Rapport de stage de M1 Mathématiques Appliquées Pour l'Ingénierie, l'Industrie et l'Innovation (MAPI3)

Analyse de données métabolomiques

Toulouse

26 août 2016

Stage effectué du 2 mai au 5 août conjointement à l'INRA, unité GenPHySE.
Maîtres de stage : Magali San Cristobal, Laurence Liaubet & Nathalie Villa-Vialaneix.

Camille Champion

Table des matières

Chapitre I

Environnement de travail

1 Établissement d'accueil

L'Institut National de la Recherche Agronomique est un organisme français de recherche public en agronomie. Il a pour but de développer des connaissances et des outils scientifiques autour des thématiques suivantes : agriculture, nutrition et environnement. L'INRA compte environ 400 unités de recherche réparties dans 19 centres et 14 départements de recherche (une unité correspond à un centre et à un département de rattachement).

L'unité de recherche GenPhySE (Génétique, Physiologie, et Systèmes d'Elevage), dans laquelle j'ai effectué mon stage, est née de la fusion de trois unités de recherche en 2014 sur le centre de Toulouse. Rattachée au département scientifique de Génétique Animale (GA), en partenariat avec le département de Physiologie Animale et Systèmes d'Elevage (PhASE), elle constitue un pôle de compétences majeur dans le domaine de la génétique animale, de la biologie intégrative et plus généralement des sciences animales.

Elle a pour mission d'étudier l'évolution et la gestion des populations animales telles que les palmipèdes, les lapins, les ovins ou encore les porcs. Les travaux de recherche des équipes qui la constituent sont notamment axés sur :

- la structure et l'organisation fonctionnelle du génome,
- la variabilité génétique des caractères d'intérêt,
- les mécanismes d'élaboration des phénotypes (adaptation, robustesse, résistance aux maladies, ...),
- la sélection génomique,
- les effets du milieu sur l'expression des gènes,
- l'évaluation et la conception de systèmes d'élevage plus durables.

2 Le projet PigHeaT

L'unité GenPhySE s'inscrit dans plusieurs projets dont le projet PigHeaT sur l'adaptation des porcs à la chaleur, sur lequel j'ai travaillé au cours de ce stage et dont une description est proposée dans les paragraphes suivants.

2.1 Objectifs

L'environnement climatique est un facteur majeur à prendre en compte dans la production porcine. En effet, il conditionne la croissance ainsi que la reproduction des porcs quelles que soient les régions dans lesquelles ils sont élevés (tempérées ou tropicales). Compte-tenu du réchauffement climatique actuel, il est important de chercher une solution pour améliorer la résistance des porcs soumis à ces nouvelles températures.

Financé par l'Agence Nationale de la Recherche (ANR) dans le cadre du programme BIO-ADAPT pour une durée de quatre ans, le projet PigHeaT a pour objectif d'identifier les régions chromosomiques liées à l'adaptation de la chaleur chez le porc et de comprendre les mécanismes physiologiques sous-jacents.

2.2 Description des données

Une expérience a été menée sur 1117 cochons issus du croisement des races « Large White » et « créole « . Ces animaux ont été soumis à deux types de stress :

- un premier groupe de 562 animaux a été élevé dans un milieu tempéré à 24°C. Il a ensuite été soumis à un choc thermique consistant à augmenter la température ambiante jusqu'à 30°C à partir de la $24^{\text{ème}}$ semaine. Des mesures ont été effectuées sur leur organisme avant, 48 heures après puis 2 semaines après ce changement,
- un deuxième groupe de 555 animaux a été élevé en Guadeloupe.

Pour mener à bien notre étude, les données phénotypiques suivantes ont été récoltées chez le porc :

- son identifiant : lorsque les porcelets quittent un site (naissance et/ou post-sevrage) pour un autre site, ils sont identifiés avec l'indicatif de marquage du site d'élevage (par exemple FR17MAG201310028),
- sa bande d'élevage : au début de l'expérience, des mâles de type Large White ou Créole ont été croisés avec des femelles dans chacune des deux régions, les cochons issus de ces croisements ont été répartis au sein de 21 bandes d'élevage au total,
- l'identifiant de ses parents,
- son sexe,
- la date de prélèvement sanguin du plasma.

3 Métabolome

Afin de mieux comprendre l'impact du stress thermique sur le métabolisme des cochons, des prélèvements sanguins ont été effectués sur chaque animal à la seringue puis centrifugés de manière à séparer les globules rouges du plasma. Ce dernier a ensuite été récupéré et stocké pour être analysé et permettre notamment d'avoir connaissance des modifications métaboliques chez ces cochons qui pourraient être liées à leur environnement.

Les métabolites, tels que le glucose, sont des composés chimiques issus du métabolisme cellulaire de l'organisme, qui participent à la croissance et au maintien des fonctions de l'organisme. La collection de métabolites synthétisés par le système biologique constitue son métabolome. De manière générale, le métabolome est homogène entre les individus grâce au processus d'homéostasie (autorégulation). Cependant, il peut s'avérer que d'un individu à l'autre, quelques modifications métaboliques s'opèrent du fait de la variabilité individuelle, notamment génétique, des modifications physiologiques ou pathologiques ou encore du stress auquel il est soumis.

4 Objectifs du stage

L'objectif de ce stage consistait à mener à bien une analyse exploratoire et prédictive de données métabolomiques afin d'identifier et de quantifier les métabolites qui sont liés à la différentiation des profils métaboliques des cochons à la chaleur.

Il s'agissait ensuite de mettre en relation les données métabolomiques et phénotypiques (poids, températures corporelles, ..) observées sur ces animaux de manière à identifier les métabolites responsables des différentes réactions physiologiques face à la chaleur. De plus, le métabolome ayant un pouvoir prédictif non négligeable sur certains phénotypes importants pour la production porcine, cette analyse permettrait de construire des modèles prédictifs des phénotypes d'intérêt.

Chapitre II

Étude métabolomique

1 Principe général

Une étude métabolomique allie technologies analytiques sophistiquées (spectroscopie) et outils statistiques (méthodes statistiques multivariées, modèles linéaires) pour l'identification des métabolites significativement modifiés entre des groupes d'individus.

Elle se décompose en plusieurs étapes illustrées dans la figure II.1 :

1. traitement des échantillons biologiques issus des prélèvements sanguins : détection des déplacements des métabolites dans le plasma,
2. construction de spectres,
3. construction du tableau de données puis analyses statistiques,
4. détection des métabolites significativement différents.

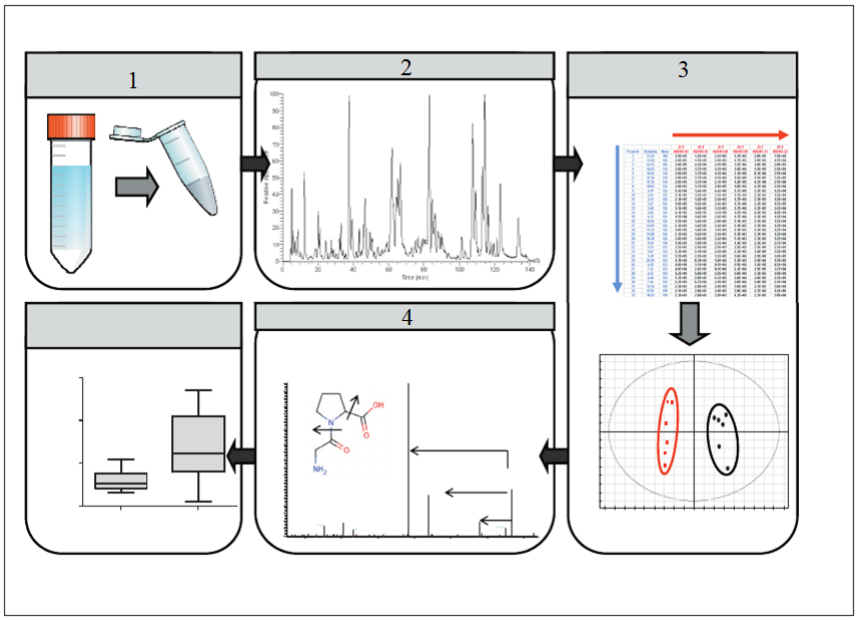

FIGURE II.1 – Étapes d'une étude métabolomique

1.1 Traitement des prélèvements sanguins

Afin d'avoir une vision globale des profils métaboliques des différents cochons, des données métaboliques ont été récoltées sur des échantillons de plasma grâce à la technologie de spectroscopie de Résonance Magnétique Nucléaire (RMN). Le principal avantage de cette méthode est sa capacité à analyser un très grand nombre d'échantillons. Elle consiste à calculer les déplacements chimiques des métabolites.

A chaque déplacement chimique est associé une intensité. Les données obtenues par la RMN se présentent sous la forme de spectres (graphiques représentatifs de l'intensité en fonction des déplacements chimiques), dont un exemple est présenté sur la figure II.2.

1.2 Construction de spectres

Un travail préalable a été effectué sur les données collectées afin que les spectres de chaque individu soient plus « lisibles » d'un point de vue graphique (voir II.2).

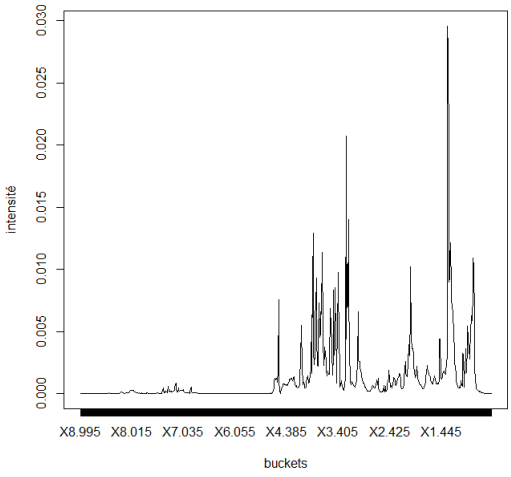

FIGURE II.2 – Profil métabolomique médian des individus

La technique de la RMN permet ainsi de produire un spectre caractérisant un mélange complexe, ce spectre étant la somme des spectres des métabolites présents dans le mélange. Chaque métabolite génère une résonance spectrale qui lui est propre avec une intensité proportionnelle à sa concentration dans le mélange. Le nombre de pics générés par un métabolite, comme leur localisation sur le spectre, est reproductible et uniquement déterminée : chaque métabolite possède sa signature sur le spectre.

Pour la suite de notre étude, il faut savoir qu'un unique spectre est associé à un cochon donné dans un environnement donné.

1.3 Analyse spectrale et statistique

L'exploitation des résultats des prélèvements sanguins pour l'analyse statistique nécessite une étape de pré-traitement des données, telle que le bucketing : le spectre est segmenté en régions consécutives non chevauchantes d'une largeur de 0.05 ppm. L'aire sous la courbe de chacune

de ces « tranches » est ensuite calculée et appelée bucket. On compte au total, 781 buckets par spectre.

Toutes les informations récoltées sur la quantité de métabolites peuvent être résumées sous la forme d'un tableau de données dont les lignes correspondent à tous les cochons et les colonnes aux différents buckets qui constituent leur spectre.

Afin de réaliser des analyses statistiques sur les jeux de données, deux fichiers m'ont été fournis, l'un contenant le tableau précédemment détaillé et l'autre contenant des informations sur la physiologie et le phénotype (identifiant de la mère et du père, bande, sexe,...) des cochons.

1.4 Détection des métabolites

En nous appuyant sur les données métaboliques collectées, nous souhaitons observer les réactions métaboliques des cochons soumis aux différentes températures, les comparer et détecter ceux significativement différents.

Dans le cadre de ce stage, nous nous sommes intéressés à l'analyse statistique des données (étape 3 de l'étude). Cette analyse vise à décrire, expliquer et prédire l'ensemble des observations (individus, déplacements chimiques) selon leur appartenance à des groupes prédéfinis (conditions de température). On distingue deux grandes approches que nous étudierons en détail :

- l'analyse exploratoire (2) permettant de détecter des valeurs aberrantes au sein de notre jeu de données ainsi que d'éventuels biais techniques dans les données qui seront à corriger,
- l'analyse discriminante prédictive linéaire ou non linéaire (3) utilisées pour expliquer les régions de températures par les données métabolomiques

2 Analyse statistique exploratoire

Comme énoncé précédemment, nous souhaitons dans un premier temps

- rechercher, identifier puis éventuellement supprimer des individus au profil métabolique atypique, pouvant être causé par des erreurs de mesure,
- rechercher puis ajuster des biais expérimentaux de manière à ne pas fausser les résultats à analyser.

grâce à une technique d'analyse statistique exploratoire. Parmi celles déjà existantes, nous avons choisi d'utiliser l'Analyse en Composantes Principales, développée par Saporta [G.S06].

2.1 Réduction de dimension par ACP

On considère qu'on a à disposition p observations $X^1, \cdots, X^p$ (correspondant aux $p = 781$ buckets) mesurées sur n échantillons ($n = 1117$).

Ces informations sont contenues dans la matrice X de taille $n \times p$ suivante :

	X^1	$\cdots$	X^j	$\cdots$	X^p
1	x_1^1	$\cdots$	x_1^j	$\cdots$	x_1^p
$\vdots$	$\vdots$		$\vdots$		$\vdots$
i	x_i^1	$\cdots$	x_i^j	$\cdots$	x_i^p
$\vdots$	$\vdots$		$\vdots$		$\vdots$
n	x_n^1	$\cdots$	x_n^j	$\cdots$	x_n^p

où $\forall i = 1, ..., n$, $\forall j = 1, ..., p : x_i^j = X^j(i)$ est la mesure de X^j sur le $i^{\text{ème}}$ individu.

Si le nombre p de variables est suffisamment petit (de l'ordre de 2 ou 3), l'observation de $E(X)$ et de $Var(X)$ permet d'avoir connaissance du comportement de chaque variable X^j ainsi que de leurs corrélations deux à deux. Dans le cas où p est grand, il sera difficile de tirer des conclusions d'une pareille démarche.

La technique de l'ACP consiste à résumer la variable p-dimensionnelle X par une variable unidimensionnelle C^1, fonction des composantes de X, et prenant en compte le maximum de l'information contenue dans X.

C^1, combinaison affine des $X^1, \cdots, X^p$ s'écrit sous la forme $C^1 = (X - E(X))^T a_1$ où $a_1 \in \mathbb{R}^p$ et est solution du problème d'optimisation suivant :

$$C^1 = \underset{C=(X-E(X))^T a, a\in\mathbb{R}^p}{\operatorname{argmax}} Var(C) \tag{II.1}$$

Le problème (II.1) admet une infinité de solution, on pose donc la contrainte supplémentaire $aa^T = 1$.

Proposition 2.1. *Les solutions de l'équation (II.1) sont données par $C^1 = (X - E(X))^T a_1$ où a_1 est un vecteur propre normé de la matrice $Var(X)$ associé à la plus grande valeur propre λ_1.*

En règle générale, une seule composante ne fournit pas un résumé suffisant de l'information contenue dans le tableau de données. Il convient alors de trouver d'autres combinaisons affines des $X^1, \cdots, X^p$, notées $C^2, \cdots, C^p$ non corrélées entre elles et de variance maximale. Celles-ci sont appelées composantes principales.

Proposition 2.2. *Les composantes principales $C^1, \cdots, C^p$ sont sans biais et vérifient $Var(C^j) = \lambda_j$ où λ_j désigne la valeur propre de la $j^{ème}$ composante.*

Remarque 1. *Les vecteurs a_j correspondant à la décomposition des composantes principales $C^1, \cdots, C^p$ sont appelées les vecteurs principaux et sont les vecteurs directeurs des axes principaux.*

Graphiquement, les individus ainsi que les variables, peuvent être représentés par des nuages de points dans des espaces de grande dimension, un point correspondant à un individu ou à une variable. La création de nouvelles composantes réduisant de la dimension des données simplifie alors considérablement leur visualisation. En effet, les nouvelles coordonnées obtenues pour chaque variable, appelées scores, peuvent aisément être visualisées sur un graphique à deux ou trois dimensions construit à l'aide des nouveaux axes. Ces nouveaux points forment des groupes en fonction de leur similarité qui peut être connue en étudiant les loadings, c'est-à-dire les coefficients de corrélations linéaires entre les variables d'origine et les facteurs (composantes principales).

2.2 Mise en œuvre : choix du nombre de composantes

Une étape déterminante de l'ACP est le choix du nombre de composantes principales. Il existe plusieurs techniques pour résoudre ce problème. Nous avons choisi une de ces méthodes qui était déjà implémentée dans le package FactoMineR et Factoextra utilisés pour l'ACP durant ce stage.

Cette méthode est basée sur le pourcentage de variance (ou d'inertie) reproduite. Celui-ci mesure la part d'information résumée par les composantes (voir le tableau II.1 résumant les informations pour les 10 premières) composantes. En pratique, on définit un seuil au-dessus duquel le pourcentage de variance cumulée est considéré comme acceptable ($\alpha = 80\%$), ce qui nous donne dans l'exemple suivant trois composantes principales.

	Pourcentage de variance reproduite	Pourcentage de variance cumulée
comp1	39.7	39.7
comp2	20.2	59.9
comp3	17.5	77.5
comp4	8.5	86
comp5	3.4	89.4
comp6	2.6	92
comp7	1.9	93.9
comp8	1.4	95.3
comp9	0.9	96.2
comp10	0.8	97

TABLE II.1 – Tableau récapitulatif des 10 premières composantes principales

En complément de cette méthode, la décroissance des valeurs propres, représentées sous la forme d'un diagramme en éboulis permet de visualiser le pourcentage de variance de chaque composante principale. Afin de sélectionner au mieux le nombre de composantes, il s'agit chercher s'il existe un coude dans le graphique et de ne conserver que les dimensions qui précèdent ce coude.

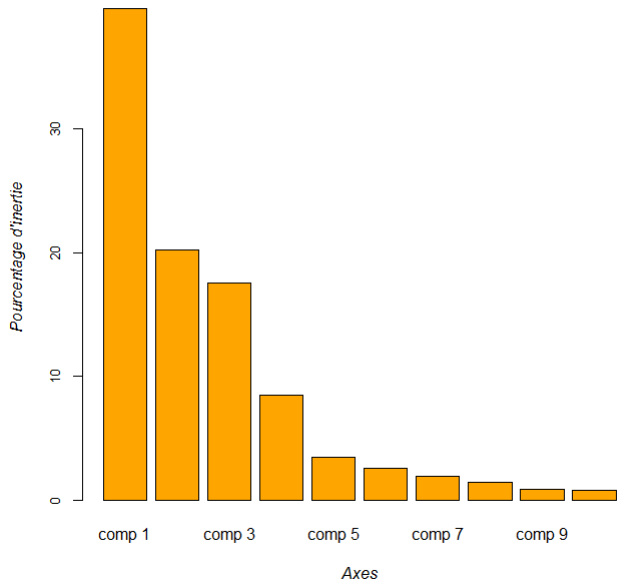

FIGURE II.3 – Éboulis des valeurs propres

Sur la figure II.3, le coude se forme à partir de la troisième composante. Nous conservons alors trois composantes principales pour toute notre analyse.

2.3 Mise en évidence d'individus atypiques

La figure II.4 représente la projection des individus sur les axes principaux obtenue par ACP.

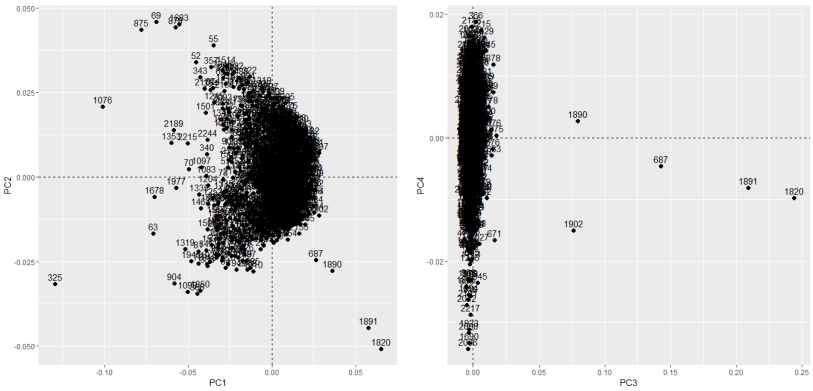

FIGURE II.4 – Projection des individus sur les axes de la 1ère et 2nde composantes à gauche (resp 2ème et 3ème à droite)

Notons qu'une dizaine d'individus s'écartent significativement des axes principaux lorsqu'ils sont projetés sur les trois premiers axes des composantes principales.

Le spectre de ces individus atypiques a une allure générale à peu près identique à celui des individus normaux, seuls quelques pics plus élevés se détachent du spectre médian (voirII.5 représentant la projection des individus sur les axes principaux obtenue par ACP). Cela confirme bien l'hypothèse de profils métaboliques anormaux, pouvant être dus à des erreurs de mesure. Nous les retirons de notre jeu de données.

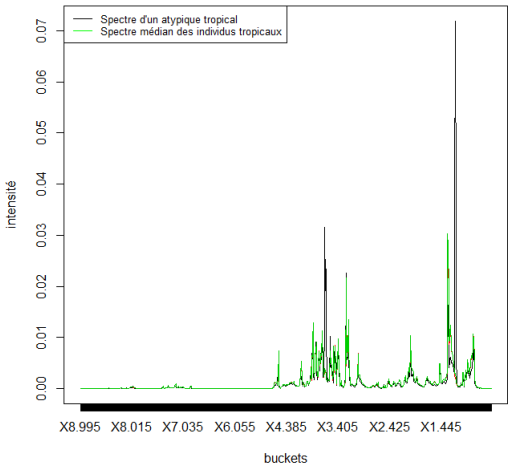

FIGURE II.5 – Spectre d'un atypique (noir) et médian des individus (vert)

2.4 Effets de certains facteurs extérieurs

Le métabolome reflétant le fonctionnement de l'organisme, nous pouvons nous attendre à ce qu'un certain nombre de caractéristiques physiologiques telles que la bande, la famille qui reflète des aspects génétiques ainsi que le sexe de chaque cochon aient des conséquences directes sur le profil métabolomique des individus.

Pour mesurer les effets de ces facteurs sur le métabolome, les individus sont projetés sur les axes principaux et, avec un jeu de couleur et la représentation d'ellipses placées autour des barycentres de chaque groupe d'individus ayant une caractéristique commune, nous faisons apparaître chaque facteur extérieur (voir figure II.6).

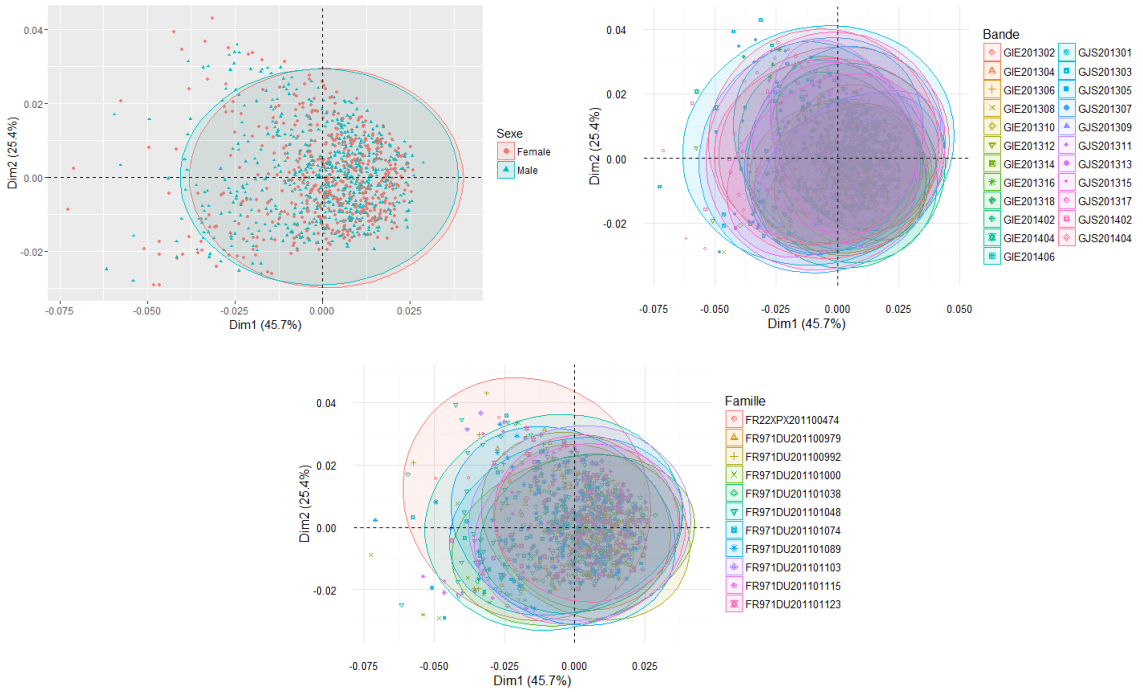

FIGURE II.6 – Influence des facteurs sexe, bande et famille sur le métabolome

Globalement, toutes les ellipses sont rassemblées autour d'un même point : l'intersection des axes principaux. Nous pouvons donc considérer que le sexe, la bande, la famille ou encore la température des cochons n'ont pas d'impact visible sur leurs profils métaboliques.

Les individus atypiques ayant été enlevés du jeu de données et les facteurs n'ayant d'influence sur le métabolome des individus, l'étape suivante consiste à identifier les variables contribuant à la différentiation entre les cochons provenant des régions tempérées ou tropicales grâce à des analyses prédictives.

3 Analyse statistique prédictive

Dans cette partie, nous allons voir deux types d'approches d'analyses discriminantes prédictives : linéaire et non linéaire. Toutes deux visent à construire un modèle prédictif pour expliquer les différents profils métabolomiques des cochons d'une région à l'autre. Pour cela, les buckets permettant de restituer au mieux cette classification sont sélectionnés et pourront être comparés selon les approches utilisées.

3.1 Prédiction par des méthodes d'analyses discriminantes linéaires

3.1.1 PLSDA

La régression des moindres carrés partiels (PLS), introduite par Wold [S.W01], est une technique récente qui généralise et combine les caractéristiques de l'ACP et de la régression multiple. Elle est notamment utilisée lorsque l'on veut prédire un ensemble de variables dépendants d'un nombre très grand de variables explicatives (prédicteurs) qui peuvent être fortement corrélées entre elles.

Dans cette partie, nous nous intéresserons à une extension de la régression PLS dédiée à la prédiction de variables préalablement classifiées : la PLSDA (Partial Least Square Analysis).

A la différence de la PLS, cette technique de classification nécessite des groupes d'appartenance de type qualitatives (classes) des différents objets qui composent le jeu de données. Elle prend donc en entrée des observations d'une variable qualitative (classe) à prédire, notée Y et codant l'appartenance des échantillons à une région (tempérée ou tropicale) ainsi que des observations de variables quantitatives $X = (X^1, \cdots, X^p)$. Étant donné que les classes sont au nombre de deux (tempérée ou tropicale), les valeurs de la variable Y sont remplacées par 1 pour les cochons élevés en région tropicale et -1 pour les autres cochons .

On suppose que X et Y peuvent être projetées sur un espace de dimensions réduites. Ces matrices sont alors décomposées en matrice de scores et de loadings selon les équations :

$$X = TP^T + E_X \tag{II.2}$$
$$Y = UQ^T + E_Y \tag{II.3}$$

où T et P représentent les scores et les loadings de X, U correspond aux scores de Y et Q^T aux poids de chaque classe, E_X et E_Y contiennent les résidus.

Le poids w_j de chaque variable X_j est mesuré par leur covariance et contenu dans la matrice W. Cette dernière est aussi utilisée pour calculer T. On définit alors la matrice W^* obtenue selon la formule :

$$T = XW^* \tag{II.4}$$

Les scores de X étant de bons prédicteurs de Y, on a également :

$$Y = TQ^T + G \tag{II.5}$$

Les équations (II.2), (II.3), (II.4) et (II.5) peuvent être combinées de manière à donner :

$$Y = XB + E_Y$$

avec

$$B = W^*Q^T$$

Les coefficients de cette matrice permettent de prédire la valeur de Y par des nouveaux échantillons n'ayant pas servi.

Afin d'estimer la capacité prédictive des variables, la méthode la plus courante est l'analyse de l'importance des variables dans la projection (Variable Importance in Projection, VIP). Le score VIP de la i$^{\text{ème}}$ variable, noté VIP_i, pour le modèle avec K composantes principales est calculé en utilisant l'équation :

$$\text{VIP}_\text{i} = \sqrt{\frac{p \sum_{j=1}^{K} (b_{ij}^2 t_{ij}^T t_{ij}) w_{ij}^2}{\sum_{j=1}^{K} (b_{ij}^2 t_{ij}^T t_{ij})}}$$

où p est le nombre de variables dans le modèle, t_{ij} et b_{ij} sont les scores et coefficients de régression pour la $i^{\text{ème}}$ composante, w_{ij} est le poids pour la $i^{\text{ème}}$ variable et la $j^{\text{ème}}$ composante et w_j le vecteur poids pour cette composante. Plus cette valeur est élevée, plus la variable considérée est importante dans la construction du modèle. La valeur seuil de sélection des variables importantes est fixée à 1.

La PLSDA est ici utilisée dans un contexte où les variables explicatives sont nombreuses ce qui peut rendre l'interprétation des données difficile.

3.1.2 OPLSDA

Dans le but d'améliorer l'interprétation des données ainsi que la séparation des classes, [OJ07] ont proposé une extension de la PLSDA, appelée OPLSDA (Orthogonal Partial Least Square Discriminant Analysis). Contrairement à la PLSDA, les composantes principales construites pour réduire la dimension de l'espace sont de deux types : la première, dite non orthogonale contient la variabilité dans X qui est corrélée au jeu de réponse Y. Elle servira à la prédiction des classes. La seconde, dite orthogonale contient la partie restante.

De même que pour la PLSDA, le coefficient VIP donné par le package ropls permet de connaître le caractère discriminant de chaque variable initiale.

Nous souhaitons prédire à l'aide de la méthode OPLSDA les classes auxquelles appartiennent les individus dont les buckets les plus discriminants viennent d'être sélectionnés. La capacité prédictive du modèle créé va être mesurée par validation simple.

La validation simple désigne le processus qui permet de tester la capacité prédictive d'un modèle dans un échantillon test par rapport à la précision prédictive d'un échantillon d'apprentissage à partir duquel le modèle est créé. Les données sont tout d'abord scindées en deux groupes. Le premier groupe est affecté à l'échantillon d'apprentissage et la fraction restante à l'échantillon test.

Après avoir effectué une validation simple sur notre jeu de modèle, nous observons la répartition des variables ainsi que les individus sur les nouveaux axes créés (voir figure II.7 représentant les éléments contenus dans la matrice score et loading pour chacune des variables).

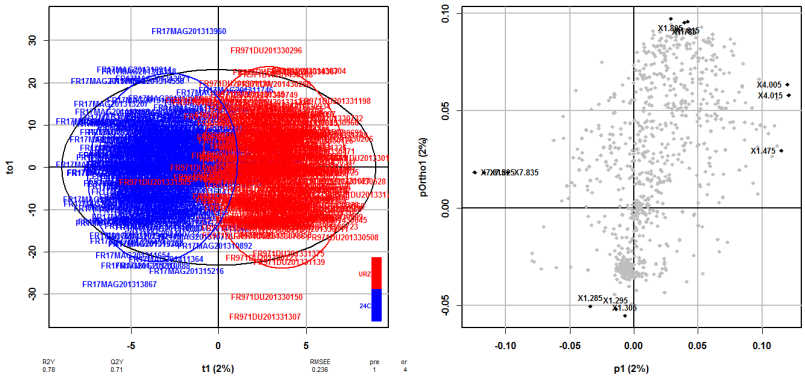

FIGURE II.7 – Projection des individus (à gauche) et des variables (à droite) dans le nouvel espace réduit

Notons sur la figure de gauche que la séparation des cochons provenant de Guadeloupe et des régions tempérées a bien été effectuée et que nous avons une première connaissance des buckets ayant eu un rôle important dans cette séparation grâce à la figure de droite. Afin d'en savoir un peu plus sur ces variables, nous utilisons le critère VIP (Variable In Projection) pour sélectionner les variables qui interviennent le plus dans la séparation des classes afin de prédire au mieux la région d'origine des cochons (voir figure II.8).

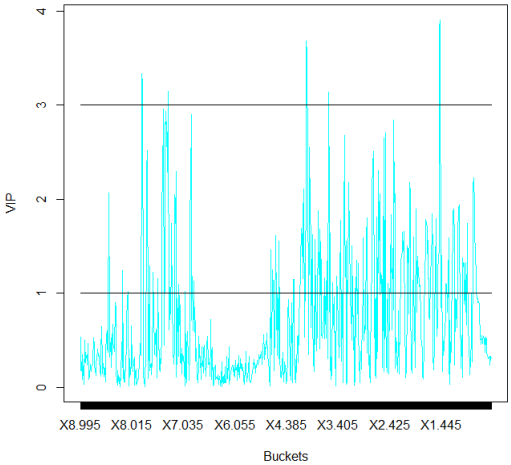

FIGURE II.8 – Graphique d'importance des variables par critère VIP

	24C	URZ
24C	272	5
URZ	3	272

TABLE II.2 – Tableau de contingence récapitulatif de la prédiction des classes

Les résultats de prédiction sont donnés sous la forme du tableau de contingence II.2. L'erreur de prédiction est de 0,01 ce qui reste très faible.

3.2 Arbres de décision et forêts aléatoires

Nous nous sommes également intéressés à la prédiction des classes par les Forêts Aléatoires, une méthode d'analyse discriminante non linéaire. Les Forêts Aléatoires ont été introduites par Breiman [L.B01] et sont basées sur l'agrégation d'arbres binaires, plus précisément décrits dans le paragraphe suivant.

3.2.1 Analyse discriminante non linéaire : les Forêts Aléatoires

Les forêts aléatoires sont la réunion d'une collection d'arbres aléatoires construits de la façon suivante : un nombre m de variables est tiré aléatoirement. Ce nombre est fixé au début de la construction et est identique pour tous les arbres. Considérons la racine de l'arbre associée à l'espace d'entrée $\mathbb{R}^p$ de notre modèle. Notre but va être de découper au mieux cette racine en deux nœuds fils de façon à améliorer l'homogénéité de Y. On cherche la variable X_i et la valeur d telles que :$\{Y_i : X_i < d\}$ et $\{Y_i : X_i > d\}$ soient de variance intra-groupe minimale.

Une fois la racine de l'arbre découpée, on se restreint à chaque nœud fils et on cherche suivant le même procédé, la meilleur façon de les découper en deux. Les arbres sont ainsi développés, jusqu'à atteindre une règle d'arrêt. Les arbres les plus redondants donnent les prédicteurs de notre modèle.

Supposons que l'on ait un échantillon d'apprentissage $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ constitué de n observations des variables explicatives $(X^1, \cdots, X^p)$ et de la variable à prédire Y. De même que pour la méthode OPLSDA, nous voulons sélectionner les variables $X^1, \cdots, X^p$ ayant le plus d'influence pour la classification de Y. Pour cette raison, on s'intéresse à la capacité prédictive des variables donnée par l'erreur OOB (Out-Of-Bag). Chaque arbre t de la forêt est construit sur une fraction IB (In-Bag) des données (c'est la fraction qui sert à l'apprentissage de l'algorithme). L'autre faction, appelée OOB(Out-Of-Bag) permet de mesurer l'erreur en terme de prédiction.

L'erreur OOB de la forêt est définie par :

$$\mathrm{E_F} = \frac{1}{n} \,\mathrm{Card}\left\{ i \in \{1, ..., n\} \,|Y_i \neq \hat{Y}_i \right\}$$

où $\hat{Y}_i$ est la classe la plus fréquente prédite par les arbres t pour lesquels i est OOB.

Deux critère permettent alors de connaître l'importance des données dans la classification :

- le critère de permutation,
- le critère de Gini.

Critère de permutation Fixons $j \in \{1, \cdots, p\}$ et calculons l'indice d'importance de la variable X^j. Notons E_{OOB_t}, l'erreur de OOB de classification d'un arbre t construit sur l'échantillon d'apprentissage X_{OOB_t}. Permutons alors aléatoirement les valeurs de la j$^{\text{ème}}$ variable X_{OOB_t}. Notons $\widetilde{X}^j_{OOB_t}$ ce nouvel échantillon et $\widetilde{E}^j_{OOB_t}$ son erreur associée. L'importance de la variable X^j est alors égale à :

$$VI(X^j) = \frac{1}{n}\sum_t (\widetilde{E}^j_{OOB_t} - E_{OOB_t})$$

Critère de Gini Une importance est calculée de manière similaire basée sur l'indice de Gini calculé de la façon suivante :

$$I_G = \sum_{c=1}^{\mathcal{K}} \hat{p_t}^c (1 - \hat{p_t}^c)$$

où $\hat{p_t}^c$ est la proportion d'observations de classe c dans le nœud t et $\mathcal{K}$ est le nombre de classes à notre disposition.

Après avoir mesuré l'importance des variables de notre modèle, nous avons représenté dans la figure II.9 les deux critères pour chaque variable considérée :

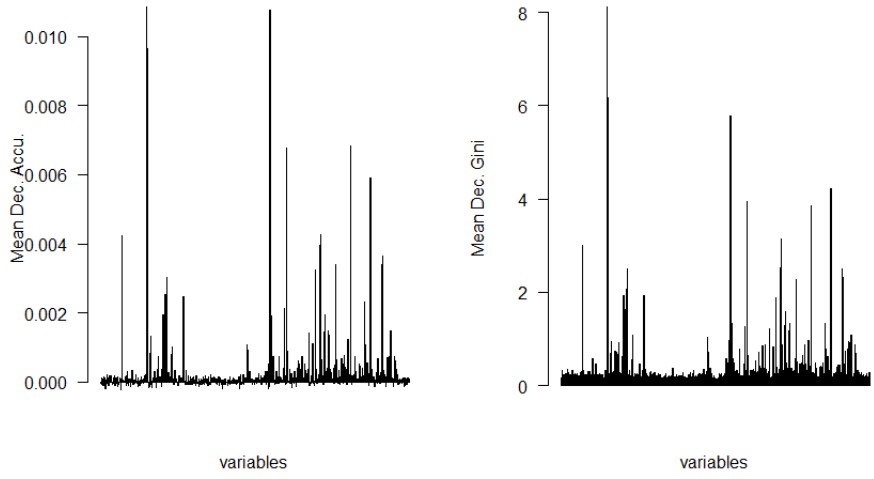

FIGURE II.9 – Importance des variables selon deux critères

De ces deux coefficients, nous pouvons en déduire les variables à sélectionner puis construire un tableau de contingence résumant la classification des données que nous avons reproduit dans le tableau II.3.

	24C	URZ
24C	488	67
URZ	54	496

TABLE II.3 – Tableau de contingence récapitulatif de la prédiction des classes

L'erreur de prédiction obtenue est faible (de l'ordre de 0,11), mais plus grande que celle de l'OPLSDA.

3.2.2 Méthode prédictive VSURF

Considérons une nouvelle méthode de prédiction se basant sur le forêts aléatoires : VSURF. Cette technique d'analyse, développée par [RTM10], a l'avantage de sélectionner automatiquement les variables les plus influentes pour la prédiction des classes. Cependant, ce processus de sélection est bien plus long que pour les autres méthodes.

VSURF se décompose en trois étapes consécutives :

1. Éliminer les variables les moins influentes
2. Parmi celles-ci, sélectionner celles qui sont les plus influentes même avec une forte redondance dans les arbres de classification
3. Trouver un échantillon plus petit de variables importantes avec le moins possible de redondance dans le but d'obtenir une meilleur prédiction des classes

Expliquons plus en détail ces étapes de sélection des variables :

- Dans un premier temps, les variables sont triées par ordre décroissant en fonction de leur taux d'importance appelé VI. Puis, elles sont ordonnées selon les écart-types de leur VI. Le seuil pour l'élimination des variables les moins importantes est donné par une estimation des écart-types des VI des variables les moins impliquées dans le modèle. Il s'agit de l'étape de seuillage.
- La deuxième étape (interp) de cette méthode vise à sélectionner toutes les variables qui permettent d'expliquer les deux classes de température. Elle repose sur la construction de modèles imbriqués de type Forêt Aléatoire à partir de la variable de plus forte importance dans un premier temps seulement, puis à partir des variables choisies par l'étape de seuillage. Les variables sélectionnées sont celles menant à la plus petite erreur OOB.
- Enfin, pour la dernière étape (pred), on considère dans un premier temps les variables sélectionnées par l'étape précédente. Une nouvelle fois, des modèles de type Forêts Aléatoires sont développés en intégrant cette fois-ci, pas à pas, les variables laissées de côté lors de l'étape précédente et en les testant au fur et à mesure du processus. Une valeur de saut moyenne (MeanJump) définie comme la l'écart moyen entre entre la moyenne des erreurs OOB d'un modèle et celle du modèle qui le suit. Ainsi, une variable est inclue dans le modèle si la moyenne des erreurs OOB calculée est plus grande que le produit du nombre de saut considéré et de sa valeur.

Le package VSURF donne la possibilité d'afficher quatre graphiques résumant les différentes étapes de la procédure de sélection.

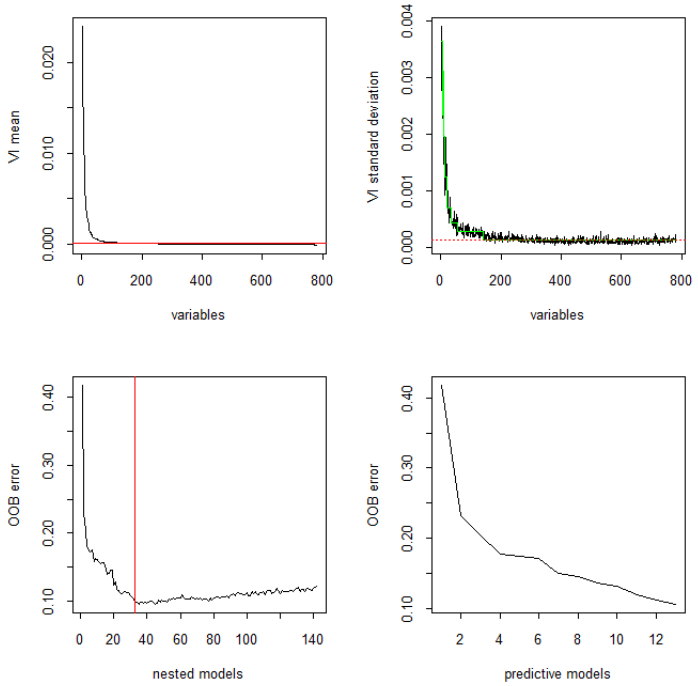

FIGURE II.10 – Étapes de sélection des variables par méthode

Le premier graphique de II.10 représente l'importance moyenne des variables rangées par ordre décroissant. La ligne rouge correspond à la valeur seuil de sélection.

Les écart-types de l'importance des variables sont tracés en noir sur le second graphique (en haut à droite) toujours dans l'ordre décroissant.

Le graphique en bas à gauche de II.10 correspond à l'étape d'interprétation. Il donne l'erreur OOB moyenne de toutes les forêts aléatoires créées. La ligne rouge verticale indique le modèle retenu.

Le dernier graphique de II.10 trace l'erreur OOB de toutes les forêts aléatoires créées à partir des variables ajoutées pas à pas. Le dernier modèle est celui qui a été choisi.

Après la dernière étape de sélection, 16 variables ont été sélectionnées. Nous pouvons alors prédire la région d'origine des individus à partir des déplacements chimiques sélectionnés.

Une nouvelle table de contingence est alors créée et donnée dans le tableau II.4.

	24C	URZ
24C	273	11
URZ	23	246

TABLE II.4 – Tableau de contingence récapitulatif de la prédiction des classes

L'erreur de prédiction obtenue faible (de l'ordre de 0,07).

Les trois méthodes présentées ci-dessus sont très fiables pour classifier les données selon chacun des critères utilisés. Nous voulons maintenant comparer les variables les plus discriminantes sélectionnées par les trois méthodes et comparer leurs rangs ou leurs valeurs d'importance selon les critères choisis.

3.3 Comparaison des méthodes de prédiction

Jusqu'à présent, 24 (OPLSDA), 18 (Forêts Aléatoires) et 15 (VSURF) variables discriminantes ont été collectées par les méthodes RandomForest et VSURF contre plusieurs centaines pour l'OPLSDA. Nous souhaitons donc diminuer le nombre de variables dites discriminantes désignées par cette méthode pour pouvoir les comparer à celles des autres méthodes. On choisit alors d'augmenter le seuil de VIP à 2,5.

Nous avons ensuite tracé un diagramme de Venn représentant les variables sélectionnées par chaque méthode. Ce dernier est donnée dans la figure II.11.

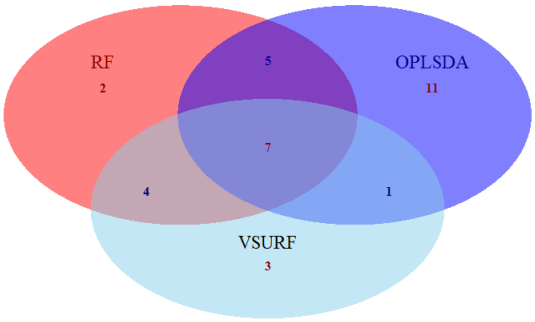

FIGURE II.11 – Comparaison des méthodes de sélection des variables par un diagramme de Venn

Parmi les variables les plus discriminantes, 7 ont été sélectionnées par toutes les méthodes.

Nous souhaitons savoir si ces variables sélectionnées permettent de séparer correctement les individus selon leur région d'origine. Pour cela, nous avons réalisé une ACP à partir de ces variables puis projeté les individus sur les composantes principales créées par la méthode. La figure II.12 illustre un tel résultat.

FIGURE II.12 – ACP à partir des variables sélectionnées en commun par chaque méthode

Maintenant que nous avons un premier aperçu des variables sélectionnées par les trois méthodes développées, nous voulons en savoir un peu plus sur les valeurs des critères qui nous ont amené à un tel résultat. Pour cela, nous avons réalisé trois graphiques permettant de mettre en évidence l'importance des buckets dans la prédiction des classes selon chacune des méthodes. Ces représentations sont données dans la figure II.13.

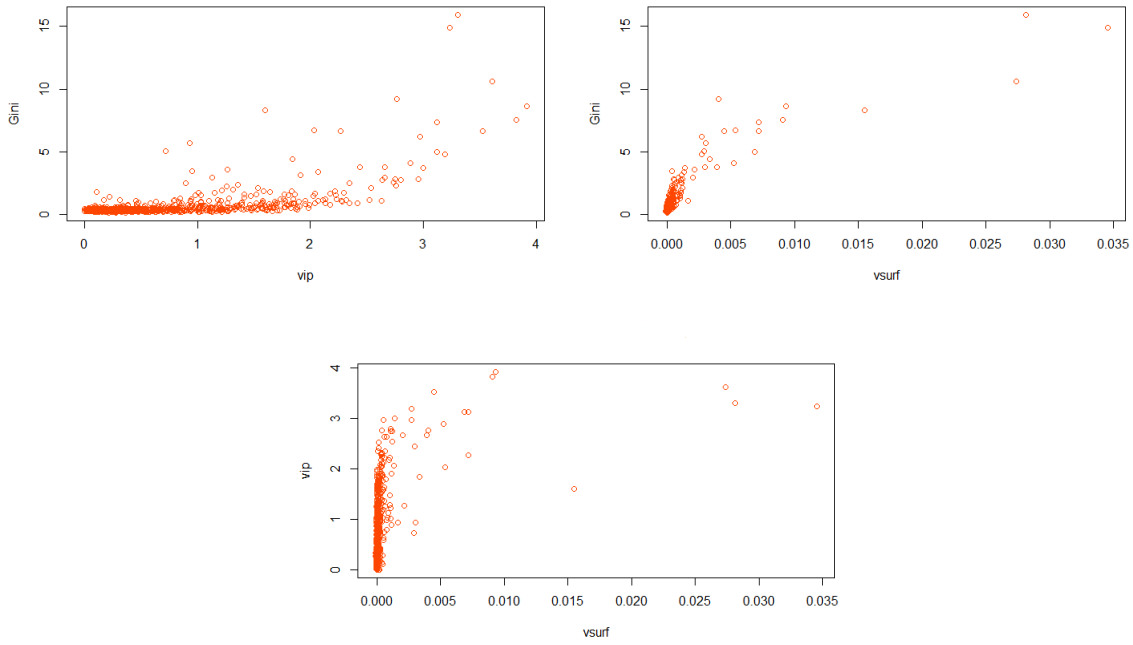

FIGURE II.13 – Comparaison critères de sélection VIP vs Gini (gauche), VSURF vs Gini (centre) et VSURF vs VIP (droite)

La première figure montre que les rangs, obtenus à partir du critère VIP et Gini, des variables les plus discriminantes sont à peu près les mêmes. Cela signifie que les variables utilisées ci-dessus pour la prédiction sont à peu près les mêmes d'une méthode à l'autre.

Dès lors, lorsque nous voudrons faire de la prédiction, nous pourrons opter pour une des trois méthodes présentées dans cette partie.

Chapitre III

Étude phénotypique

1 Présentation des données

Le métabolome influençant fortement les phénotypes des individus, nous souhaitons connaître plus en détails les conséquences des deux localisations géographiques sur la physiologie des animaux.

L'objectif est double :

- d'une part, d'explorer les phénotypes afin de mettre en évidence l'effet environnement sur la physiologie de l'animal,
- d'autre part, de prédire au mieux un phénotype donné à partir des données métaboliques. La qualité de prédiction de ces phénotypes peut avoir une importance pour la production de cochons, le rendement des éleveurs dépendant directement de la physiologie des animaux.

Pour cela, de nouvelles données phénotypiques ont été mises à disposition afin d'approfondir notre étude. Ceux-ci sont détaillés dans le tableau III.1.

	Semaine	Notation	Signification
Vitesse de croissance	11-13	CRS11S13	Vitesse de croissance entre la semaine 11 et 13
	13-15	CRS13S15	Vitesse de croissance entre la semaine 13 et 15
	15-17	CRS15S17	Vitesse de croissance entre la semaine 15 et 17
	17-19	CRS17S19	Vitesse de croissance entre la semaine 17 et 19
	19-21	CRS19S21	Vitesse de croissance entre la semaine 19 et 21
	21-23	CRS21S23	Vitesse de croissance entre la semaine 21 et 23
	11-23	CRS23	Vitesse de croissance entre la semaine 11 et 23
Poids	10	PDSs10	Poids de l'animal en semaine 10
	11	PDSs11	Poids de l'animal en semaine 11
	13	PDSs13	Poids de l'animal en semaine 13
	15	PDSs15	Poids de l'animal en semaine 15
	17	PDSs17	Poids de l'animal en semaine 17
	19	PDSs19	Poids de l'animal en semaine 19
	21	PDSs21	Poids de l'animal en semaine 21
	23	PDSs23	Poids de l'animal en semaine 23
Température corporelle	19	TMPs19	Température rectale de l'animal en semaine 19
	21	TMPs21	Température rectale de l'animal en semaine 21
	23	TMPs23	Température rectale de l'animal en semaine 23
	19-23	TMPmoyRECT	Moyenne des températures rectales
	19	TMPCUTs19	Température cutanée de l'animal en semaine 19
	23	TMPCUTs23	Température cutanée de l'animal en semaine 23
	19-23	TMPmoyCUT	Moyenne des températures cutanées
Épaisseur	19	EPs19	Épaisseur du lard de l'animal en semaine 19
	23	EPs23	Épaisseur du lard de l'animal en semaine 23
	19-23	EPgain	Gain d'épaisseur entre la semaine 19 et 23
Alimentation	19-23	CAMJ	Consommation Alimentaire Moyenne Journalière
	19-23	CAMJR	Consommation Alimentaire Moyenne Résiduelle
	19-23	FCR	Indice de consommation

TABLE III.1 – Tableau récapitulatif des phénotypes, de leur notation et de leur signification

2 Analyse exploratoire des phénotypes

Dans un premier temps, nous avons regroupé ces phénotypes selon les catégories présentées dans la première colonne du tableau III.1. Puis, nous avons étudié les corrélations entre les phénotypes d'une même catégorie au travers de graphiques. Nous y avons notamment fait apparaître les effets bande, région, sexe et famille.

Ces graphiques nous donnent ainsi une première idée de la répartition des individus selon les phénotypes.

Cas particulier du phénotype poids et vitesse de croissance :

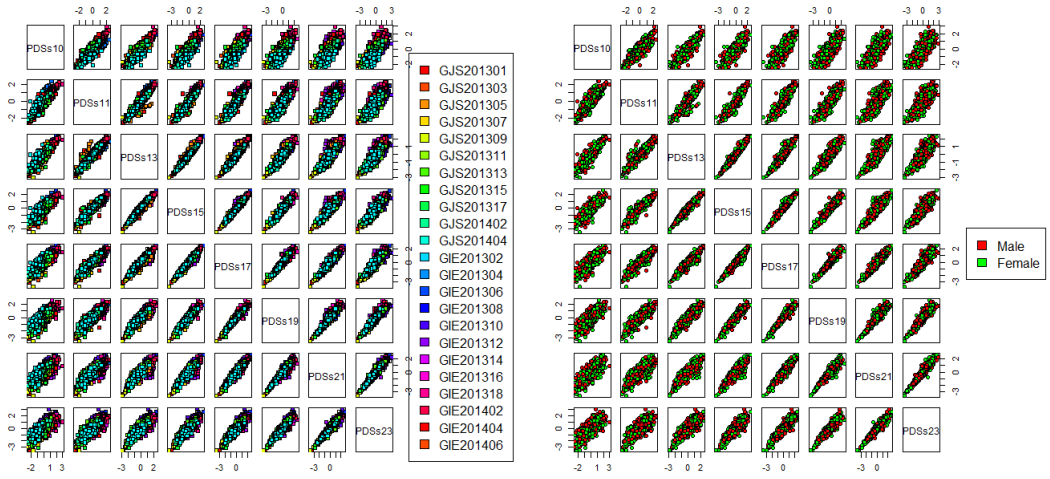

FIGURE III.1 – Corrélation phénotype poids avec effets bande-sexe

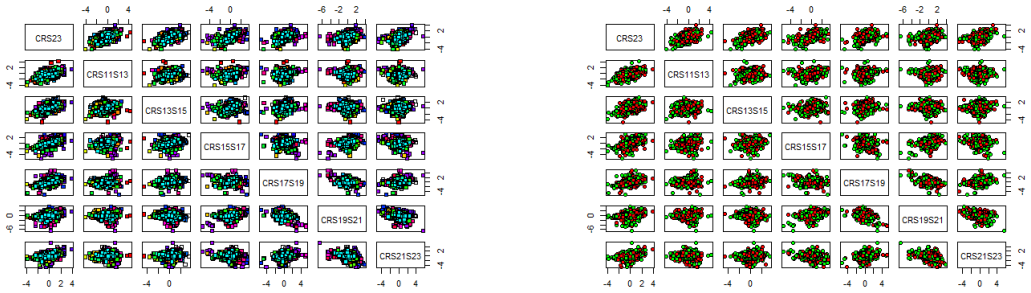

FIGURE III.2 – Corrélation phénotype vitesse de croissance avec effets bande-sexe

Le graphique III.1 nous permet de remarquer que les variables correspondantes au phénotype poids sont très corrélées. Par contre, dans le cas du phénotype vitesse de croissance (graphique III.2), les corrélations sont parfois négatives ou positives, ce qui peut se traduire par le fait que les pics de croissance des animaux ne se font pas au même moment.

Pour la suite de cette étude, les données phénotypiques sont tout d'abord centrées et réduites. Elles sont ensuite intégrées dans un nouveau tableau de données dont les lignes représentent les individus et les colonnes ces nouvelles variables.

Nous avons effectué une première Analyse en Composantes Principales en regroupant toutes les données phénotypiques de poids quelques soient les individus. Nous avons ensuite projeté les variables phénotypiques sur les deux premières composantes principales calculées par la méthode. Comme attendu, la première composante principale restitue la quasi totalité de l'information après réduction de dimension et les variables poids entre la semaine 10 et 23 se répartissent

uniformément autour du premier axe sur la figure III.3. En revanche, la vitesse de croissance d'une semaine à l'autre est assez irrégulière.

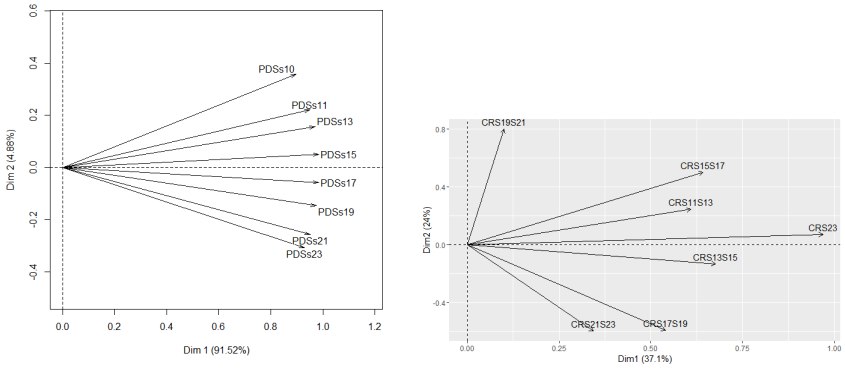

FIGURE III.3 – Répartition des phénotypes poids et vitesse de croissance

Cependant, de tels graphiques pourraient seulement souligner une courbe de poids et de vitesse de croissance différente d'un environnement à l'autre. Pour s'en assurer, nous avons projeté les individus sur les axes et nous avons tracé deux ellipses correspondantes aux de classes de température (tempérée ou tropicale).

FIGURE III.4 – Répartition des individus pour les phénotypes poids et vitesse de croissance dans les deux régions

Les deux ellipses tracées sur la figure III.4 se séparent assez bien.

Nous devons, dès à présent, distinguer les phénotypes pour des individus provenant d'une région ou de l'autre. Après les avoir séparés en deux catégories, nous avons effectué une nouvelle ACP sur les phénotypes poids et vitesse de croissance.

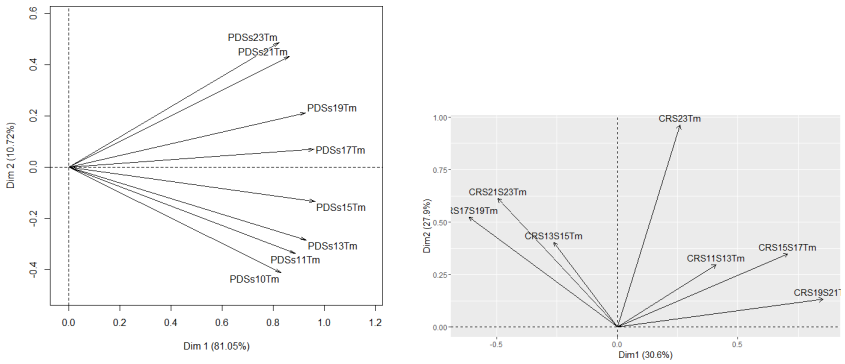

FIGURE III.5 – Répartition des phénotypes poids et vitesse de croissance en région tempérée

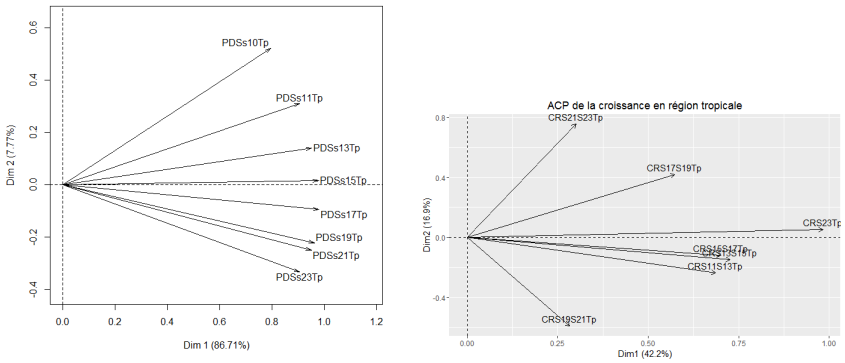

FIGURE III.6 – Répartition des phénotypes poids et vitesse de croissance en région tropicale

Nous retrouvons sur les figures III.5 et III.6 la même répartition des variables poids autour du premier axe. Par contre, la répartition des variables vitesses de croissance est assez différente d'une région à l'autre.

Observons enfin les courbes d'évolution de ces phénotypes dans les deux types de région (figure III.7).

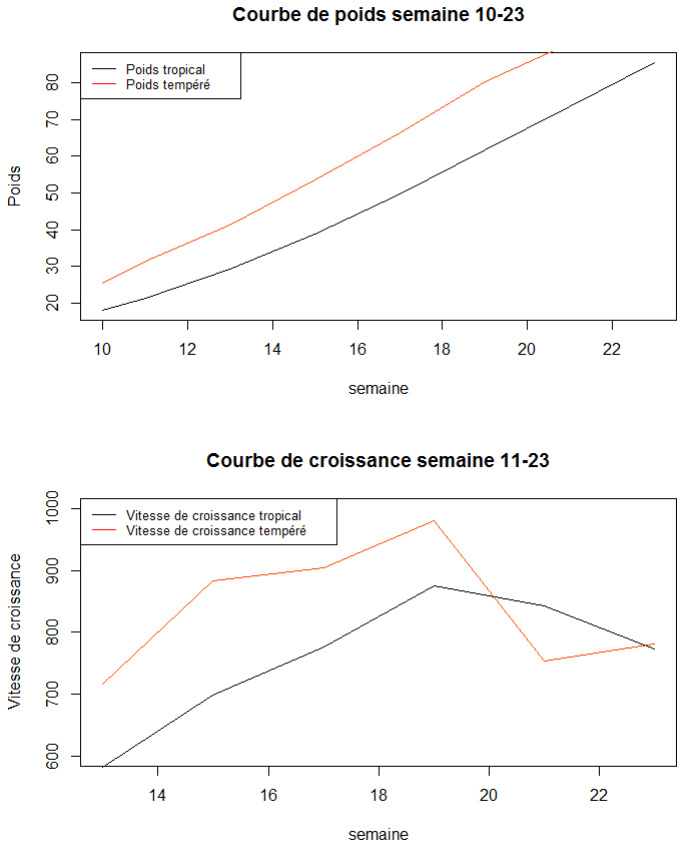

FIGURE III.7 – Courbes de poids et de vitesse de croissance dans les deux régions

Globalement, les cochons provenant des régions tempérées sont plus lourds et grandissent plus vite que ceux provenant des régions tropicales.

3 Relation données métabolomiques et phénotypiques

Nous souhaitons à présent identifier les buckets permettant d'expliquer les évolutions physiologiques des animaux d'une région à l'autre que nous avons détaillé ci-dessus.

Afin de mettre en relation les données métabolomiques et phénotypiques, une analyse en régression des moindres carrées de type SPLS a été effectuée.

Les résultats associés sont représentés sous la forme des figures III.8 et III.9. Ces deux figures mettent en évidence l'influence des buckets sélectionnés sur les différents phénotypes. Le code couleur donné en haut à gauche de chaque figure indique le degré de corrélation entre ces données.

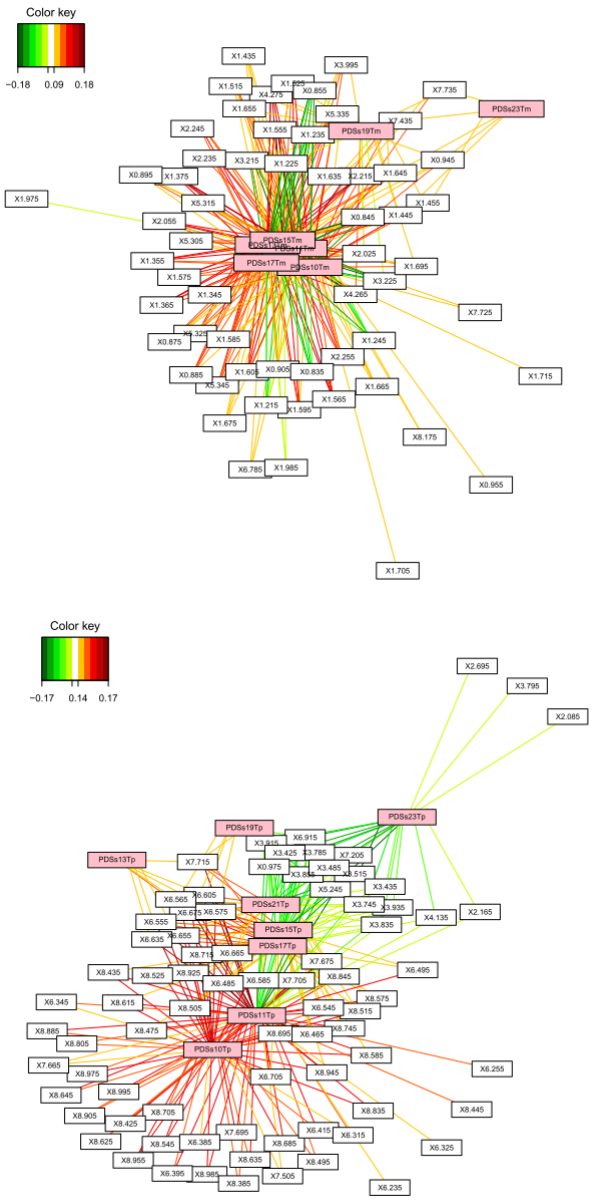

FIGURE III.8 – Corrélations données métabolomiques et phénotype poids

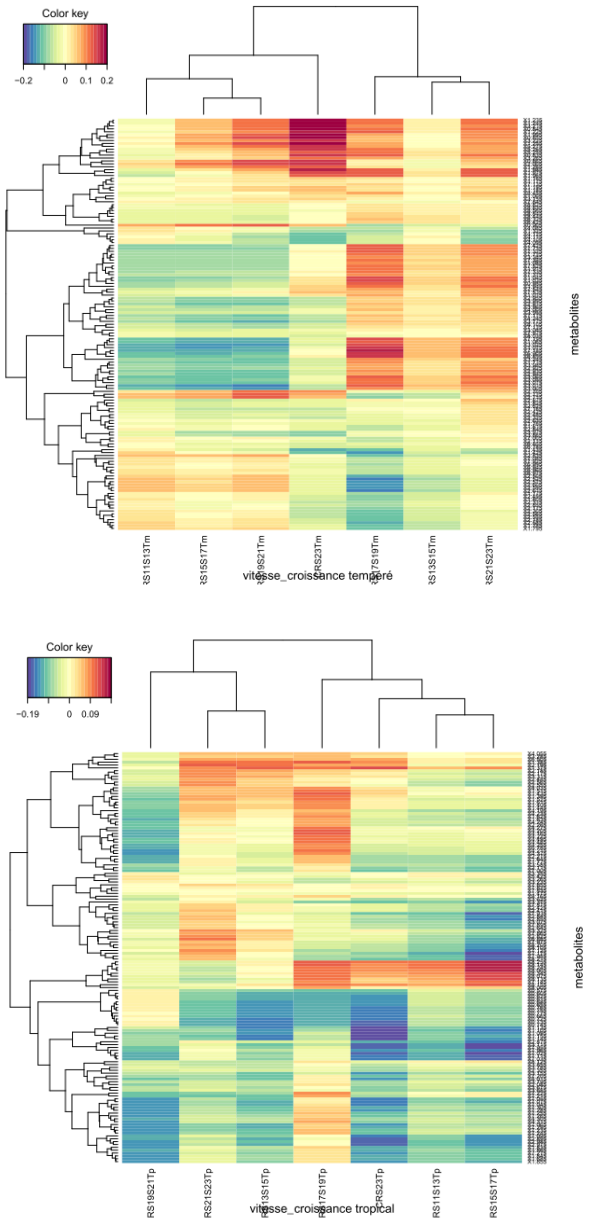

FIGURE III.9 – Corrélations données métabolomiques et phénotype croissance

Les corrélations entre données phénotypiques et métabolomiques étant trop faibles, ce résultat ne peut être pris en compte dans notre étude.

Chapitre IV

Conclusion

Les objectifs de ce stage étaient de mener à bien une étude exploratoire et prédictive sur des données métabolomiques et phénotypiques afin de mettre en évidence l'impact d'une hausse de température sur l'organisme des porcs. Les données métabolomiques et phénotypiques ont été préalablement nettoyées puis représentées sous la forme de spectres donnant ainsi accès aux profils métaboliques de chaque animal.

A l'aide de la méthode d'Analyse en Composantes Principales, ces profils ont été explorés. Les animaux dont le profil métabolique différait significativement ont été retirés de notre jeu de données. Les facteurs environnementaux et phénotypiques (température, bande d'élevage, sexe) impliquant la création de certains biais techniques ont été analysés puis corrigés.

Le deuxième travail effectué a consisté à prédire les conditions de température auxquelles ont été soumis les cochons seulement à partir de leurs données métabolomiques. Trois méthodes de prédiction ont pour cela été développées puis comparées.

Étant donné que le métabolome des cochons évolue au cours du temps, cela peut engendrer chez l'animal des modifications phénotypiques. Il s'agissait donc de déterminer les réactions phénotypiques des cochons soumis à la chaleur, puis de mettre en évidence les métabolites expliquant de telles différences.

Bibliographie

[G.S06] G.Saporta. Probabilités, analyse de données et statistiques. 2006.

[L.B01] L.Breiman. Random forests. 2001.

[OJ07] M.Rantalainen E.Holmes O.Cloarec, J.KNicholson and J.Trygg. Opls discriminant analysis : combining the strengths of pls-da and simca classification. 2007.

[RTM10] J.M Poggi R.Genuer and C. Tuleau-Malot. Variable selection using randomforests. 2010.

[S.W01] L.Eriksson S.Wold, M.Sjöström. Pls-regression : a basic tool of chemometrics. 2001.