est développé ces dix dernières années, et ce notamment en France (Muséum-Aquarium de Nancy, s.d.).

La matrice étudiée comporte **8 variables quantitatives**, représentant différentes mesures physiques du mollusque à savoir : LongestShell (longueur de la coquille en mm), Diameter (diamètre perpendiculaire à la longueur en mm), Height (hauteur avec viande dans la coquille en mm), WholeWeight (poids entier en grammes), ShuckedWeight (poids entier une fois décoquillé en gramme) , VisceraWeight (poids boyaux après saignée en grammes), ShellWeight (poids coquille après séchage en grammes), Rings (nombre d'anneaux, donne l'âge lorsqu'on additionne 1,5 au chiffre). La base de données reprend également une variable catégorielle nommée *type* qui reprend **trois types de mollusques** : M pour Male, F pour Female et I pour Infant. Enfin, le jeu de données est divisé en **4177 observations** divisées équitablement (37% de mâles, 32% de jeunes, 31% de femelles) entre I (infant), M (male), F (female).

Finalement, il est évident que ce type de données est relativement important pour étudier la dynamique des populations d'ormeau. En effet, comme énoncé précédemment, le modèle mathématique réalisé sur base de cette base de données, et une fois couplé avec des facteurs extérieurs, permet de connaître l'âge des individus. Cela n'est pas anodin car ces coquillages sont menacés par le changement climatique, et plus précisément l'acidification des océans, mais aussi la surexploitation. Par conséquent, connaître les écosystèmes les plus favorables au développement de l'espèce pourrait favoriser leur protection (BOREA, 2017).

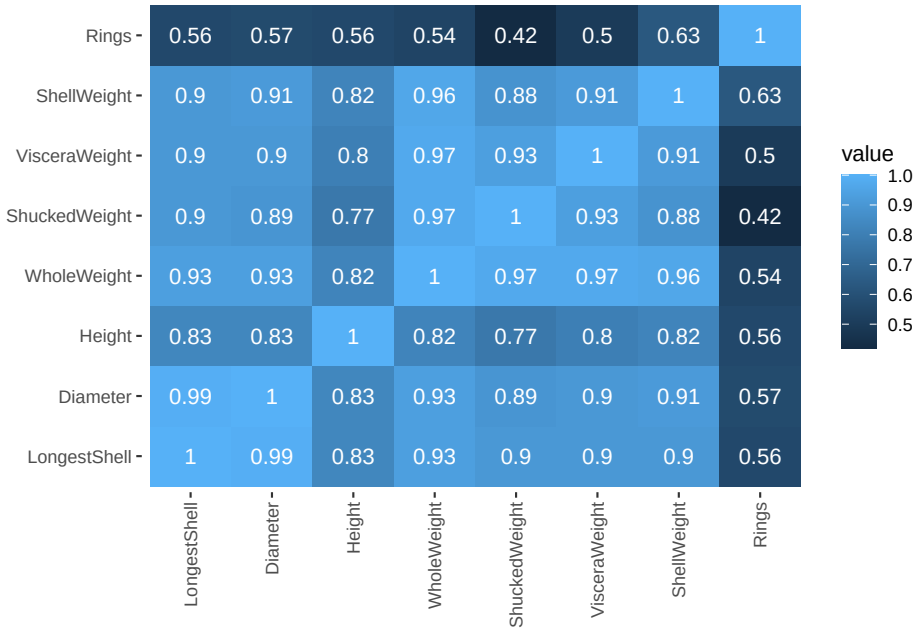
2.2 Analyse descriptive

Afin d'avoir un aperçu géneral des données, la fonction `summary()` est utilisée. Celle-ci permet d'avoir des informations sur des quantiles, minimum et maximum, médiane et moyenne de chaque variable. Afin de rester concis, seules les moyennes de chaque variables selon leur catégorie respective sont présentées. Néanmois, l'output complet de la fonction `summary()` est présenté en annexe. Les longueurs sont en millimètres tandis que les poids sont en grammes.

```
         LongestShell Diameter Height WholeWeight ShuckedWeight VisceraWeight
moyennes        0.524    0.408   0.14       0.829         0.359         0.181
         ShellWeight Rings
moyennes       0.239 9.934
```

On observe que le poids total (`WholeWeight`) est principalement composé par le poids total une fois décoquillé (`ShuckedWeight`) ainsi que par le poids de la coquille. Le poids des boyaux après la saignée (`VisecraWeight`), ne représente quant à elle qu'environ 20% du poids total.

Dans un second temps, il est intéressant d'analyser la matrice de correlation afin d'avoir une idée des correlations entre les différentes variables quantitatives. Les variables fortement correlées auront une correlation de 1 et dans le cas contraire, de -1. Une correlation de 0 indiquera une indépendance des variables.



Les variables *Diameter* et *LongestShell* s'avèrent être les plus corrélées, cela s'explique sûrement par le fait qu'au plus la coquille grandit, au plus le diamètre grandit également. En outre, *Shellweight*, *VisceraWeight*, *Shuckedweight* sont fortement liés à *WholeWeight*, cela semble logique car ces variables sont des constituants du poids total, ainsi elles l'influencent parfois positivement, parfois négativement. Outre cela, il est possible de constater certaines subtilités, c'est par exemple le cas entre *VisceraWeight* et *ShellWeight* par rapport à *ShuckedWeight*. En effet, il est possible de constater que *VisceraWeight* est plus corrélé à *ShuckedWeight*, cela est tout à fait pertinent car *ShuckedWeight* représente le poids décoquillé, c'est donc les composantes du mollusque en tant que tel qui ont le plus d'influence pour ce type de poids.

Enfin, il est possible de remarquer que la variable *Rings* est la moins corrélée aux autres variables, possiblement car il s'agit d'excroissances sur la coquille qui ne se forment pas selon un processus linéaire.

3 Analyse en composante principales

L'analyse descriptive permet d'avoir une idée globale des relations entre les variables et des moyennes générales, mais reste assez floue vu le nombre de variables et d'individus que comporte la base de données choisie. L'Analyse en Composantes Principales va permettre de réduire ces dimensions afin de faciliter l'interprétation. L'objectif de cette partie est donc de trouver des liens clairs entre les différentes variables et d'observer comment les individus sont influencés par celles-ci.

3.1 Application de l'Analyse en Composantes Principales (ACP)

L'ACP est réalisée grâce à la fonction `PCA()` du package `FactoMineR`. Cette dernière centre et réduit les données afin de former la matrice standardisée $\mathbf{Z}$. Autrement dit, pour n observations

$$\mathbf{Z} = \frac{X_{ij} - \bar{X_j}}{\sqrt{n} s_j}$$

Ou $\bar{X_j}$ est la moyenne et s_j la variance. La matrice de corrélation est également calculée et est définie de la manière suivante

$$\mathbf{Z}'\mathbf{Z} = \mathbf{R}$$

Les vecteurs propres ainsi que les valeurs propres correspondantes de la matrice $\mathbf{R}$ sont calculés. Les valeurs propres (`VaP`) sont présentées ci-dessous par ordre croissant. Celles-ci sont essentielles à la PCA puisque le choix du sous-espace représentant les données se fait sur base des deux valeurs propres les plus élevées, associées aux vecteurs propres les plus grands.

On observe que la première composante apporte énormément d'informations comparé aux autres composantes. La quantité d'information apportée par la deuxième, troisième et quatrième composante est non négligeable, ce qui n'est pas le cas pour les quatres dernières.

```
      [,1]   [,2]   [,3]  [,4]   [,5]   [,6]   [,7]   [,8]
VaP 6.7124 0.6956 0.2584 0.166 0.0849 0.0635 0.0127 0.0064
```

On s'intèresse ensuite au pourcentage de la variance ainsi qu'au pourcentage de la variance cumulée. Cela permet de mettre en évidence la quantité d'information retenue par chaque composante.

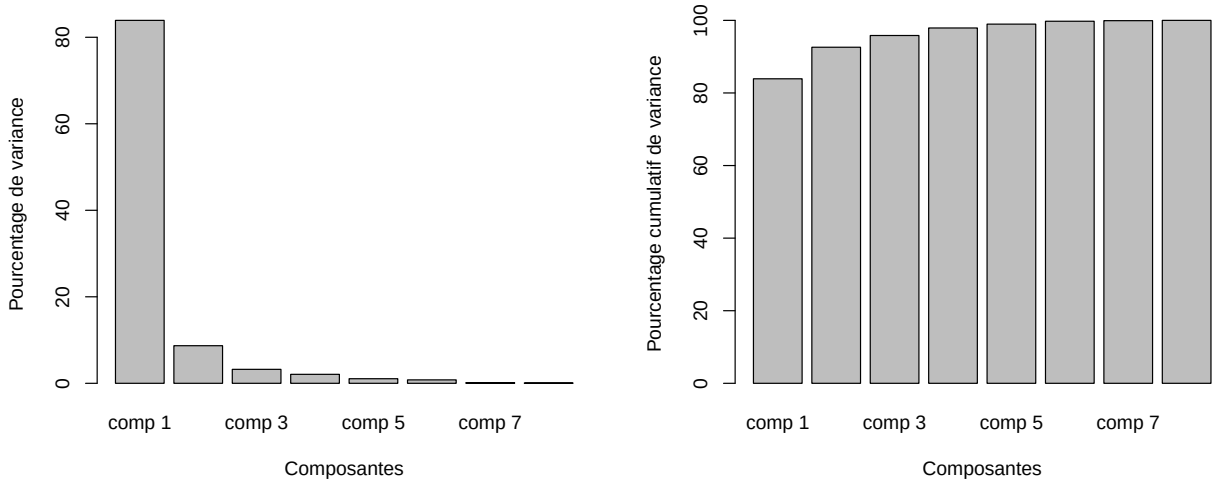


Figure 1: Pourcentages de variance et de variance cumulée

La première composante correspond à une valeur propre de 6,71 et capte à elle seule 83,9% de la variance totale, tandis que la deuxième composante en capte 8,69%. Cela veut dire que les deux premières composantes captent à elles seules 92,6% de la variance totale. Il est à notifier que la troisième et quatrième composante captent respectivement 3,23% et 2,07% de la variance totale, ce qui est relativement faible en soit, mais largement supérieur aux dernières composantes.

3.2 Choix des composantes principales retenues

Pour interpréter et visualiser les résultats de l'analyse en composantes principales, il faut choisir le nombre de composantes à garder. Une limite arbitraire souvent fixée est de garder toutes les composantes dont la valeur propre est supérieure à 1. Ceci est visible sur la figure 2.

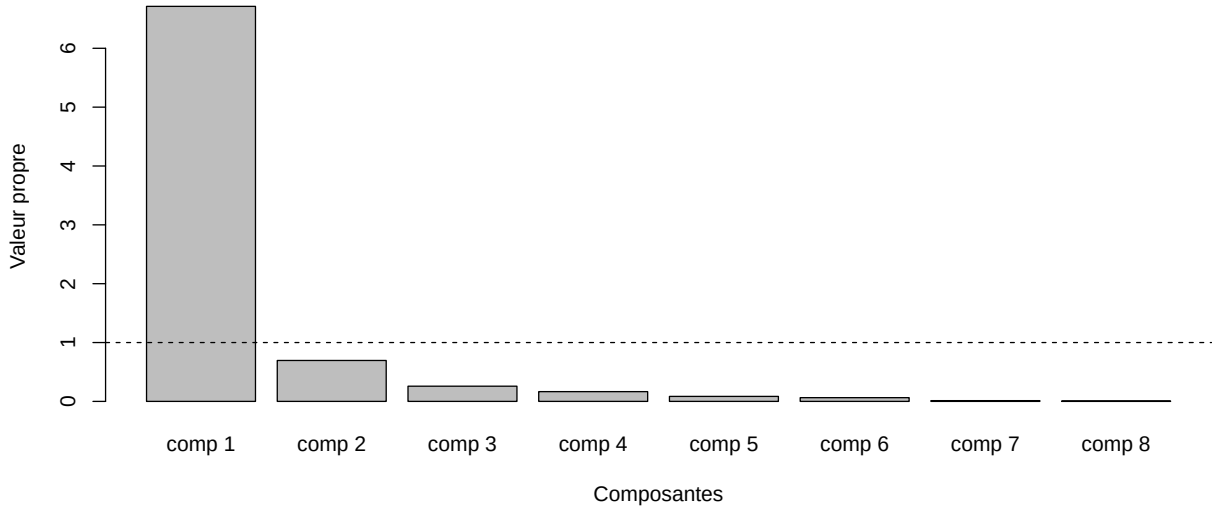


Figure 2: Composantes principales à garder

Seule une seule composante possède une valeur supérieur à 1. Cependant, il faut au minimum deux composantes afin de pouvoir visualier les résultats en 2D. Il semble dès lors logique de choisir la composante avec la seconde valeur la plus élevée, la composante numéro 2.

Il existe une autre méthode afin de choisir les bonnes composantes. Il est intéressant de choisir les composantes situées avant un "coude" dans le graphe des valeurs propres ou des pourcentages de variance. En effet, ce coude témoigne du fait que le gain d'information en passant à la composante suivante est relativement faible, et est dond relativement négligeable. Cette seconde méthode confirme l'utilité de choisir les deux premières composantes, comme le témoigne le coude présent sur la figure 3 à la troisième composante.

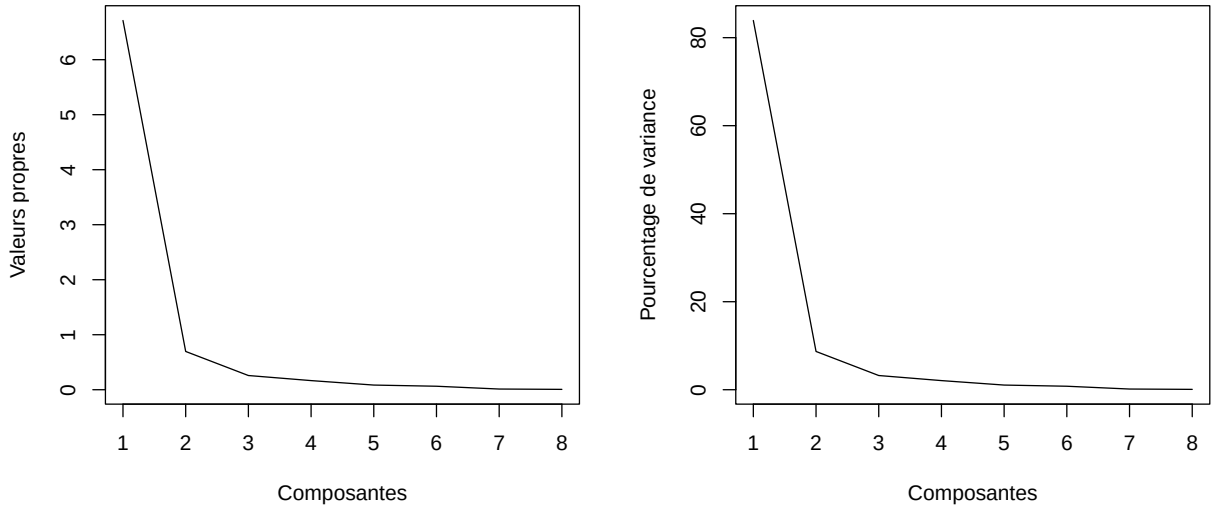


Figure 3: Valeurs propres et pourcentages de variance

3.3 Résultats et discussion

3.3.1 Analyse des variables

Le choix de garder deux composantes pincipales simplifie la représentation, mais a comme conséquence de ne représenter correctement que certaines variables. Les variables considérées comme mal représentées dans le plan choisi sont celles qui auront un angle de plus de 45° avec ce plan ($cos^2 > 0{,}5$).

Les valeurs ci-dessous représentent la qualité de représentation dans le premier plan factoriel. On observe que toutes les variables sont bien représentées à l'exception de la variable `Height` qui a un cos^2 inférieur à 0.9, toutefois cela reste une valeur acceptable. Concernant les contributions, toutes les variables sauf `Rings` contribuent avec une valeur de l'ordre de 15% dans la dimension 1, tandis que `Rings` contribue à 84 % dans la dimension 2 et uniquement 5% dans la dimension 1. Le détail de toutes les dimensions est présenté en annexe.

```
                Height ShellWeight VisceraWeight LongestShell Diameter
cos2 [-]         0.780       0.924         0.933        0.933    0.935
contribution [%] 12.058      13.952        14.080       14.315   16.198
                ShuckedWeight WholeWeight  Rings
cos2 [-]                0.944       0.974  0.986
contribution [%]       16.576      22.068 90.753
```

Finalement, le cercle de corrélation est présenté sur la figure 4. Cela permet de visualiser les résultats discutés ci-dessus. On observe que toutes les variables sont bien représentées. En effet, la longueur des flèches est proche de 1, correspondant à cos^2, sauf celle de la variable `Height`. De plus, comparé aux autres variables, `Rings` est la seule dont la représentation est de bonne qualité dans la dimension 2. Il est à noter que le pourcentage de variance capté par les diférents axes est visible le long de ceux-ci. Une figure de plus grande taille est prensentée en annexe pour plus de lisibilité.

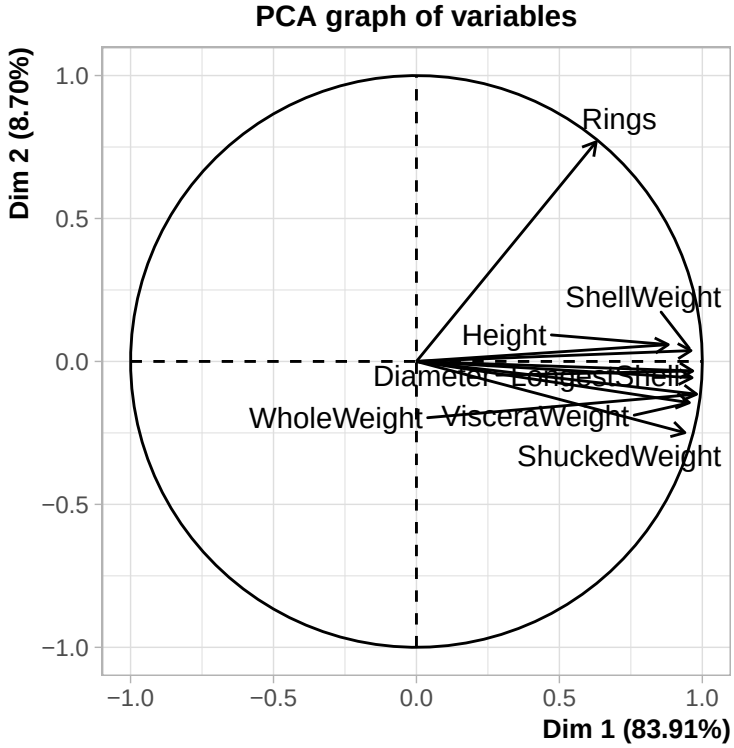


Figure 4: Cercle de corrélation

3.3.2 Analyse des individus

Etant donné le nombre important d'individus (>4000), il n'est pas pertinent de présenter les résultats concernant la qualité de représentation pour tous les individus. Une alternative est la représentation visuelle et son interprétation.

Pour le score cos^2 ainsi que pour la contribution, les tendances sont les mêmes au sein des trois catégories. Pour cos^2, la majorité des individus sont très bien représentés dans le premier plan factoriel même si certains ont des valeurs inférieures à $0,5$. Concernant la contribution, celle-ci est relativement faible avec une valeur moyenne inférieure à 10%. Certains individus sont mieux représentés avec des valeurs atteignant 75%, et pour le type `infant`, cette valeur maximale vaut 50%.

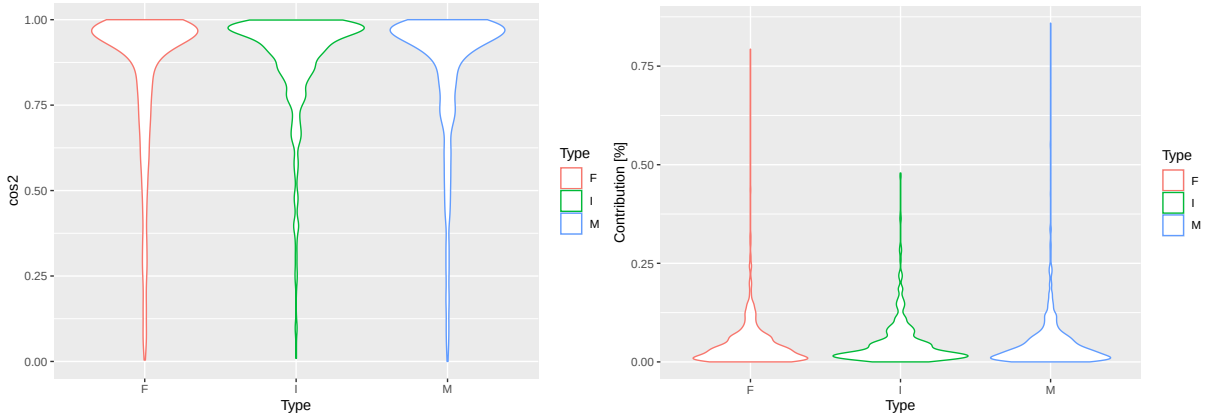


Figure 5: Distribution du score cos^2 ainsi que la contribution pour le premier plan factoriel pour les individus

La carte des individus mise en comparaison avec le cercle des corrélations s'avère utile pour visualiser les liens possibles entre les individus et les variables (représentées dans le cercle des corrélations). Ainsi, cette section s'efforcera de mettre en parallèle les résultats obtenus avec des explications logiques détaillant les diverses corrélations plus ou moins marquées.

Pour commencer, il est utile de préciser que l'ormeau garde sa coquille toute sa vie, mais il y a une phase de latence durant son stade « I » Infant. De plus, durant ce stade, comme la coquille est « neuve », il n'y pas de cernes, cela explique la corrélation inverse voir l'anticorrélation entre la catégorie « I » dans le nuage d'individus par rapport à la variables Rings dans le cercle des corrélations (les points de couleur verte se trouvent pour la majorité dans le sens inverse de la flèche). Dans une moindre mesure, cette tendance peut se répliquer aux autres variables. En effet, cela semble évident car durant les premiers stades de développement, l'ormeau possède des mesures physiques relativement faibles (petit poids, petite coquille, petit diamètre).

En outre, il est possible de remarquer une correlation marquée entre ShellWeight, Height, LongestShell, Diameter, WholeWeight, VisceraWeight, ShuckedWeight car toutes ces variables évoluent de la même manière lorsque l'individu grandit. Cela peut être visualisé en comparant le nuage d'invividus et le cercle des corrélations. En effet, les points appartenant à la classe "F" et "M" (individus d'un âge plus avancé que ceux de la classe "I") se placent dans le même quadrant et dans la même dimension que les variables précédemment citées. La variable rings quant à elle est moins groupée aux autres variables précédemment citées, cela peut être lié au fait que la coquille grandit moins vite que le poids par exemple, ou encore que la formation des anneaux sur la coquille soit plus longue que la croissance du mollusque. Néanmoins, la tendance visible sur le nuage d'individus (l'enchainement des individus se fait des individus "I" vers "F" et "M") semble suivre la dynamique de la variable rings dans le cercle des corrélation (le tracé global des individus est dirigé vers le quadrant en haut à droite tout comme la variable rings), cela confirme donc que la croissance (et particulièrement l'âge) évolue en synergie avec le nombre d'anneaux. Cette tendance a d'ailleurs déjà été expliquée lors de l'interprétation de la matrice des corrélations dans la section précédente.

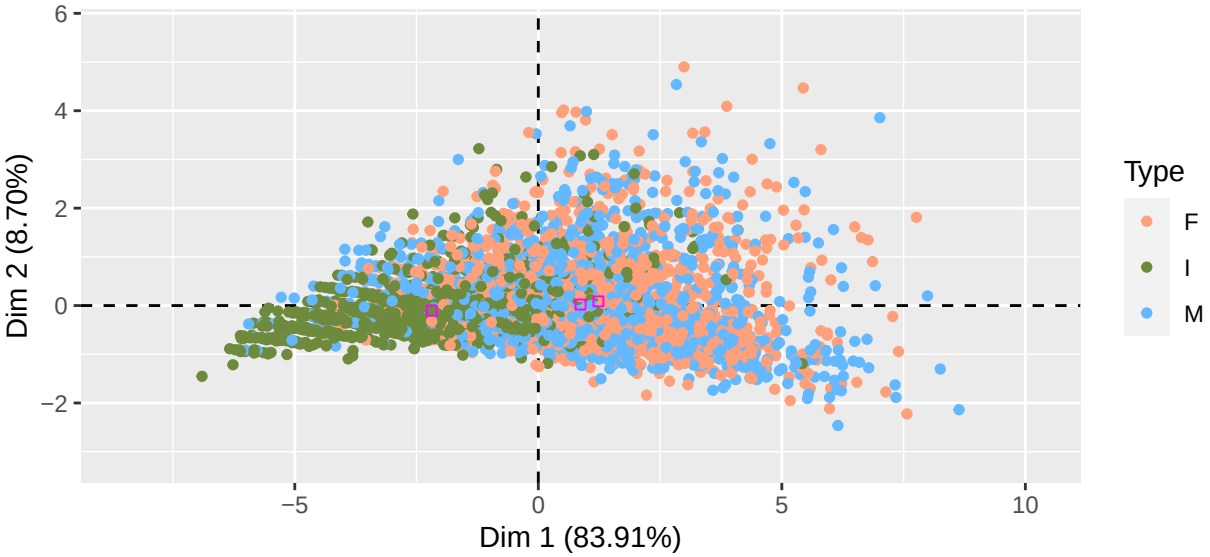


Figure 6: Carte des individus

4 Analyse de classification : Analyse discriminante linéaire

Dans le cadre de ce projet, la classification est réalisée à l'aide d'une Analyse discriminante linéaire (LDA). Cependant, il aurait été possible de faire usage de la méthode K-means ou encore d'une méthode de classification hiérarchique (hierarchical cluster analysis). Ces deux méthodes permettent de créer de nouvelles classes, dans le cas présent, celles-ci auraient pu être des distinctions plus poussées en fonction des caractéristiques physiques de l'individu.

Néanmoins, l'objectif principal de ce projet est de prédire l'âge de l'ormeau en fonction de ses caractéristiques physiques. En effet, comme cela a déjà été explicité dans l'introduction, coupler l'âge aux caractéristiques environnementales (non traitées dans le cadre du projet) permet de connaître la répartition des individus par écosystème étudié et la dynamique des populations d'ormeaux. Par conséquent, l'utilisation d'une LDA parait être le choix le plus pertinent car celle-ci permettra de prédire l'âge sur base des caractéristiques physiques, mais aussi de donner les variables les plus déterminantes pour la séparation des classes d'individus.

4.1 Analyse discriminante linéaire

L'analyse discriminante linéaire décrit des tendances similaires à l'ACP, car le premier axe canonique capte 98% de l'inertie totale. Il est également possible de constater que certain individus semble être relativement éloignés de "la masse". Cela peut être causé par des erreurs de prise de mesures sur le terrain, ou à l'inverse, que ces individus présentent des caractéristiques phyiques singulières.

De manière qualitative, cette classification a permis de différencier les individus plus efficacement que lors de l'analyse en composante principale, néanmoins, celle-ci est loin d'être optimale. En effet, il est possible de voir que les *infant* se différencient mieux des mâles et des femelles, qui sont encore relativement mélangés. Ce phénomène peut s'expliquer par le fait que les mâles et femelles ont des caractéristiques physiques relativement semblabes comparé aux juvéniles. Cela semble en adéquation avec le but principal du modèle qui a pour but de différencier les différents âges des individus (les mâles et femelles peuvent être assimilés à un stade adulte comparé à la classe *infant*). Ainsi, il serait nécessaire de prendre en compte davantage de caractéristiques lors de la prise de mesures pour fournir un modèle qui différencie à la fois les âges et les différents sexes de manière plus optimale.

4.2 Prédiction

Dans cette section, le but est de prédire la classe d'une nouvelle observation en l'associant avec la classe dont le centre en coordonnées canoniques est le plus proche. La proximité des centres n'est pas le seul facteur influancant la classification car il faut tenir compte des probabilités à priori. Dans le cas présent, les probabilités sont relativement similaires car l'échantillon de base est équilibré (37% de mâles, 32% de

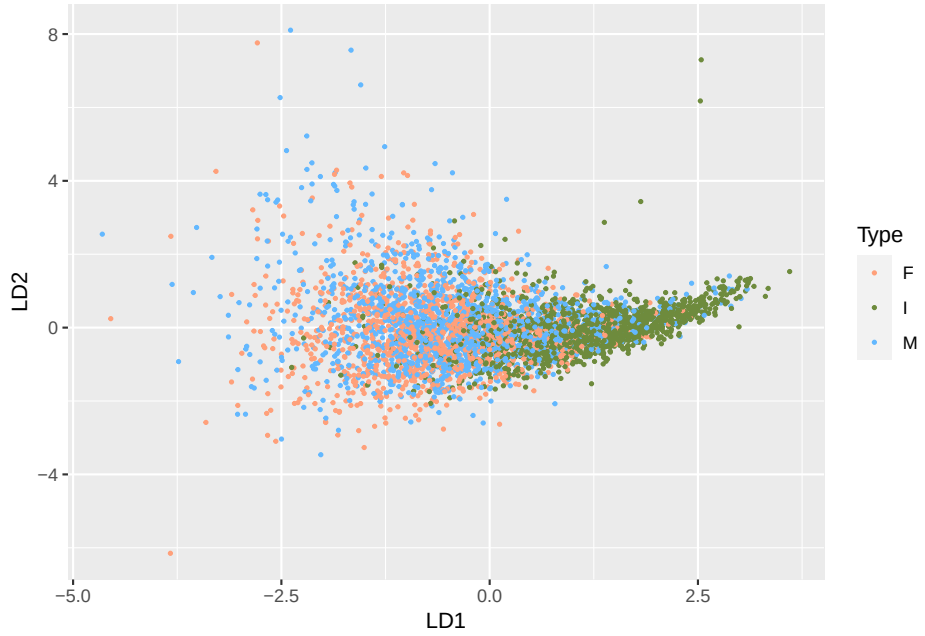


Figure 7: LDA - Variables canoniques

jeunes, 31% de femelles). Ainsi, ce facteur n'influencera que très peu la prédiction.

Ci-dessous, il est possible de voir les caractéristiques physiques choisies aléatoirement pour deux nouvelles observations (cela peut donc être répliqué avec d'autres dimensions).

```
    LongestShell Diameter Height WholeWeight ShuckedWeight VisceraWeight
new          0.4      0.3   0.09         0.5           0.2           0.1
    ShellWeight Rings
new        0.15    12
```

```
     LongestShell Diameter Height WholeWeight ShuckedWeight VisceraWeight
new2          0.8      0.6   0.15         0.7           0.5           0.3
     ShellWeight Rings
new2        0.45    22
```

```
           new
Prédiction "I"
```

Ainsi, pour cette première observation, la classe prédite en fonction des caractéristiques physiques aléatoires est celle des juvéniles (*infant*).

```
           new2
Prédiction "F"
```

En revanche, lorsque les dimensions sont modifiées légèrement, la classe prédite change. De ce fait, en ayant augmenté la valeur des caractéristiques physiques par rapport à la première nouvelle observation, il est possible de voir que la prédiction ne se porte plus sur la classe des juvéniles, mais sur celle des femelles (*females*).

4.3 Qualité

A la suite des prédictions, il semble pertinent d'étudier la qualité de la discrimination réalisée. Dans le cas présent, l'exactitude est calculée en fonction de la fréquence à laquelle le modèle répartit correctement les nouvelles observations dans les différentes classes.

Le tableau de contingence ci-dessous met en relation les "vraies valeurs" issues de la base de de données (lignes du tableau), avec les classes prédites (colonnes du tableau).

```
   Dataset.pred
      F    I    M
```

```
  F  445  189  673
  I   60 1032  250
  M  394  317  817
```

```
[1] 0.549198
```

Enfin, à l'issue de cette méthode de vérification, il est possible de voir que la qualité vaut 0.549198, ce qui signifie que 54.92% des individus ont été bien classé. Ce chiffre ne met pas forcément en exergue une classification optimale, mais cela peut s'expliquer par le fait que les individus (surtout les males et femelles) sont relativement mélangés, cela est d'ailleurs visible sur la figure de la LDA. Par conséquent, le modèle éprouve des difficultés lors de la classification, ce qui peut expliquer l'exactitude d'environ 55%. A l'avenir, il serait judicieux de revoir les critères de classification en modifiant ou en affinant les caractéristiques physiques dans le but d'accentuer la séparation entre individu, et ainsi augmenter l'exactitude de l'analyse discriminante linéaire.

5 Conclusion

Pour plus de lisibilité, cette conclusion est divisée en deux sections, une comprenant un rétrospective de l'ACP et une autre dirigée sur l'analyse discriminante linéaire.

5.1 Analyse en Composantes Principales (ACP):

Cette analyse a tout d'abord permis de mettre en relation la matrice de corrélation avec l'analyse descriptives des différentes composantes de la base de données, ainsi que d'autres outils visuels comme le cercle des corrélations. Le cercle des corrélations qui met en relation les différentes variables liées à l'analyse, n'a donc fait que confirmer les tendances visibles dans la matrice de corrélation. Ensuite,en mettant en parralèle le cercle des corrélations et la carte d'individus via le premier plan factoriel, il a été possible de relier les individus aux variables. Cette comparaison a mis en exergue les corrélations entre les mâles et femelles avec certaines caractéristiques physiques telles que les différents poids, les tailles, tout en montrant le phénomène inverse avec les ormeaux juvéniles. Cependant, en raison du nombre important d'observations, et par soucis de lisibilité, cette analyse s'avère simplifiée. En effet, seules les composantes les mieux représentées (composante 1 et 2 dans le cadre de ce projet) et qui par conséquent apportent le plus d'informations ont été retenu. De ce fait, à l'avenir, il serait intéressant de considérer plus de deux composantes pour en apprendre davantage sur les autres variables.

5.2 Analyse Discriminante Linéaire (LDA):

La LDA a permis de classifier de manière plus efficace les indivdus entre eux (comparé à l'ACP), mais elle n'est pas parfaite pour autant. En effet, la classe des juvéniles semble mieux se différencier que les deux autres. Diverses hyptohèses ont été émise quant à l'orignie de cette difficulté, mais celle qui semble la plus probable est que les données relevées sur le terrain avaient pour but principal de différencier le stade de vie des ormeaux, et non pas les spécificités qu'il pourrait y avoir entre les sexes.

Pour terminer ce rapport, il est pertinent de mentionner que les outils utilisés dans le cadre de ce projet s'avèrent être précieux pour l'analyse de données. Cependant, la qualité des représentations et de l'information récoltée sont fortement liées à la base de données initiale. Cet effet a notamment été expliqué et démontré dans le cadre de l'Analyse Discriminante Linéaire où la classification reposait sur les critères de base, parfois non optimaux pour la différentiation des classes. Outre la LDA, l'analyse de classification aurait pu être réalisé à l'aide d'autres outils comme le **K-means clustering** et le **hierarchical cluster analysis**. Cependant, ces outils nécessitent la création de nouveaux critères de classification nécessitant une connaissance plus profonde des caractéristiques intrinsèques des ormeaux. Ainsi, dans le cadre d'un projet plus conséquent, il serait utile de travailler conjointement avec des spécialistes pour affiner les critères, et ainsi perfectionner le modèle pour fournir une prédiction la plus optimale possible dans l'intérêt de tous.

6 Bibliographie

« Kaggle : Tout ce qu'il faut savoir sur cette plateforme ». Kaggle. 2021. Consulté le 2 décembre 2022 sur : https://datascientest.com/kaggle-tout-ce-quil-a-savoir-sur-cette-plateforme

« Ormeau ». UCI. 1995. Consulté le 2 décembre 2022 sur : https://archive-beta.ics.uci.edu/dataset/1/abalone

« Ormeau Haliotis tuberculata ». Muséum-Aquarium de Nancy. S.D. consulté le 5 décembre 2022 sur https://especeaquatique.museumaquariumdenancy.eu/fiche_espece/204

« Impact de l'acidification océanique sur la reproduction et le développement de l'ormeau Haliotis tuberculata». BOREA. 2017. Consulté le 4 décembre 2022 sur https://borea.mnhn.fr/fr/impact-l%E2%80%99acidification-oc%C3%A9anique-reproduction-d%C3%A9veloppement-l%E2%80%99ormeau-haliotis-tuberculata

Dans le cadre de ce rapport, certaines ressources sont tirées du cours **LSTAT2110 - Analyse des données** enseigné par Johan Segers, donné lors de l'année académique 2022-2023, ainsi que des travaux pratiques supervisés par Lise Léonard.

7 Annexe

7.1 Analyse descriptive

7.1.1 Détail de la fonction summary

```
 Type      LongestShell       Diameter          Height        WholeWeight
 F:1307   Min.   :0.075   Min.   :0.0550   Min.   :0.0000   Min.   :0.0020
 I:1342   1st Qu.:0.450   1st Qu.:0.3500   1st Qu.:0.1150   1st Qu.:0.4415
 M:1528   Median :0.545   Median :0.4250   Median :0.1400   Median :0.7995
          Mean   :0.524   Mean   :0.4079   Mean   :0.1395   Mean   :0.8287
          3rd Qu.:0.615   3rd Qu.:0.4800   3rd Qu.:0.1650   3rd Qu.:1.1530
          Max.   :0.815   Max.   :0.6500   Max.   :1.1300   Max.   :2.8255
 ShuckedWeight    VisceraWeight     ShellWeight         Rings
 Min.   :0.0010   Min.   :0.0005   Min.   :0.0015   Min.   : 1.000
 1st Qu.:0.1860   1st Qu.:0.0935   1st Qu.:0.1300   1st Qu.: 8.000
 Median :0.3360   Median :0.1710   Median :0.2340   Median : 9.000
 Mean   :0.3594   Mean   :0.1806   Mean   :0.2388   Mean   : 9.934
 3rd Qu.:0.5020   3rd Qu.:0.2530   3rd Qu.:0.3290   3rd Qu.:11.000
 Max.   :1.4880   Max.   :0.7600   Max.   :1.0050   Max.   :29.000
```

7.2 Analyse en composantes principales

7.2.1 Score pour cos^2

```
                  Dim.1       Dim.2        Dim.3        Dim.4        Dim.5
LongestShell  0.9295860 0.003243314 0.0002494918 0.0605664821 1.051356e-05
Diameter      0.9343663 0.001115656 0.0004345503 0.0573109036 2.848945e-04
Height        0.7760802 0.003453757 0.2092020069 0.0109444486 2.733858e-04
WholeWeight   0.9606611 0.013122027 0.0109879172 0.0097088981 2.082101e-05
ShuckedWeight 0.8818350 0.062121924 0.0112120853 0.0055737682 1.349191e-02
VisceraWeight 0.9117832 0.020813830 0.0100686034 0.0116760487 8.154263e-03
ShellWeight   0.9227010 0.001433795 0.0067469633 0.0098979354 5.860135e-02
Rings         0.3954264 0.590308665 0.0095415025 0.0003113583 4.112524e-03
```

7.2.2 Score pour la contribution

```
                  Dim.1      Dim.2       Dim.3      Dim.4       Dim.5
LongestShell  13.848707  0.4662527  0.09653645 36.4880652  0.01237622
Diameter      13.919922  0.1603846  0.16814155 34.5267533  0.33536857
Height        11.561821  0.4965055 80.94702091  6.5934448  0.32182095
WholeWeight   14.311655  1.8863977  4.25158045  5.8490917  0.02450982
ShuckedWeight 13.137326  8.9305299  4.33831832  3.3578971 15.88224146
VisceraWeight 13.583485  2.9921567  3.89586820  7.0341947  9.59893551
ShellWeight   13.746136  0.2061197  2.61061827  5.9629765 68.98361705
Rings          5.890949 84.8616534  3.69191583  0.1875767  4.84113042
```

7.2.3 Cercle de corrélation

7.3 LDA

```
Call:
lda(Type ~ ., data = abalone, method = "mle")

Prior probabilities of groups:
        F         I         M
0.3129040 0.3212832 0.3658128

Group means:
  LongestShell  Diameter    Height WholeWeight ShuckedWeight VisceraWeight
F    0.5790933 0.4547322 0.1580107   1.0465321     0.4461878    0.23068860
```

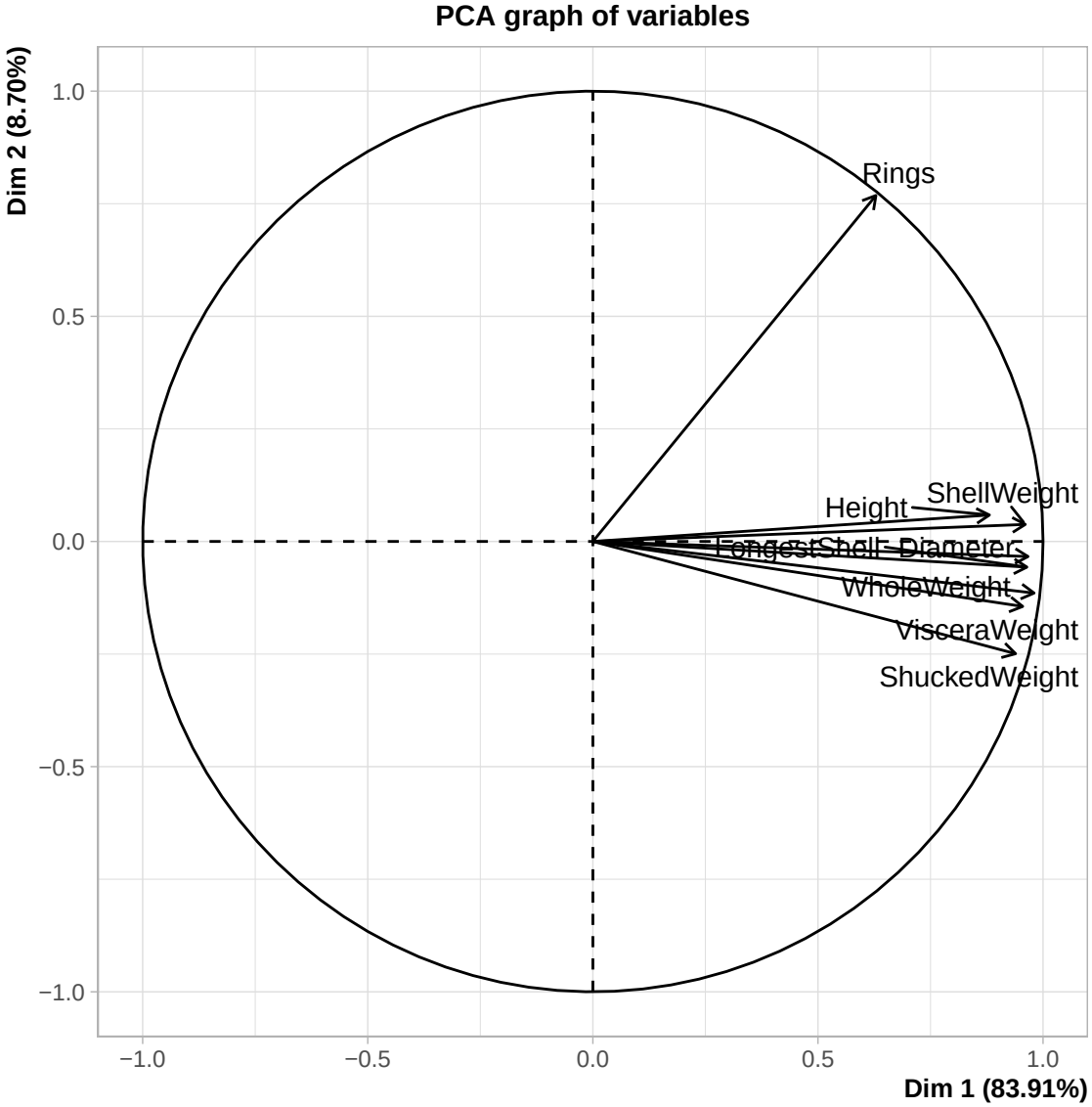


Figure 8: Cercle de corrélation

```
I    0.4277459 0.3264940 0.1079955   0.4313625     0.1910350    0.09201006
M    0.5613907 0.4392866 0.1513809   0.9914594     0.4329460    0.21554450
  ShellWeight     Rings
F   0.3020099 11.129304
I   0.1281822  7.890462
M   0.2819692 10.705497

Coefficients of linear discriminants:
                     LD1         LD2
LongestShell    5.6745367  -0.7582557
Diameter      -10.8752953  -9.5435257
Height         -4.5245127  -7.3608445
WholeWeight    -0.4281403  -1.4828022
ShuckedWeight   0.4788686  15.4342514
VisceraWeight  -5.4738338 -10.6774713
ShellWeight     1.7725261  -2.4309962
Rings          -0.1094434   0.1486725

Proportion of trace:
   LD1    LD2
0.9836 0.0164
```