Analyse de multiples blocs de données compositionnelles : application à l'étude de la sous-nutrition chronique

Antoine Ménard

AGROCAMPUS OUEST

☐ **CFR Angers**

☑ **CFR Rennes**

Année universitaire : 2018-2019 Spécialité : Agronomie…… Spécialisation (et option éventuelle) : Sciences des données……	**Mémoire de fin d'études** ☑ d'Ingénieur de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage ☐ de Master de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage ☐ d'un autre établissement (étudiant arrivé en M2)

Analyse de multiples blocs de données compositionnelles : application à l'étude de la sous-nutrition chronique

Par : Antoine MENARD

Soutenu à ***Rennes*** ***le*** ***4 septembre 2019***

Devant le jury composé de :

Président :François HUSSON

Maître de stage : Vincent GUILLEMOT

Enseignant référent : Francois Husson

Autres membres du jury (Nom, Qualité) :

David CAUSEUR, relecteur

Table des matières

1 Introduction

L'institut Pasteur est une fondation à but non lucratif dont l'objectif est de permettre et de développer la prévention et les traitements des maladies infectieuses. Pour y parvenir, la fondation a 4 missions principales : le développement de la recherche biomédicale, la santé des populations et des personnes, l'enseignement et l'innovation technologique. L'institut est présent dans le monde entier à travers le réseau International des Instituts Pasteur, qui comporte 32 établissements répartis sur tous les continents. Certains d'entre eux sont dans des zones sensibles d'un point de vue sanitaire où ils ont une mission de veille sanitaire, pour identifier le plus tôt possible les malades infectieuses émergentes. L'institut Pasteur Paris fait partie de ce réseau d'instituts et regroupe aujourd'hui plus de 2500 collaborateurs dans 130 unités de recherches différentes.

1.1 Le Hub de Bioinformatique Biostatistique

Créé en 2015, le Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI) est une structure qui a pour rôle le soutien des chercheurs de l'institut Pasteur pour le traitement, l'analyse et la modélisation des grandes quantités de données générées par les équipes de recherche présentes sur le campus. Le Hub de Bioinformatique et Biostatistique est la plate-forme de services du C3BI. Les ingénieurs du Hub apportent un soutien en bioinformatique et biostatistiques aux équipes de recherche du campus, participent au développement d'outils d'analyse de données et forment les personnels des Institut Pasteur à Paris. Il est aujourd'hui composé de 50 ingénieurs répartis en 7 groupes d'expertise. Mon tuteur de stage le Dr. Vincent Guillemot, rattaché au groupe d'expertise STATS, qui traite les questions liées aux statistiques.

1.2 Le projet Afribiota, cadre de mon stage

Aujourd'hui, un quart des enfants de la planète âgés de moins de 5 ans sont affectés par un retard de croissance. On détermine si un enfant est en retard de croissance en fonction de la valeur d'un Z-score calculé sur la base de sa taille rapportée à son âge –Height-for-Age Z-score (HAZ). Si la valeur du HAZ est plus de 2 écarts types en dessous de la valeur médiane de référence standard de l'OMS (définie en 2006), on considère que l'enfant est en retard de croissance. Cela a de nombreuses conséquences sur la santé et le développement des enfants à court et long terme : les risques de maladies augmentent, la mortalité augmente et on observe un retard de développement psychomoteur [Dewey and Begum (2011)]. Ce retard est souvent lié à une malnutrition chronique qui génère des carences en micro et macronutriments mais aussi aux infections répétées auxquelles sont exposés les jeunes enfants. Cependant, il est encore difficile de proposer des traitements efficaces car les causes et mécanismes complexes mis en jeu n'ont pas été complètement élucidés. C'est pour combler ce manque qu'a été mis en place le projet Afribiota, pour une compréhension plus fine des mécanismes pathophysiologiques responsables du retard de croissance [Vonaesch et al. (2018)].

L'hypothèse avancée par cette étude est que le retard de croissance est dû à un syndrome caractérisé par une inflammation chronique du petit intestin mais encore mal compris à ce jour : l'entéropathie environnementale pédiatrique (PEE) [Vonaesch et al. (2018)]. Cette PEE serait à l'origine d'une dysbiose, un déséquilibre ou une mauvaise adaptation de l'écosystème intestinal, qui comprend notamment le microbiote et les métabolites de ce milieu. Ce dérèglement altérerait les capacités de digestion des aliments ainsi que les capacités d'absorption des nutriments. L'objectif principal du projet est de détecter et décrire cette dysbiose en identifiant des facteurs de risques et des biomarqueurs du retard de croissance. Il est en effet important de mieux diagnostiquer la maladie pour pouvoir développer des traitements plus ciblés et efficaces que ceux actuellement disponibles.

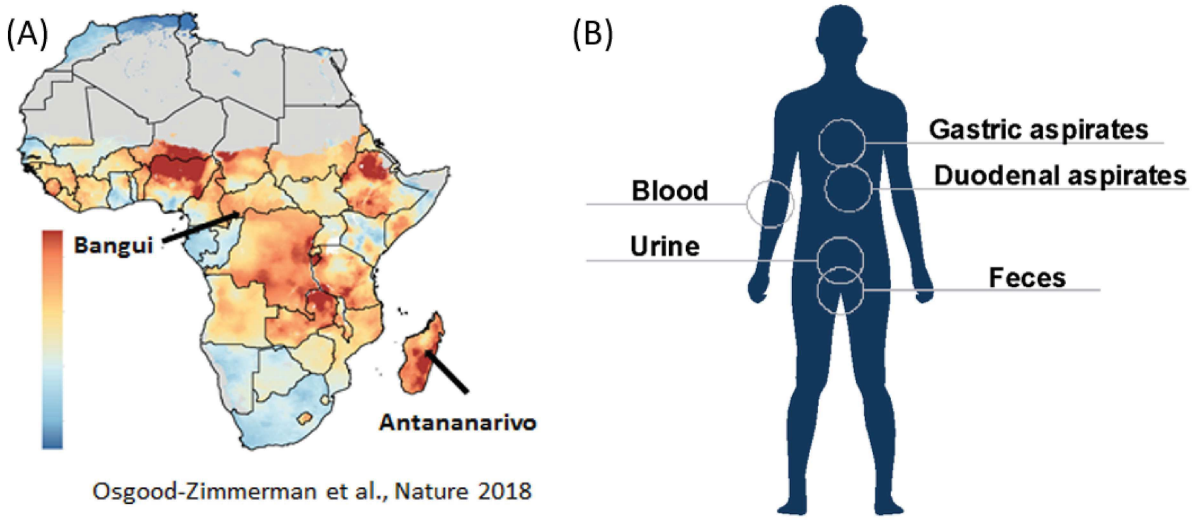

FIGURE 1 – (A) La malnutrition chez les enfants entre 0 et 5 ans en Afrique. (B) Prélèvements biologiques effectués sur les enfants. Les trois compartiments concernés par l'étude Afribiota sont indiqués sur la droite de la figure : ils concernent les aspirations gastriques, duodénales et les selles.

Deux pays ont été choisis comme cadre de l'étude : Madagascar (à Antananarivo) et la République Centrafricaine (à Bangui), où respectivement 49.2 % et 40.7 % des enfants de moins de 5 ans souffrent de retard de croissance [Global Nutrition Report (2018)], voir Figure 1(A). Pour détecter et identifier le phénomène de dysbiose chez les jeunes enfants, nous allons étudier la composition du microbiote intestinal de ces derniers ainsi que l'abondance de différents acides biliaires au sein de plusieurs compartiments du système digestif. Les prélèvements ont été réalisés au niveau de l'intestin grêle, de l'estomac et des selles (cf. Figure 1 (B)). Pour compléter ces observations et déterminer de nouveaux facteurs de risques nous disposons d'informations sur l'environnement socio-économique et les conditions de vie des enfants.

Lors de mon stage, mon rôle a été de participer à l'intégration de ces données dans l'étude du retard de croissance et de répondre à la question suivante : comment parvenir à intégrer ces blocs de données dans une méthode d'analyse commune pour mieux comprendre et identifier les liens entre la composition du microbiote intestinal, la composition intestinale en acides biliaires et le retard de croissance ?

2 Jeux de données et notations

2.1 Notations

Les matrices sont notées en lettres majuscules et en gras (e.g., $\mathbf{X}$), les vecteurs colonnes sont notés en lettres minuscules et en gras (e.g., $\mathbf{x}$), and leurs éléments en minuscules italiques (e.g., $x_{i,j}$). Les sous-matrices, vecteurs (colonnes ou lignes) et éléments d'une même matrice se réfèreront tous à la même lettre (e.g., $\mathbf{A}$, $\mathbf{a}$, a). L'opérateur de transposition est noté par l'exposant « $^\top$ », la matrice identité est notée $\mathbf{I}$, un vecteur ou une matrice de uns est noté $\mathbf{1}$.

2.2 Données de métabolomique : acides biliaires

Les acides biliaires sont des molécules qui jouent plusieurs rôles dans le bon fonctionnement de l'appareil intestinal. Sécrétés dans le foie, ils sont stockés dans la vésicule biliaire sous forme conjuguée puis sécrétés au moment du repas. On les retrouve dans la bile intestinale où ils agissent comme un détergent permettant la fragmentation des lipides, ce qui facilite la digestion. Ce sont aussi des agents antibactériens qui évitent la prolifération de certaines bactéries

intestinales. Une partie d'entre eux est transformée en acides biliaires secondaires par des bactéries intestinales. Puis, Au niveau de l'iléon, la majeure partie des acides biliaires sont réabsorbés et retransportés au niveau du foie, une partie d'entre eux sera éliminée dans l'urine et les selles [Poupon (2004)].

Chez les enfants avec un retard de croissance, on s'attend à observer un dérèglement des abondances des différents types d'acides biliaires. Notre objectif ici est de détecter des variations de la composition en acides biliaires entre les enfants normonutris et les enfants malnutris.

Les données concernant la composition en acides biliaires des échantillons sont obtenues par spectrométrie de masse couplée à la chromatographie en phase liquide [Han et al. (2015)]. Par cette méthode, on obtient l'abondance en acides biliaires dans chaque échantillon (protocole réalisé à l'Université de Victoria, Canada). Cependant, les quantités d'acides biliaires que l'on détecte sont dépendantes de facteurs difficilement contrôlables comme la durée de jeûne précédent le prélèvement ainsi que la composition du dernier repas. Ainsi, on s'intéressera uniquement aux valeurs d'abondances relatives des différents acides détectés dans chaque échantillon. Ce bloc de données sera noté par la suite $\mathbf{X}^{(1)}$. À l'intersection d'un ligne i et d'une colonne j, on note $x_{i,j}^{(1)}$ l'abondance relative d'un acide i dans l'échantillon j. Il comporte $n^{(1)} = 940$ lignes, soit 940 échantillons et $p^{(1)} = 77$ colonnes, qui correspondent aux acides biliaires. On a deux lots d'extraction qui comportent respectivement 353 et 587 échantillons traités dans deux années consécutives suite à l'arrivage des échantillons.

2.3 Données de métataxonomique : le microbiote intestinal

Le microbiote intestinal correspond à l'ensemble des micro-organismes présents dans les intestins. Il a de très nombreux rôles au sein de l'appareil digestif : il permet de contrôler la prolifération de micro-organismes pathogènes, intervient dans le contrôle de l'inflammation intestinale, des processus de réabsorption, mais aussi dans de nombreuses réactions métaboliques comme celle qui consiste à transformer les acides biliaires primaires en acides biliaires secondaires [Jandhyala (2015)].

La métataxonomie consiste à étudier une partie spécifique de l'ensemble des génomes présents dans le microbiote. On commence par extraire l'ensemble de l'ADN présent dans les intestins (humain, du microbiote et lié à ce qui a été mangé). Puis on amplifie et on décode les séquences d'une région spécifique des ARNs 16S. À partir de ces séquences amplifiées et décodées, des traitements bioinformatiques sont réalisés pour obtenir une liste d'Amplicon Sequence Variant (ASV), des séquences sélectionnées à partir de modèles qui prennent en compte les erreurs de séquençage. La méthode utilisée ici est la méthode Dada2 basée sur la base de données Silva (version 132) [Callahan et al. (2016)].

Ces ASVs pourront ensuite être annotées à partir d'une base de données de référence. Annoter un ASV correspond à identifier à quelle famille, à quel genre voire à quelle espèce de microbe correspond cet ASV. Pour cela, on confronte l'ASV à la base de données et on cherche à quelle séquence de référence elle est similaire à 97% ou plus. Une séquence correspondant à une souche, on peut ainsi connaître le nombre de fois que chaque souche a été observée dans chaque échantillon. Cependant, pour des raisons expérimentales et techniques, le nombre total de séquences extraites ne peut pas être relié au nombre réel total de molécules présentes dans l'échantillon et le nombre de séquences détectées peut très fortement varier selon l'échantillon. Par exemple, certaines substances inhibent l'extraction et dans certains cas, on n'arrive pas à casser la membrane de certaines bactérie et on ne peut donc pas extraire l'ADN de ces dernières. Ainsi, on ne s'intéresse pas aux comptes absolus mais bien à des fréquences observées dans l'échantillon [Gloor et al. (2017)]. Si on combine les données récoltées pour tous les

échantillons, on obtient une table de fréquences correspondant à 4882 souches détectées. L'extraction d'ADN a été faite sur place en Afrique et à l'institut Pasteur de Paris, la construction des librairies de séquence a été faite par une compagnie Canadienne (Microbiome Insights) et la génération des tableaux de comptage a été réalisée à Paris à l'Institut Pasteur.

On a affaire à des données avec une très grande proportion de valeurs nulles. En effet, certains ASVs n'ont été détectés que dans un faible nombre d'échantillon, voire un seul échantillon. Pour éviter d'éliminer une trop grande quantité d'information en écartant les ASVs faiblement représentés du tableau, nous avons choisi d'agréger les données de comptage de souches à un niveau taxonomique plus élevé : le genre (grâce au package *phyloseq*, [Paul J. McMurdie (2013)]). Agréger la table de comptage revient à considérer le jeu de données comme étant une matrice de comptage de genres au lieu de considérer une matrice de comptages de souches. Ainsi, on perd un niveau de précision pour l'identification des biomarqueurs mais on conserve une plus grande partie de l'information. Une fois que l'on a agrégé les données puis filtré les genres trop peu présents, on obtient un bloc de données $\mathbf{X}^{(2)}$ de $n^{(2)} = 911$ lignes et $p^{(2)} = 437$ colonnes (437 genres). À l'intersection d'une ligne i et d'une colonne j, on note $x_{i,j}^{(2)}$ la fréquence du genre j dans l'échantillon i par rapport au nombre total de genres qui ont été observées dans l'échantillon i.

On note que $n_1 > n_2$, c'est lié au fait que nous disposions seulement d'une partie des échantillons de l'étude, l'extraction a été terminée durant mon stage mais nous n'avons pas eu le temps de les intégrer aux analyses. On utilisera les échantillons communs à $\mathbf{X}^{(1)}$ et $\mathbf{X}^{(2)}$ pour les analyses qui prennent en compte ces deux blocs simultanément.

2.4 Données cliniques supplémentaires

Dans le cadre du projet Afribiota, une grande quantité de données complémentaires sur les enfants a été acquise : sous la forme d'informations anthropométriques comme la taille, l'âge, le poids ou bien encore la circonférence du crâne mais aussi sous la forme de tests biologiques pour connaître la réponse inflammatoire intestinale par exemple. Parmi toutes ces variables, certaines sont connues pour avoir une importance biologique dans le cadre de l'étude de la sous-nutrition. En plus du sexe, de l'âge, du pays d'origine et de la variable qui indique le retard de croissance, on va s'intéresser à des biomarqueurs de l'inflammation intestinale comme la calprotectine fécale (qui reflète aussi l'état de perméabilité intestinale) et l'α_1-antitrypsine. On s'intéresse aussi à l'anémie (taux d'hémoglobine en dessous de 11g/l), aux helminthes (organismes connus pour induire un tamponnement de la réponse immunitaire) et à l'âge de l'arrêt de l'allaitement par la mère [Vonaesch et al. (2018)].

Un questionnaire standardisé a été rempli par les familles pour connaître les antécédents médicaux de l'enfant, la grossesse de la mère, le régime alimentaire, le niveau d'hygiène et les caractéristiques du ménage dans lequel l'enfant vit. Au total, on dispose d'un jeu de données de $n^{(3)} = 926$ lignes et $p^{(3)} = 774$ variables, qui seront généralement utilisées comme variables illustratives. On notera ce bloc $\mathbf{X}^{(3)}$ avec $x_{i,j}^{(3)}$ la modalité ou la valeur que prend l'échantillon i pour la variable j.

2.5 Résumé des données multiblocs du projet Afribiota

La table 1 résume le nombre d'individus et le nombre de variables pour les trois blocs de données que nous avons considérés.

Sur les 926 enfants, 490 viennent de Madagascar et 436 viennent de Centrafrique. Pour les enfants malgaches (respectivement Centrafricains), 254 (respectivement 242) sont atteints de retard de croissance et 236 (respectivement 194) ne sont pas atteints.

Bloc	Type de données	n	p
$\mathbf{X}^{(1)}$	Métabolomique	940	77
$\mathbf{X}^{(2)}$	Métataxonomique	911	437
$\mathbf{X}^{(3)}$	Clinique	926	774

TABLE 1 – Résumé des dimensions des les trois blocs de données considérés dans mon stage.

3 Les données compositionnelles

Parmi les données dont nous disposons pour cette étude, les données d'analyse métabolomique des sels biliaires et de métataxonomique sont des données compositionnelles qui nécessitent une approche spécifique qui est présentée dans cette partie.

3.1 Définition et propriétés

Les données compositionnelles sont un type de données qui prennent la forme de proportions : les observations sont décrites par leur composition en différents éléments ou en différentes parties. Comme précisé précédemment, les données de métabolomique et de métataxonomique sont des données compositionnelles : elles décrivent des échantillons par leur composition relative en acides biliaires ou bien en genres. Deux propriétés essentielles aux données compositionnelles sont qu'elles sont positives mais aussi que pour un échantillon donné, la somme de ses valeurs est constante et unique pour tout le jeu de données (100 dans le cas de compositions exprimées en pourcentages par exemple). Cette dernière caractéristique est appelée contrainte de fermeture. [Pawlowsky-Glahn and Egozcue (2006)].

De plus, pour que les résultats observés soient reproductibles et généralisables, les analyses que l'on va mener se doivent d'être cohérentes d'un point de vue subcompositionnel : les conclusions de nos analyses doivent être stables selon le groupe de variables considérées. En effet, les techniques de séquençage modernes utilisées ne garantissent pas que tous les composants d'un milieu soient détectés [Gloor et al. (2017)]. On peut alors avoir des tables de compositions différentes d'une expérience à une autre, les individus ne seront pas exactement décrits par le même groupe de variables. Or, on veut que nos conclusions sur les relations entre les composants communs aux deux expérience soient identiques et généralisables. Par exemple, les données de métataxonomique que nous avons à disposition représentent en fait uniquement les fréquences des genres du compartiment pour lesquels on a pu extraire et amplifier l'ADN. Le jeu de données est donc un sous-ensemble de la table qui contiendrait réellement tous les genres présents dans l'échantillon. Lors des analyses, on veut que les liens entre les colonnes (les genres observés) soient indépendants du groupe de genres pris en compte [Greenacre (2018)].

La contrainte de fermeture et la nécessité de cohérence subcompositionnelle rendent problématique l'utilisation des outils statistiques classiques. Avec la fermeture, si une des proportions d'un échantillon est augmentée, toutes les autres vont diminuer indépendamment d'un lien éventuel entre les différentes variables de la composition, ce qui rend inutilisable les coefficients de corrélation classiques. Alors, si on s'intéresse à la moyenne d'une variable pour un jeu de données complet et pour une sous-composition de ce jeu, on va avoir des résultats différents. Ainsi, il est déconseillé d'utiliser la moyenne, l'écart type ou bien encore le coefficient de corrélation sur des données compositionnelles.

3.2 Log Transformation

Pour remédier à l'absence de cohérence subcompositionnelle des outils statistiques classiques sur nos données, nous considérerons dans ce rapport une solution simple mais efficace : travailler sur les ratios entre les composants. En effet, ces derniers sont stables quelle que soit

la sous-composition étudiée, considérer des ratios de valeurs compositionnelles permettra de s'affranchir de la contrainte de fermeture.

Comme les ratios de compositions peuvent prendre des valeurs très petites ou très grandes et présenter des distributions dissymétriques, nous appliquons aux ratios une transformation logarithmique. Il existe plusieurs méthodes pour réaliser une log-transformation sur des données compositionnelles, celle que nous avons choisie est la transformation des données en Log Ratio Centrés pondérés, que l'on nomme CLR [Aitchison (1986)]. La log-transformation de type CLR a pour avantage de nécessiter peu de temps de calcul et il est plus aisé d'interpréter les résultats et les relations entre des CLR qu'entre d'autres types de log-ratios (additifs ou isométriques) [Greenacre (2018)]. Pour une valeur d'un tableau, faire cette transformation revient à calculer le logarithme du ratio entre la valeur observée et la moyenne géométrique pondérée de l'ensemble du tableau.

Pour une valeur $x_{i,j}$ d'un jeu de données à i observations et j colonnes, on peut calculer $x^*_{i,j}$ sa valeur log-transformée :

$$(\text{CLR}) : x^*_{i,j} = \log\left(x_{i,j} / \prod_j x_{i,j}^{c_j}\right) = \log\left(x_{i,j}\right) - \sum_j c_j \log\left(x_{i,j}\right)$$

avec c_j le poids associé à la colonne j dans le calcul de la moyenne géométrique.

On utilise une pondération pour donner moins d'importance aux parts des observations qui sont présentes en très faibles proportions dans les échantillons. Les c_j sont respectivement les moyennes des j colonnes dans les données initiales. Lors de la mise en place des algorithmes traitant des données compositionnelles sous *R*, les principaux packages qu'on a utilisés sont *FactoMineR* [Lê et al. (2008)] et *easyCODA* [Greenacre (2018)].

3.3 Analyse des Log Ratios (LRA)

Les jeux de données que l'on utilise pour le projet sont des jeux de données avec un nombre de variables relativement conséquent : 77 acides biliaires et 437 genres. Un des objectifs a donc été dans un premier temps de visualiser l'information contenue dans les données par une méthode de réduction de dimension : l'Analyse des Log Ratios (LRA). La LRA peut être vue comme une adaptation de l'Analyse en Composantes Principales (ACP) au cas des données compositionnelles. En effet, il s'agit de réaliser une ACP pondérée sur le tableau issu de la log-transformation centrée [Greenacre (2018)].

Lors d'une ACP pondérée, on prend en compte des poids pour pondérer l'importance d'une colonne ou d'une ligne dans le calcul. Dans le cas de nos données compositionnelles, on veut éviter que des acides biliaires ou des genres qui sont présents en très faibles proportions sur l'ensemble des individus aient une trop grande importance dans l'analyse. On choisit donc de pondérer la colonne j par la moyenne de ses valeurs, ainsi on évite de donner trop d'influence à des parts trop peu abondantes. On associe ensuite un poids $r_i = 1/n$ à tous les individus. C'est un système de pondération proche de celui qui est utilisé en analyse des correspondances, la seule différence est que, pour la LRA, la matrice d'intérêt est celle des logarithmes des valeurs de compositions [Greenacre (2016)] alors que si l'on appliquait une analyse des correspondances, la matrice d'intérêt serait simplement la matrice des compositions. On peut ensuite interpréter les résultats de la LRA de la même manière que lors d'une ACP classique mais en se rappelant bien que l'on a ici des log-ratios d'abondances relatives, on observe des variations dans les abondances ou fréquences des différentes parts.

Pour résumer, la LRA revient à effectuer une décomposition en valeurs singulières sur la matrice

S définie comme suit [Aitchison (1986)] :

$$\mathbf{S} = \mathbf{D_r}^{1/2} \left(\mathbf{I} - \mathbf{1r}^\top \right) \log(\mathbf{N}) \left(\mathbf{I} - \mathbf{1c}^\top \right)^\top \mathbf{D_c}^{1/2},$$

où **N** est la matrice de données compositionnelles ne contenant aucune valeur nulle. Les dimensions de la LRA sont obtenues à partir des vecteurs singuliers à gauche de **S**.

3.4 Gestion des valeurs nulles

Les données de métabolomique et de métataxonomique comportent respectivement 40,5 % et 88.7 % de valeurs nulles qui marquent une absence de différents acides ou genres dans certains échantillons. C'est une caractéristique très courante, notamment dans des données de métataxonomique. Or, cela rend impossible la transformation CLR ($\log(0)$ n'existe pas). Il faut donc trouver un moyen pour contourner ce problème sans pour autant perdre ou altérer l'information contenue dans les données. On ne peut pas écarter les variables ou les individus qui présentent des valeurs nulles car cela reviendrait à écarter la quasi totalité des données et cela biaiserait les analyses. Une solution alternative est proposée dans cette partie.

3.4.1 Remplacement des valeurs nulles

Plutôt que d'éliminer les lignes ou colonnes présentant des valeurs nulles, on va chercher à remplacer ces dernières avant de poursuivre les analyses. Plusieurs méthodes de remplacement sont proposées dans la littérature [Martín-Fernández et al. (2003); Tsagris and Stewart (2018)] mais nous avons choisi la méthode de remplacement simple proposée par Greenacre [Greenacre (2018)], car elle est applicable sur des jeux de données ayant un grand nombre de zéros et ne nécessite pas de connaître les limites de détections pour chacun des acides ou genres étudiés.

Le remplacement simple proposé par Greenacre se fait en deux étapes. On commence par remplacer les valeurs nulles par colonnes : pour une colonne donnée, les valeurs nulles sont remplacées par la moitié de la plus petite valeur positive présente sur la colonne. Ensuite, on applique une opération de fermeture par ligne (on recalcule la nouvelle composition) sur les données ainsi complétées pour revenir à des données de compositions.

On applique la transformation CLR sur les données ainsi modifiées. Je tiens à préciser ici que ce problème de remplacement et de gestion des valeurs nulles est une problématique d'actualité et aucune solution idéale n'a été pour le moment identifiée. Ce rapport présente la méthode qui nous a semblé la plus viable après avoir comparé les résultats avec différentes méthodes.

3.4.2 Analyse de la sensibilité

Remplacer les valeurs nulles par d'autres valeurs comme on le fait avec le remplacement simple de Greenacre pose un problème car on introduit de la variabilité qui n'était pas présente au départ dans le jeu de données initial. On risque de modifier les relations entre les parts des compositions à l'échelle du jeu de données. Une méthode pour essayer d'évaluer l'impact de ce remplacement sur la structure des données est l'analyse de sensibilité [Greenacre (2018)].

Pour ce faire, on va faire varier la valeur utilisée lors du remplacement des zéros et observer si la structure des données n'est pas trop modifiée. Plutôt que de remplacer les zéros par la valeur minimale de chaque colonne divisée par 2, on divise la valeur minimale par une valeur k que l'on va faire varier. On peut ainsi générer autant de jeux de données que l'on génère de valeurs de k. Puis, grâce à la LRA, on va pouvoir observer et mesurer les différences de structures qui sont engendrées par cette modification. On se sert du jeu de données pour lequel $k = 2$ comme jeu de référence pour extraire des composantes principales. Puis, on projette dans cet espace les

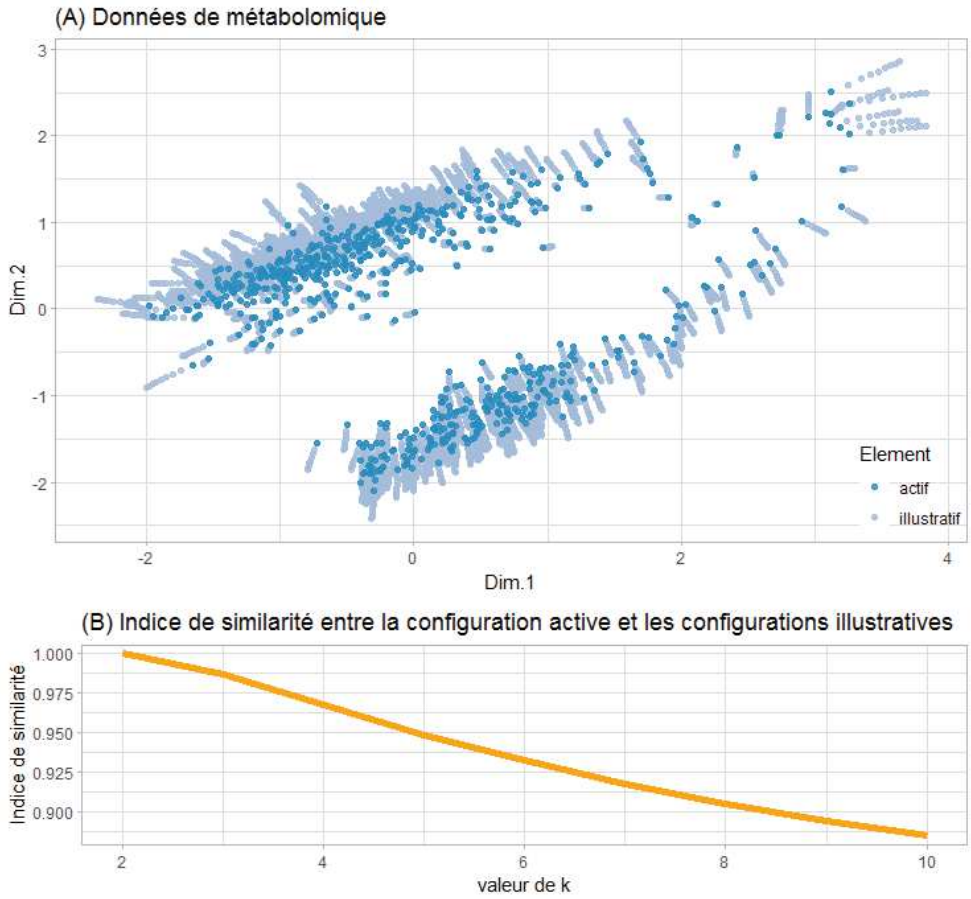

FIGURE 2 – Analyse de la sensibilité pour le jeu de données de métabolomique pour k allant de 2 à 10. (A) Les représentations des échantillons issus d'une LRA en faisant varier la valeur de remplacement. (B) Évolution de l'indice de similarité lorsque k varie.

individus issus des autres jeux de données générés avec d'autres valeurs de k, ce sont des individus illustratifs. Avec cette représentation, on peut observer si la modification de la valeur de remplacement influe fortement ou pas sur l'information contenue dans les données. Pour quantifier les différences que l'on observe entre la représentation de référence et les représentations obtenues avec différentes valeurs de k, on peut calculer un coefficient de similarité entre les matrices de coordonnées. On utilise l'indice de similarité de Procrustes [Indahl et al. (2016)], qui calcule la similarité entre deux matrices. Plus la valeur du coefficient est proche de 1, plus les jeux de données ont des structures de corrélations proches. Si notre méthode est viable, on s'attend à ce que la similarité entre les différentes représentations soit élevée.

C'est ce qu'on observe sur la Figure 2 qui représente l'étude de sensibilité réalisée sur les données de métabolomique. Les représentations sont proches, il semble qu'il y ait seulement un effet de dilatation des représentations, on conserve donc bien la nature des relations entre nos individus. L'indice de similarité entre les configurations reste élevé et proche de 1 même lorsque k varie. Les résultats que l'on a obtenus pour les données de métataxonomique nous ont mené aux mêmes conclusions, la structure est bien conservée lorsque k varie.

3.5 Analyse différentielle

Pour pouvoir identifier les biomarqueurs du retard de croissance sur notre cohorte, on veut caractériser et identifier les différences entre les enfants atteints ou non-atteints de retard de croissance, cette étape est appelée analyse différentielle.

Lors du déroulement de mon stage, nous avons voulu comparer les résultats que l'on obtient lorsque l'on analyse les données compositionnelles brutes non transformées et les données

transformées par la méthode CLR présentée précédemment. Toutes les analyses présentées par la suite ont été réalisées à la fois sur des données compositionnelles brutes et les données transformées, pour comparer les résultats obtenus par une approche qui ne prend pas en compte les spécificités des données compositionnelles à celle avancée lors du stage.

3.5.1 Test de Wilcoxon-Mann-Whitney

Le test de Wilcoxon-Mann-Whitney [Mann and Whitney (1947); Wilcoxon (1945)] est un test non paramétrique qui a pour but de comparer deux populations. L'hypothèse nulle de ce test ($\mathscr{H}_0$) est que les distributions associées à ces deux populations sont les mêmes, l'hypothèse alternative ($\mathscr{H}_1$) étant que les deux populations sont stochastiquement différentes. Dans notre cas d'application, on a un échantillon de n_1 valeurs pour les enfants atteints et un échantillon de n_2 valeurs pour les enfants sains. Sous $\mathscr{H}_0$, nous supposons qu'un acide biliaire ou un genre a la même abondance relative chez les enfants des deux populations.

Le calcul de la statistique de test qui va déterminer si l'on rejette ou non $\mathscr{H}_0$ se fait en utilisant les rangs des valeurs prises pour l'ensemble des observations. Pour une variable données, on trie par ordre décroissant toutes les valeurs observées puis on associe à chaque valeur un rang. En cas d'égalité, on associe le même rang moyen aux observations égales. On calcule ensuite les quantités U_1 et U_2 : U_1 correspond au nombre de fois qu'une valeur du groupe 1 est inférieure à une valeur du groupe 2, les égalités comptant pour 0.5 point. On calcule U_2 par analogie. La statistique de test U^* est le minimum entre U_1 et U_2.

Ici, comme les groupes comportent un grand nombre d'échantillons (n_1 et n_2 sont supérieurs à 40), on peut approximer la loi de U^* par une loi normale de moyenne $\mu = \frac{n_1 n_2}{2}$ et de variance $\sigma^2 = \frac{n_1 n_2}{12}\left((n+1) - \sum_{i=1}^{k} \frac{t_i^3 - t_i}{n(n-1)}\right)$. Cette formule permet de gérer les égalités, t_i étant le nombre de valeurs ayant le même rang i, et k étant le nombre de rangs distincts.

Enfin, on calcule la probabilité critique associée à la valeur de la statistique selon la loi normale précédemment définie.

3.5.2 Régression logistique

Pour identifier des biomarqueurs, on s'est aussi basé sur l'utilisation de modèles de régression logistique, on cherche à savoir si les acides ou genres ont un effet sur le retard de croissance. La variable réponse de notre modèle est la variable binaire *stunted*, qui marque le retard de croissance. Pour chaque acide et chaque genre, on calcule un modèle de régression à plusieurs facteurs sans interactions. Comme on l'a vu, il est important de prendre en compte l'effet lot dans le modèle dédié aux acides biliaires.

Le modèle peut se résumer ainsi dans le cas de la métabolomique avec $x_{\cdot,j}^{(1)}$ les abondances relatives pour un acide biliaire donné :

$$\begin{aligned} \text{stunted} \sim\ & x_{\cdot,j}^{(1)} + \text{age} + \text{sexe} + \text{pays} + \text{anemie} + \text{calprotectine} + \text{alphaantitrypsine} \\ & + \text{age\_allaitement} + \text{lot} \end{aligned} \tag{1}$$

Dans le cas de la métataxonomique, avec $x_{\cdot,j}^{(2)}$ les fréquences pour un genre donné :

$$\begin{aligned} \text{stunted} \sim\ & x_{\cdot,j}^{(2)} + \text{age} + \text{sexe} + \text{pays} + \text{anemie} + \text{calprotectine} + \text{alphaantitrypsine} \\ & + \text{age\_allaitement} \end{aligned} \tag{2}$$

Pour chaque modèle, on récupère la probabilité critique associée au test de significativité de l'effet de $x_{\cdot,j}^{(\cdot)}$.

3.5.3 Correction pour les tests multiples

Quelle que soit la méthode utilisée pour calculer les probabilités critiques, et comme de nombreux tests sont effectués, toutes les probabilités critiques calculées sont corrigées de sorte à contrôler le *taux de faux positifs*, False Discovery Rate (FDR) en anglais [Benjamini and Hochberg (1995)]. Les acides ou les genres pour lesquelles la probabilité critique ajustée est inférieure à 0,05 sont ceux qui nous intéressent et que l'on identifie comme biomarqueurs.

3.6 Intégration dans des méthodes multi-blocs

Un des autres objectifs de mon stage a été d'intégrer simultanément les données compositionnelles du projet Afribiota. Nous avons choisi deux méthodes multiblocs pour détecter des liens possibles entre les abondances biliaires et la fréquence des genres : (1) l'Analyse Factorielle Multiple (AFM) et (2) l'Analyse Canonique des Corrélations Régularisée Généralisée (RGCCA).

3.6.1 Analyse Factorielle Multiple (AFM)

L'AFM est une méthode de réduction de dimensions adaptée à l'analyse simultanée de plusieurs tableaux de données. L'AFM va fournir une représentation consensus des distances entre les individus dans un espace de dimensions réduites en prenant en compte l'ensemble des variables des deux tableaux. L'idée principale derrière cette méthode est de concaténer les blocs de données et de réaliser une ACP sur le grand tableau ainsi formé [Abdi et al. (2013)]. On pourra ainsi repérer les structures communes qui existent entre les différents points de vue sur nos échantillons.

Il est nécessaire d'équilibrer l'influence de chaque bloc dans l'analyse, on veut éviter que des tableaux de grande dimension aient une trop grande importance dans l'analyse et on veut faire ressortir les structures communes entre les blocs de données. Pour parvenir à cet équilibre, on pondère les valeurs de chaque tableau par la racine carrée de la première valeur propre issue de la LRA de ce tableau seul. Cela permet d'équilibrer l'influence des tableaux de tailles différentes mais aussi de faire ressortir les tableaux pour qui présentent des structures bien discriminantes dès la premières dimensions des ACP séparées.

Grâce à la transformation CLR, on peut directement intégrer les données compositionnelles transformées dans l'AFM. Lors de l'ACP globale, on utilisera les tableaux transformés et pondérés. On peut intégrer les poids que l'on attribuait aux colonnes et aux lignes lors d'une LRA. Ici, les poids que l'on attribue aux lignes sont identiques d'un tableau à l'autre, on conserve donc $r_i = 1/I$ pour toutes les lignes. On attribue ensuite à chaque colonne le poids qu'on lui attribuerait lors d'une LRA séparée.

L'AFM fournit une représentation consensus des structures des tableaux, on espère observer des groupes d'individus sur les dimensions extraites. De la même manière qu'en ACP on pourra décrire ces groupes en fonction de leur position dans l'espace et des variables qui contribuent le plus à la construction des axes [Lê et al. (2008)]. Dans un cas idéal, on pourrait définir des groupes d'échantillons à la fois par des abondances relatives d'acides biliaires et des genres spécifiques.

3.6.2 Analyses Canonique des Corrélations Régularisée Généralisée (RGCCA)

Comme l'AFM, la méthode RGCCA (Regularized Generalized Canonical Correlation Analysis) est une autre méthode d'analyse multiblocs adaptée pour intégrer différents tableaux de données. La méthode se base sur des principes proches de ceux de l'approche PLS, basée sur l'utilisation de variables latentes utilisées pour expliquer et permet de prendre en compte une

matrice de design qui définit les relations entre nos blocs de données [Tenenhaus and Tenenhaus (2011)].

Avant de faire les analyses, on doit indiquer si il existe un lien entre chaque paire de blocs de données. Dans notre cas, on a considéré nos blocs de données comme étant tous liés, car on s'attend à voir des interactions entre les acide biliaires et les genres observées. Puis, on va chercher à extraire une ou plusieurs composantes pour chacun des blocs avec deux objectifs :

— les composantes extraites doivent bien expliquer leur propre bloc,
— les composantes des blocs connectés doivent être les plus corrélées possibles

Le problème d'optimisation dans le cas général est le suivant :

$$\underset{\mathbf{a}_1,\ldots,\mathbf{a}_K}{\text{Maximiser}} \sum_{j>k} c_{j,k} g\left(\text{cov}\left(\mathbf{X}_j\mathbf{a}_j, \mathbf{X}_k\mathbf{a}_k\right)\right) \tag{3}$$

$$\text{tel que } \tau_j \left\|\mathbf{a}_j\right\|^2 + \left(1-\tau_j\right) \text{Var}\left(\mathbf{X}_j\mathbf{a}_j\right) = 1, \forall j = 1,\ldots,K,$$

avec :

— le vecteur $\mathbf{a}_j \in \mathbb{R}^{p_j}$ sont les coefficients appliqués au bloc numéro j ;
— les paramètres τ_j sont tous fixés égaux à 1 dans notre cas, on veut faire en sorte que la composante explique bien le bloc dont elle est issue mais aussi que les composantes soient bien corrélées entre elles.
— la fonction g correspond au « schéma » utilisé, dans notre cas $g(x) = |x|$;
— les $c_{j,k}$ indiquent les liens entre blocs j et k, tous nos blocs sont liés donc les $c_{j,k}$ sont tous égaux à 1.

Une fois qu'on a extrait une ou plusieurs composantes pour chaque bloc, on peut représenter les individus selon les différentes composantes de chacun des blocs. Pour mesurer la qualité de la réduction de dimension effectuée par RGCCA, on utilise la variance moyenne expliquée –Average Variance Explained (AVE)– qui évalue la partie de la variabilité expliquée par les composantes. On peut calculer un AVE pour chacun des blocs séparément mais aussi pour le modèle complet qui prend en compte tous les blocs.

4 Résultats : Identification de biomarqueurs du retard de croissance

Avant de présenter les résultats, il est important de préciser que l'on va quasiment exclusivement se focaliser sur les échantillons issus des selles des enfants. En effet, les prélèvements duodénaux et gastriques sont des prélèvements qui, pour des raisons éthiques, n'ont pas pu être réalisés chez les jeunes enfants sans retard de croissance. Lorsque l'on compare des populations d'enfants sains ou non, on compare uniquement les échantillons issus des selles.

4.1 Analyse exploratoire

Pour commencer l'identification des biomarqueurs du retard de croissance, nous avons utilisé quelques outils de visualisation pour observer et étudier certaines de nos hypothèses.

4.1.1 Premières observations

Pour la métabolomique, on s'intéresse tout d'abord aux abondances relatives des acides biliaires primaires et secondaires en fonction de l'état de santé des enfants pour les échantillons issus des selles. On retrouve bien une des hypothèses de départ sur la Figure 3 , les acides biliaires primaires sont bien relativement plus présents chez les enfants atteints de retard de croissance et inversement pour les acides biliaires secondaires. Cela semble bien indiquer que

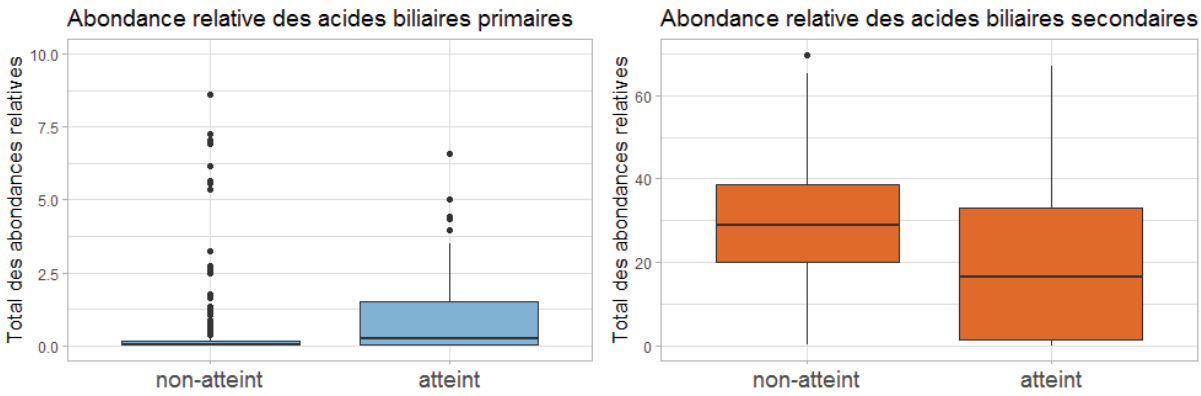

FIGURE 3 – Abondances relatives des acides primaires et secondaires selon l'état de santé des enfants.

les réactions de transformations des acides biliaires primaires en acides secondaires sont ralenties et moins efficaces chez les enfants atteints de retard de croissance.

Ensuite, on a voulu comparer les individus en fonction de différents facteurs d'intérêts : leur sexe, leur pays d'origine et leur classe d'âge. Les différences visibles sont assez peu marquées mais l'âge et le pays d'origine semblent bien déterminer des variations de la composition intestinale en acides biliaires dans nos données, c'est ce à quoi on s'attendait au départ.

Puis, pour vérifier la qualité des données, on a aussi comparé les échantillons selon le lot dans lequel il était au moment de la récolte des données. Il semble qu'il existe une différence entre les individus des deux lots, il faudra par la suite être prudent et tenter de prendre en compte cette différence de composition due au protocole expérimentale et à la récolte des données.

4.1.2 Heatmaps et corrélations croisées

Pour continuer, on a réalisé des heatmaps des jeux de données et des matrices des corrélations croisées pour essayer d'identifier des groupes d'individus et groupes d'acides ou de genres singuliers en se basant sur des méthodes permettant de visualiser l'ensemble de l'information sans utiliser de méthode de projection dans un espace réduit [Leo Lahti (2019)].

Pour des raisons de lisibilité et d'intérêt des résultats, aucune heatmap n'est intégrée au rapport. En effet, que ce soit d'un point de vue de la **métataxonomique** ou des compositions en **acides biliaires**, les heatmaps que l'on obtient pour différents sous groupes d'échantillons ne nous permettent pas de déterminer des groupes d'individus ayant des compositions spécifiques. On retrouve seulement l'effet lot pressenti sur les heatmaps issues du jeu de données de métabolomique lors de la section 4.1.1.

Les **matrices de corrélations croisées** sont des matrices pour lesquelles on calcule la corrélations de Spearman entre toutes les paires d'acides biliaires et de genres possibles [Leo Lahti (2019)]. On représente ensuite la matrice des corrélation obtenue en ne conservant que les acides biliaires et genres pour lesquels au moins une interaction est significativement non nulle au seuil de 5%. Si la corrélation observée est significativement positive ou négative, on pourra penser que l'abondance relative d'un acide biliaire donné est liée à la fréquence d'un genre au sein du compartiment et qu'il y a sûrement une interaction spécifique entre cet acide et ce genre.

Ces heatmaps et matrices de corrélations croisées sont calculées pour 5 sous groupes d'individus différents : tous les échantillons, seulement les échantillons venant de Madagascar, seulement ceux de République Centrafricaine, seulement les échantillons atteints de retard de croissance et seulement le échantillons sains.

La Figure 4 est un exemple de ce que l'on obtient lorsque l'on calcule une matrice de corrélations croisées pour les individus atteints de retard de croissance uniquement : 5 acides biliaires

et 67 genres ont été retenus. Cette matrice ne nous permet pas de mettre en évidence des associations entre des groupes d'acides et des groupes de microbes. Les acides aux structures proches ont le même profil de corrélations et on ne distingue pas de genres au profil spécifique. De plus les corrélations qui ont été retenues sont de faible intensité : les valeurs extrêmes sont -0,27 pour les valeurs négatives et 0,29 pour les corrélations positives.

FIGURE 4 – Matrices des corrélations croisées entre les acides biliaires (en ordonnées) et les genres (en abscisse) lorsque l'on s'intéresse aux individus atteints d'un retard de croissance uniquement

C'est un constat qui est généralisable à l'ensemble des matrices d'associations croisées réalisées, on n'a pas pu mettre en évidence de lien direct entre des groupes d'acides et de genres, quel que soit le sous groupe d'échantillons considéré.

4.1.3 Analyse des Log Ratios

Pour compléter cette section, on peut voir en Figure 5 les résultats de la mise en place d'une LRA. On remarque donc que dans les deux cas (A) et (B) de la Figure 5, la LRA retranscrit une bonne partie de la variabilité totale et que le premier axe discrimine bien les différents types d'échantillons, les échantillons de selles se détachent des échantillons gastriques et duodénaux. On s'attendait à cette distinction, car la composition en acides biliaires évolue tout au long du système digestif. Au contraire, on a du mal à bien distinguer les échantillons duodénaux et gastriques. Cette différence n'est pas plus marquée si on s'intéresse aux dimensions suivantes. Il apparaît aussi que la deuxième dimension pour les acides biliaires sépare les échantillons issus des selles en trois groupes distincts.

On a représenté les résultats d'une LRA réalisée avec les échantillons du type *selles* uniquement en Figure 6 (C) et (D).

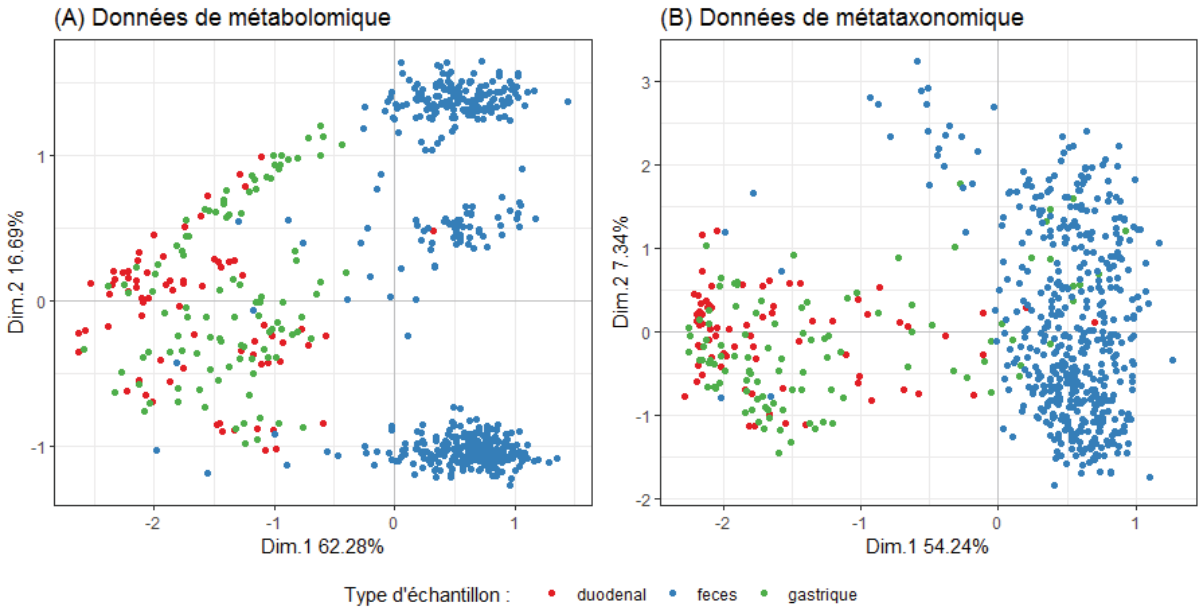

FIGURE 5 – Analyse des Log Ratios sur les deux jeux de données pour tous les échantillons en fonction du compartiment.

Les représentations obtenues expliquent toujours de bonnes proportions de la variabilité totale mais cette fois on n'observe pas de distinction claire entre les individus atteints ou non de retard de croissance. Les deux groupes que l'on distingue sur la Figure 6 (C) sont en réalité deux groupes liés à l'effet lot que nous avions déjà vu précédemment [Greenacre (2018)] sur les deux jeux de données pour l'ensemble des échantillons.

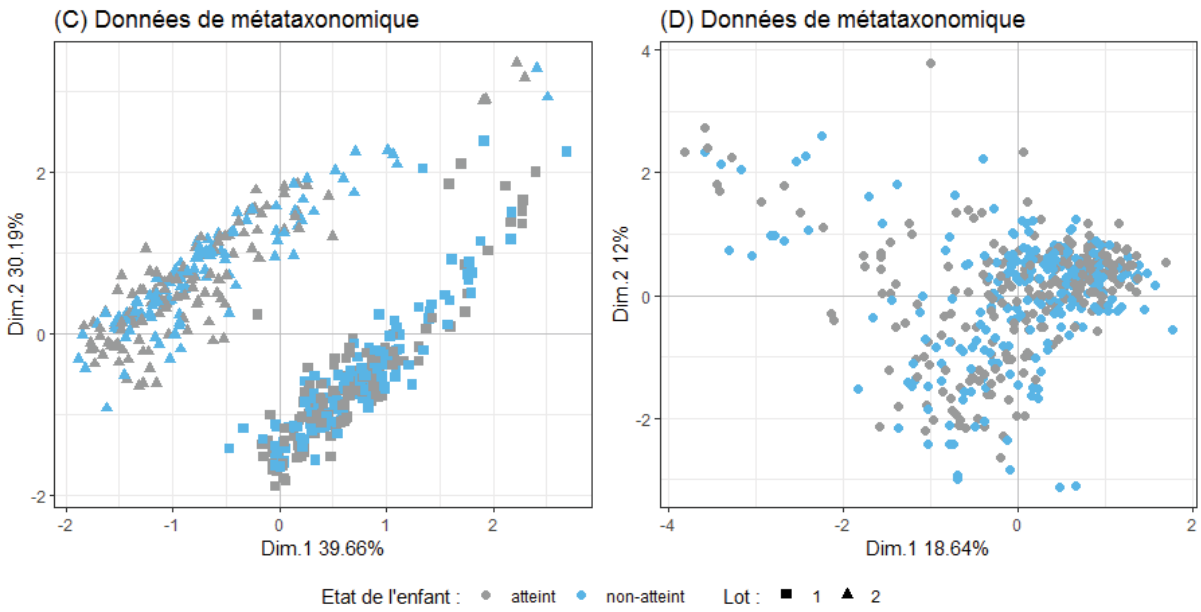

FIGURE 6 – Analyse des Log Ratios sur les deux jeux de données pour les échantillons provenant des selles uniquement.

Au vu de ces différentes observations, il apparaît que les échantillons que nous avons à notre disposition sont assez homogènes, les compositions en genres et en acides biliaires des enfants atteints ou non ne semblent pas être différentes. Pour la suite de notre étude et l'analyse différentielle, on va se focaliser sur des variations plus locales, en se concentrant sur les différences que l'on peut observer sur des acides ou genres particuliers car si des différences sont présentes, elles ne seront visibles qu'à un niveau plus fin.

4.2 Analyse différentielle

Dans cette partie dédiée à la détection de biomarqueurs du retard de croissance, nous avons aussi réalisé les tests à différents niveaux, en se concentrant sur différents sous groupes d'individus (par pays et par âge). À chaque étape on compare les profils d'acides ou de genres retenus comme étant des biomarqueurs selon les groupes d'individus étudiés. Comme indiqué dans la section 3.5, on va aussi comparer les biomarqueurs identifiés lorsque l'on a une approche qui ne prend pas en compte l'aspect compositionnel (Données brutes) des données à notre approche basée sur la transformation CLR (Données CLR).

4.2.1 Tests de Wilcoxon

Dans cette partie, on réalise des tests de Wilcoxon comme décrits dans la section 3.5.1. Pour les données de métabolomique, on réalise un test par acide dans lequel on compare les échantillons atteints d'un retard de croissance aux échantillons sains. Comme on réalise des tests multiples, on corrige les probabilités critiques obtenues par une correction de type FDR. On répète ce processus pour les données de métataxonomique et l'identification de genres. On obtient des listes d'acides et de genres détectés par groupe d'échantillons étudiés.

	Acides biliaires		Genres	
Groupe d'échantillons	Données brutes	Données CLR	Données brutes	Données CLR
Tous	27	25	23	29
Madagascar	23	21	10	21
Centrafrique	7	3	38	72
2-3 ans	17	17	18	21
3 ans et plus	15	11	18	21
Mada 2-3 ans	15	13	5	15
Mada 3ans et plus	4	5	13	15
RCA 2-3 ans	3	5	30	65
RCA 3 ans et plus	3	1	16	33

TABLE 2 – Table de comptage du nombre d'acides biliaires et de genres retenus lors des tests de Wilcoxon sur différents sous groupes d'échantillons avec des données log-transformées (Données CLR) ou non (Données brutes).

Le premier constat est qu'un nombre conséquent d'acides biliaires et de genres ont été détectés comme biomarqueurs potentiels, que ce soit sur les données transformées ou les données brutes. Les listes détaillées seront analysées au cas par cas pour déterminer l'intérêt et le sens biologique des observations mais on peut déjà faire ressortir plusieurs constats généraux.

En ce qui concerne **les acides biliaires**, on voit que les individus atteints ou non ont des profils d'abondances qui diffèrent au niveau global. Si on se penche sur les résultats obtenus par pays, on voit que 21 acides ressortent pour les échantillons malgaches contre 3 seulement pour les échantillons Centrafricains. Selon l'âge, les profils d'acides détectés changent peu, la majorité des acides sont communs aux deux classes d'âge. Plus spécifiquement, les acides semblent bien caractériser les différences entre les enfants malgaches entre 2 et 3 ans alors qu'on ne peut pas ou très peu distinguer les enfants Centrafricains selon leur âge. Parmi les acides biliaires qui ont été détectés, on retrouve des acides biliaires primaires (acide chénodéoxycholique, acide cholique et leur dérivés) mais aussi des acides biliaires secondaires (acide lithocholique, acide déoxycholique et leur dérivés).

Ensuite, pour **les genres**, on peut avoir des conclusions similaires au niveau global et des classes d'âge mais des conclusions différentes au niveau des pays d'origine. Cette fois, ce sont les échantillons Centrafricains qui sont caractérisés par 72 genres contre 21 pour les échantillons

de Madagascar. Parmi les genres détectés, on retrouve des genres de la classe des *Clostridium* comme des microbes des *Ruminococcaceae* ou *Anaerotruncus*. Si on voit que leur abondance est diminuée chez les enfants atteints, ce sont des microbes qui sont des marqueurs classiques d'une infection et d'un dérèglement du bon fonctionnement de la digestion.

D'une manière générale, la sélection d'acides et de genres discriminants est assez stable selon que l'on s'intéresse aux résultats des tests sur les données brutes ou les données transformées à part lorsque l'on se penche sur les genres qui caractérisent les individus Centrafricains. On retrouve certains acides communs et des acides spécifiques à une méthode d'analyse. Les acides les plus discriminants comme l'acide lithocholique sont bien détectés par les deux analyses mais des acides pour lesquels les variations sont plus subtiles peuvent être détectées dans un cas mais pas dans l'autre. Pour compléter et essayer d'expliquer cette observation, on pourra s'intéresser plus spécifiquement à ces variables en particulier. On pourrait par exemple détecter des outliers spécifiques à un acide ou un genre susceptibles de fausser les tests. En effet, le test de Wilcoxon est sensible au fait que notre jeu de données comporte beaucoup de valeurs égales. Ce phénomène est plus présent pour les analyses sur les genres, la variabilité des genres sélectionnés dans les différents sous groupes ou pour les deux cas d'étude s'explique sûrement par la grande proportion de valeurs nulles dans les données, certains genres ne sont présents que chez un nombre restreint d'individus.

Malgré les limites du test de Wilcoxon, nos conclusions vont dans le sens de l'hypothèse de départ : les enfants malnutris en retard de croissance présentent un métabolome et un microbiote altérés, certains microbes vont être trop peu présents dans le milieu, les transformations des acides biliaires ne vont pas pouvoir se réaliser et la digestion des aliments et la croissance de l'enfant vont être ralenties.

4.2.2 Régression logistique

Dans cette partie, on réalise des régressions logistiques comme présenté dans la section 3.5.2 en suivant le même processus de tests que pour les tests de Wilcoxon. On obtient des listes d'acides et de genres ayant un effet sur l'état de santé de l'enfant dans le modèle après correction des probabilités critiques.

	Acides biliaires		Genres	
	Données brutes	Données CLR	Données brutes	Données CLR
Tous	0	0	9	18
Madagascar	0	0	2	3
Centrafrique	0	0	23	34
2-3 ans	0	0	0	12
3 ans et plus	0	0	2	18
Mada 2-3 ans	0	0	0	0
Mada 3ans et plus	0	0	0	15
RCA 2-3 ans	0	1	0	26
RCA 3 ans et plus	0	1	16	0

TABLE 3 – Table de comptage du nombre d'acides biliaires et de genres qui ont un effet sur l'état de croissance des enfants en fonction du groupe d'échantillon étudié avec des données log-transformées (Données CLR) ou non (Données brutes).

La première chose que l'on voit lorsque l'on observe ces résultats est que pour les acides biliaires, très peu voire aucun d'entre eux ne semble avoir un effet sur la variable qui indique l'état de santé de l'enfant. Les seuls qui semblent avoir un effet ont été détectés pour les échantillons centrafricains d'une classe d'âge donnée : l'acide DCA3Gluc pour les enfants de 2-3

ans et l'acide Taurocholique pour les enfants de 3 ans et plus. Pour les données de métataxonomique, les tests font ressortir des genres ayant un effet au niveau global, au niveau des différentes classes d'âge et au niveau des pays. À nouveau, et comme observé avec les tests de Wilcoxon, les genres semblent avoir un effet sur la variable stunted surtout pour les individus Centrafricains. Dans les genres sélectionnés, on retrouve à nouveau des genres de la famille des *Clostridia* ou des familles très proche des *Clostridia* (*Ruminococcaceae_UCG*,*Anaerotruncus*, *Veillonella*).

Ici, les deux méthodes (Données CLR ou Données brutes) apportent des conclusions plus éloignées l'une de l'autre que ce que l'on avait pu voir en 4.2.1. En effet, prendre en compte l'aspect compositionnel permet de dégager plus de genres discriminants que ce que l'on avait obtenu sur les données brutes. Des vérifications plus fines sont nécessaires pour compléter ces informations et vérifier le rôle des genres et acides détectés mais on peut déjà confirmer les premières conclusions de la section 4.2.1, la malnutrition perturbe la composition du métabolome et du microbiome, et cette perturbation varie selon le pays dans lequel les enfants vivent.

4.3 Analyse Factorielle Multiple (AFM)

On réalise une AFM comme décrit dans la section 3.6.1 en utilisant les deux jeux de données que nous avons à notre disposition, le bloc de métabolomique et le bloc de métataxonomique. Sur la figure 7, on peut voir la représentation consensus fournie par l'AFM.

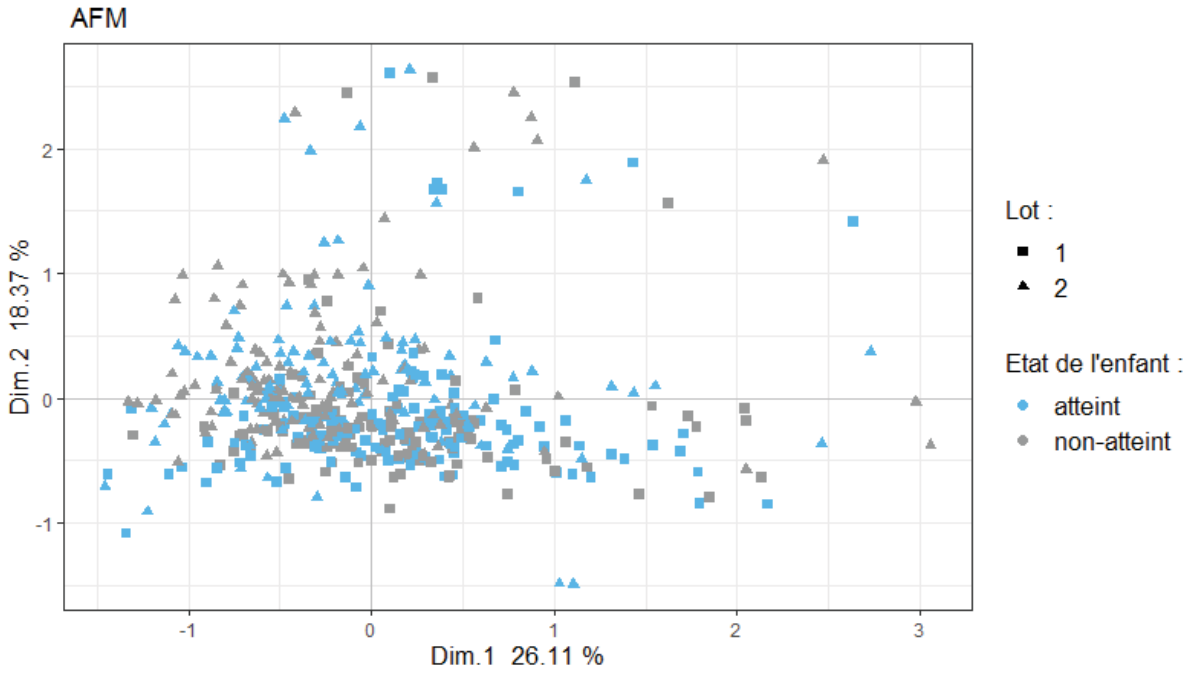

FIGURE 7 – Analyse factorielle multiple pour l'ensemble des jeux de données pour les échantillons provenant des selles.

On ne peut pas dire que les individus sains ou non se détachent les uns des autres, et on n'observe pas de groupe d'individus caractéristique sur les deux premières dimensions. Il semble tout de même que le premier axe sépare les individus des deux lots de l'expérience. Si on s'intéresse aux variables qui contribuent le plus à la construction des deux premiers axe, on s'aperçoit que le premier axe est plus déterminé par la composition en acides biliaires (acides *Cholique*, *Chénodéoxycholique*, *Lithocholique*) là où le second est déterminé par la composition en genres (*Prevotella_9* et des *Clostridias*).

L'AFM ici ne nous a pas permis d'affiner ou de préciser des relations probables entre des acides biliaires et des genres biomarqueurs, le grand nombre de variables prises en compte et l'absence

de séparation entre nos individus empêchent d'apporter des interprétations supplémentaires. On pouvait s'y attendre après avoir vu ce qu'on a obtenu sur les LRA séparées. Cependant, on a pu intégrer plusieurs blocs de données compositionnelles dans un processus d'analyse multibloc qui permet l'obtention et l'interprétation d'une représentation consensus.

4.4 Analyses Canonique des Corrélations Régularisée Généralisée (RGCCA)

Pour compléter l'intégration par des méthodes multiblocs, on a réalisé une RGCCA sur 3 jeux de données : les deux utilisés en 4.3 et un jeu de données cliniques. Ce dernier est composé des variables d'intérêt : *stunted*, *age*, *sexe*, *pays*, *calprotectine*, *alphaantitrypsine* et *lot*. Comme ce sont des données qualitatives, on intégrera ce bloc sous la forme d'un tableau disjonctif complet. La représentation Figure 8 nous montre les représentations des individus selon les différentes composantes extraites pour chacun des blocs.

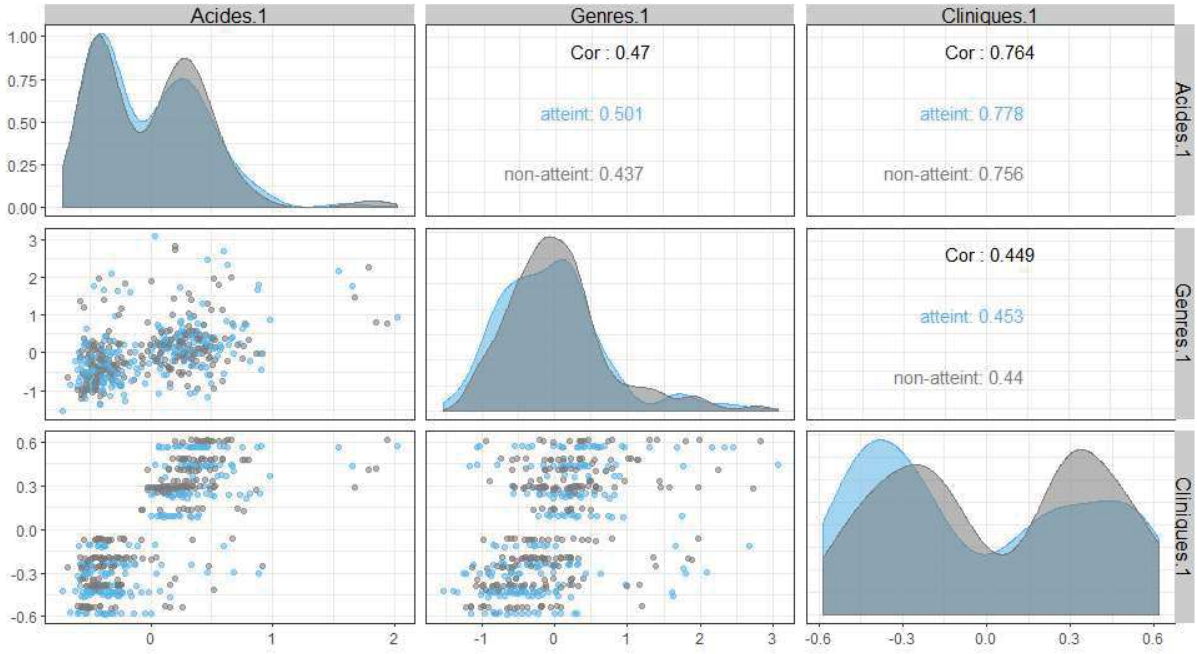

FIGURE 8 – RGCCA réalisée pour les 3 jeux de données (Acides pour les données d'acides biliaires, Genre pour les données de genres et Clinique pour les données cliniques) pour les échantillons provenant des selles. On a extrait une unique composante par blocs.

Une fois encore, on ne peut pas distinguer les enfants atteints de retard de croissance des autres enfants, les densités représentées se superposent pour les deux groupes d'individus. On observe bien la séparation en deux groupes liés au lot pour les représentations faisant intervenir la composante liée aux données de métabolomique (*Acides.1*). Si on s'intéresse aux deux groupes formés par la composante *Clinique.1*, on s'aperçoit qu'ils sont aussi liés au numéro du lot expérimental.

Pour cette méthode, nos conclusions sont similaires à celle obtenues avec la MFA, on n'a pas pu affiner les relations entre acides et genres mais on peut tout de même dire que ces deux représentations se complètent. La RGCCA offre des représentations spécifiques à des composantes choisies et permet d'intégrer un design et des traitements spécifiques à chaque bloc. L'AFM fournit des représentations consensus et extrait des composantes globales aux deux groupes.

5 Conclusions et perspectives

Après toutes les analyses menées sur nos jeux de données, on a pu en partie répondre à nos questions de départ. On a pu intégrer l'information contenus dans les jeux de données que nous avions à disposition par l'intermédiaire de différentes approches. On a montré l'importance de la prise en compte de l'aspect compositionnel des données avant de mener les analyses et grâce à la log-transformation CLR, on a pu s'affranchir des principales contraintes théoriques. Une fois la transformation et les filtrations nécessaires réalisées, on a pu chercher des biomarqueurs du retard de croissance. Dans un premier temps, l'étude séparée des jeux de données a fait ressortir la relative homogénéité des échantillons au niveau global, on a du mal à séparer les enfants en retard de croissance des enfants sains. C'est l'étude dans le détail qui nous a permis de détecter un certain nombre d'acides biliaires et de genres qui discriminent les individus selon leur état de santé. Les résultats que l'on observe vont dans le sens de l'hypothèse formulée par le projet Afribiota : on a pu repérer des marqueurs d'une inflammation intestinale ainsi que des altérations du métabolomes des enfants. L'altération de la digestion et le retard de croissance sont probablement liés à un dérèglement de la composition en acides biliaires primaires et secondaires ainsi qu'à un dérèglement de la composition du microbiote intestinal. Ce dérèglement semble spécifique au pays d'origine des enfants ; si des traitements doivent être administrés, il sera important de prendre en compte ce paramètre.

Cependant, ces conclusions sont limitées dans leur précision et on n'a pas pu mettre en évidence de lien entre des groupes d'acides biliaires et de genres. Intégrer les deux jeux de données dans des analyses et méthodes de visualisation commune comme l'AFM et la RGCCA nous a surtout permis de confirmer la présence d'un très fort effet lot, qui semble influer sur une grande partie de nos analyses sur les acides biliaires. De plus, la présence d'un très grand nombre de valeurs nulles dans les données initiales rend plus complexe les analyses et la méthode qui consiste à les remplacer n'est pas une méthode idéale. Ces limitations font que la différence déjà subtile entre les individus sains et en retard de croissance est difficile à caractériser et identifier, on n'a pas pu identifier de réelle tendance, seulement des variations très locales.

Cela n'a pas été présenté dans le rapport mais on a essayé de corriger l'effet lot observé, cependant cela n'a pas permis de changer nos conclusions. Pour essayer de limiter l'effet du grand nombre de valeurs nulles, on a aussi essayé de réaliser différentes filtrations plus drastiques mais il est compliqué de trouver une filtration adéquate qu'on peut justifier d'un point de vue biologique et statistique.

Pour la suite de ce projet, d'un point de vue de la santé, il sera intéressant de se pencher plus en détail sur les listes de biomarqueurs potentiels que l'on a détecté. L'âge ici semble avoir un rôle mais on a du mal à caractériser les différences entre les enfants des différents groupes d'âge ; s'intéresser à des enfants plus jeunes (entre 0 et 3 ans) permettrait de compléter cette étude. Il sera aussi important de chercher d'autres méthodes de gestion de valeurs nulles dans les compositions : par exemple en utilisant la distribution de Dirichlet [Tsagris and Stewart (2018)] pour modéliser les données, ou bien en obtenant plus d'informations a priori sur les niveaux de détections spécifiques à nos variables, ou encore en agrégeant les données de métataxonomiques à des niveaux supérieurs pour diminuer l'importance des valeurs nulles. Pour compléter l'intégration multiblocs de l'AFM, il faudra aussi adapter l'algorithme pour prendre en compte les données cliniques comme un bloc de variables qualitatives.

Au cours du stage, j'ai été amené à présenter une partie de mon travail à une conférence scientifique de bioinformatique, JOBIM 2019 à Nantes. Ce stage m'a permis de me confronter au monde de la recherche bioinformatique, ainsi qu'à l'importance et la complexité de la mise en commun de données et de fichiers.

Remerciements

Je tiens à remercier Vincent Guillemot et Pascale Vonaesch pour leur suivi au long de ces 6 mois et pour m'avoir permis de réaliser ce stage à l'institut Pasteur. Aussi, je tiens à remercier l'ensemble des équipes du Hub pour leur accueil ainsi qu'aux stagiaires avec qui j'ai pu échanger, découvrir de nouvelles thématiques et progresser dans mon travail. Enfin, je voudrais remercier mes camarades et professeurs de Master sans qui je n'aurai pas pu prendre part à ce projet.

Références

Abdi, H., Williams, L. J., and Valentin, D. (2013). Multiple factor analysis : principal component analysis for multitable and multiblock data sets : Multiple factor analysis. *Wiley Interdisciplinary Reviews : Computational Statistics*, 5(2) :149–179.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Springer Netherlands, Dordrecht. OCLC : 858944307.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2 : High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13 :581.

Dewey, K. G. and Begum, K. (2011). Long-term consequences of stunting in early life : Long-term consequences of stunting. *Maternal & Child Nutrition*, 7 :5–18.

Global Nutrition Report (2018). `https://globalnutritionreport.org/reports/global-nutrition-report-2018/`.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome Datasets Are Compositional : And This Is Not Optional. *Frontiers in Microbiology*, 8.

Greenacre, M. (2016). *Correspondence Analysis in Practice*. Chapman & Hall/CRC.

Greenacre, M. (2018). *Compositional Data Analysis in Practice*. Chapman & Hall/CRC.

Han, J., Lin, K., Sequeira, C., and Borchers, C. H. (2015). An isotope-labeled chemical derivatization method for the quantitation of short-chain fatty acids in human feces by liquid chromatography–tandem mass spectrometry. *Analytica Chimica Acta*, 854 :86–94.

Indahl, U. G., Næs, T., and Liland, K. H. (2016). A similarity index for comparing coupled matrices. *Submitted for review*.

Jandhyala, S. M. (2015). Role of the normal gut microbiota. *World Journal of Gastroenterology*, 21(29) :8787.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR : A package for multivariate analysis. *Journal of Statistical Software*, 25(1) :1–18.

Leo Lahti, S. S. (2012-2019). microbiome r package.

Mann, H. B. and Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1) :50–60.

Martín-Fernández, J. A., Barceló-Vidal, C., and Pawlowsky-Glahn, V. (2003). Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. *Mathematical Geology*, 35(3) :253–278.

Paul J. McMurdie, S. H. (2013). phyloseq : An r package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE*, 8(4) :e61217.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2006). Compositional data and their analysis : an introduction. *Geological Society, London, Special Publications*, 264(1) :1–10.

Poupon, Raoul ; Chignard, N. . R. O. . B. V. . H. C. (2004). La fonction biliaire et sa régulation. *M/S : médecine sciences*, 20(12) :1096–1099.

Tenenhaus, A. and Tenenhaus, M. (2011). Regularized Generalized Canonical Correlation Analysis. *Psychometrika*, 76(2) :257–284.

Tsagris, M. and Stewart, C. (2018). A Dirichlet Regression Model for Compositional Data with Zeros. *Lobachevskii Journal of Mathematics*, 39(3) :398–412.

Vonaesch, P., Randremanana, R., Gody, J.-C., Collard, J.-M., Giles-Vernick, T., Doria, M., Vigan-Womas, I., Rubbo, P.-A., Etienne, A., Andriatahirintsoa, E. J., Kapel, N., Brown, E., Huus, K. E., Duffy, D., Finlay, B., Hasan, M., Hunald, F. A., Robinson, A., Manirakiza, A., Wegener-Parfrey, L., Vray, M., and Sansonetti, P. J. (2018). Identifying the etiology and pathophysiology underlying stunting and environmental enteropathy : study protocol of the AFRIBIOTA project. *BMC Pediatrics*, 18(1) :236.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6) :80.

	Diplôme : Ingénieur Spécialité : Agronome Spécialisation / option : Science des données Enseignant référent : François HUSSON

Auteur(s) : Antoine MENARD Date de naissance* : 30/09/1996	Organisme d'accueil : Institut Pasteur Adresse : 25-28 Rue du Dr Roux, 75015 Paris
Nb pages : Annexe(s) :	
Année de soutenance : 2019	Maître de stage : Vincent GUILLEMOT

Titre français : Analyse de multiples blocs de données compositionnelles : application à l'étude de la sous-nutrition chronique

Titre anglais : Analysis of multiple compositional datasets : application to chronic undernutrition

Résumé (1600 caractères maximum) :

Le retard de croissance touche un quart des enfants de moins de cinq ans dans le monde. C'est pour mieux comprendre les liens entre malnutrition et retard de croissance que s'est mis en place le projet Afribiota. Ce rapport présente le travail réalisé pour intégrer les jeux de données compositionnelles de générés par cette étude pour pouvoir identifier des biomarqueurs du retard de croissance. Le processus et l'importance de l'intégration de l'aspect compositionnel des données sont présentés. Puis, une méthodologie d'intégration par bloc et des méthodes d'intégration multiblocs sont proposées, ainsi que différentes méthodes d'analyses différentielle. Il en résulte que la sous-nutrition chronique engendre un dérèglement de l'écosystème intestinal, un dérèglement des populations bactériennes et des abondances en acides biliaires.

Abstract (1600 caractères maximum) :

Growth delay syndrome affects a quarter of the children who are under five. It is to better understand the links between growth delay and undernutrition that the Afribiota project was set up. This reports presents the work realised in order to integrate the compositional datasets generated by this study and to detect biomarkers of growth delay. The processus and the importance of the integration of the compositional aspect of the data are detailed. Then, a methodology of integration for a single block and two methodologies for multiple blocks are proposed, as well as differential analysis methods. It result that undernutrition generate modifications of the intestinal ecosystem, and more paticularly an imbalance of bacterial populations and a modification of the bile acid relative abondances.

Mots-clés :

Key Words:

* *Elément qui permet d'enregistrer les notices auteurs dans le catalogue des bibliothèques universitaires*

Document à intégrer au mémoire