Projet LSTAT2110(A) – Analyse des données

Analyse de l'emprunte carbone du cycle de vie de denrées alimentaires

Auteurs:

Baily Thomas
(30201700) BIRA2M
Heroufosse Gauthier
(50621700) BIRE1M

Professeur:
Johan Segers

Assistant:
Jean-Loup Dupret

Année académique 2021 - 2022

Contents

1 Introduction

Ce rapport est le produit d'un projet réalisé dans le cadre du cours LSTAT2110A "Analyse de données". Le projet en question vise à implémenter et interpréter les méthodes vues au cours pour analyser une base de données dans un domaine d'application au choix. Notre analyse porte sur la base de données *Environment Impact of Food Production*, issue du site *Kaggle* et disponible à l'adresse suivante: https://www.kaggle.com/selfvivek/environment-impact-of-food-production.

Deux enjeux fondamentaux du XXI^{e} siècle sont la sécurité alimentaire et le changement climatique. Ensemble, ces deux enjeux viennent soulever la question suivante : comment fournir une alimentation saine, nutritive, suffisante et dont la production est respectueuse de l'environnement à l'échelle mondiale ? Actuellement, nous vivons dans un système alimentaire à deux vitesses. Alors que, à l'image de nos populations occidentales, de nombreux individus vivent dans une logique d'abondance à la limite de la surconsommation, une part significative de la population mondiale n'a pas accès à une alimentation suffisante tant d'un point de vue quantitatif que qualitatif. En outre, de nombreuses denrées alimentaires abondamment consommées dans les "pays riches" ont un impact environnemental non négligeable, qui se traduit notamment par leurs empruntes carbones. L'emprunte carbone d'une activité est un indicateur qui mesure la quantité de gaz à effet de serre émis dans l'atmosphère par cette activité. Identifier et comprendre l'impact carbone des différents aliments et des différentes étapes de leur cycle de vie est une étape cruciale permettant des actions ciblées contre le réchauffement climatique.

L'objectif du projet est double:

i) Visualiser différentes denrées alimentaires en les projetant dans un espace de dimensions réduit. Cela facilitera l'étude des liens entre ces denrées et les émissions de gaz à effet de serre des étapes de leur cycle de vie, ainsi que les relations internes entre les denrées alimentaires et les émissions des différentes étapes du cycle de vie.

ii) Classer les différentes denrées alimentaires en groupes les plus homogènes possibles sur base des émissions de gaz à effet de serre des différentes étapes de leur cycle de vie.

Pour atteindre ces objectifs, nous allons, après avoir présenté la base de données et réalisé quelques statistiques descriptives, implémenter une analyse en composantes principales (ACP) et une analyse de classification (clustering).

2 Présentation et analyse descriptive du jeu de données

La base de données contient les 43 aliments les plus cultivés dans le monde et 22 variables indiquant leur impact sur l'utilisation et de l'eau ainsi que leur empreinte carbone. Parmi ces 22 variables, 8 ont été retenues pour cette étude qui s'intéresse aux émissions de gaz à effet de serre des différentes étapes du cycle de vie des denrées alimentaires. Ces variables sont: `Land use change`, `Animal Feed`, `Farm`, `Processing`, `Transport`, `Packaging`, `Retail` et `Total_emissions`. Elles représentent les émissions de gaz à effet de serre par kg de produit alimentaire (kg d'équivalent CO_2 par kg de produit) à différentes étapes du cycle de vie de ces aliments. Les 4 aliments pour lesquels la variable `Land use change` prend des valeurs négatives n'ont pas été retenus pour la présente étude. Après modification de la base de données initiale, nous obtenons une matrice de données de dimension 39x8.

Quelques statistiques de base telles que la moyenne, la médiane et l'écart-type ont été calculées dans le but de décrire les variables étudiées (Table 1). De manière interessante, les étendues et les coefficients de variation relativement élevés indiquent que les émissions de gaz à effet de serre des étapes du cycle de vie peuvent varier fortement d'un aliment à l'autre. En effet, le coefficient de variation est calculé comme le rapport entre l'écart-type et la moyenne. Nous remarquons également que, pour toutes les variables, la moyenne est plus élevées que la médiane. Cela pourrait s'expliquer par le fait que, pour une étape donnée, la majeure partie des aliments étudiés émettent relativement peu de gaz à effet de serre par rapport à quelques aliments

qui présentent des valeurs plus élevées, tirant ainsi la moyenne vers le haut. Cette hypothèse est soutenue par le fait que les médianes sont, pour toutes les variables, beaucoup plus proches de la valeur min que de la valeur max. Cela se confirme visuellement sur les boxplots des différentes variables (Figure 2). La variable `Total_emissions` n'est pas représentée sur le boxplot pour éviter d'écraser davantage les graphiques.

Table 1: Statistiques descriptives

	Land use change	Animal Feed	Farm	Processing	Transport	Packging	Retail	Total_emissions
nbr.val	39	39	39	39	39	39	39	39
nbr.null	14	29	0	16	1	9	25	0
nbr.na	0	0	0	0	0	0	0	0
min	0	0	0.1	0	0	0	0	0.3
max	16	2.9	39	1.3	0.8	1.6	0.3	60
range	16	2.9	39	1.3	0.8	1.6	0.3	59
sum	57	20	142	10	7.6	9.9	3	249
median	0.2	0	0.8	0.1	0.1	0.1	0	2.4
mean	1.5	0.5	3.6	0.26	0.19	0.25	0.077	6.4
SE.mean	0.55	0.15	1.2	0.061	0.025	0.053	0.018	1.8
CI.mean.0.95	1.1	0.31	2.4	0.12	0.051	0.11	0.037	3.5
var	12	0.91	55	0.14	0.025	0.11	0.013	119
std.dev	3.5	0.95	7.4	0.38	0.16	0.33	0.11	11
coef.var	2.4	1.9	2	1.5	0.81	1.3	1.5	1.7

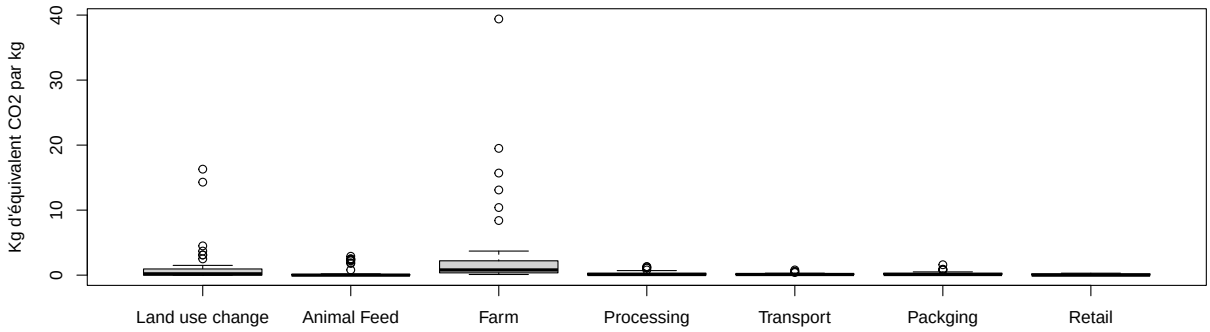

Figure 1: Boxplots des différentes variables

Pour se faire une première idée du lien entre les différentes variables, nous pouvons visualiser les corrélations entre celles-ci. La scatter plots matrice est assez difficile à lire car elle contient 56 graphiques (voir Annexe A). Visuellement, nous pouvons toutefois deviner des corrélations entre les variables. On remarque par exemple que la variables `Total_emissions` est corrélée positivement aux variables `Farm`, `Packaging` et `Processing`. La variable `Farm` semble également corrélée positivement avec les variables `Processing`, `Land use change` et, dans une moindre mesure, avec la variable `Animal Feed`.

Ces premières intuitions sont vérifiées en calculant les coefficients de corrélation (Table 2). Nous remarquons par exemple que la corrélation entre les variables `Farm` et `Total_emissions` vaut 0.97. En dehors des corrélations évidentes susmentionnées, il semblerait que la majorité des variables soient relativement faiblement (mais positivement) corrélées. Certaines variables, comme `Transport` et `Packaging` semblent même indépendantes tant leur coefficient de corrélation est faible (0.0039 dans le cas de ces 2 variables).

Table 2: Matrice des corrélations

	Land use change	Animal Feed	Farm	Processing	Transport	Packging	Retail	Total_emissions
Land use change	1	0.22	0.65	0.48	0.053	0.21	0.15	0.8
Animal Feed	0.22	1	0.57	0.44	0.25	-0.0099	0.5	0.57
Farm	0.65	0.57	1	0.71	0.23	0.17	0.38	0.97
Processing	0.48	0.44	0.71	1	0.28	0.38	0.44	0.73
Transport	0.053	0.25	0.23	0.28	1	0.00039	-0.0068	0.22
Packging	0.21	-0.0099	0.17	0.38	0.00039	1	0.048	0.22
Retail	0.15	0.5	0.38	0.44	-0.0068	0.048	1	0.38
Total_emissions	0.8	0.57	0.97	0.73	0.22	0.22	0.38	1

3 Analyse en composantes principales

Une ACP a été réalisée pour visualiser les différentes denrées alimentaires en les projetant dans un espace de dimensions réduites de façon optimale. Elle permet aussi la visualisation des différentes variables. Le but de cette ACP est d'étudier les liens entre les aliments et les émissions de gaz à effet de serre des étapes de leur cycle de vie ainsi que les relations internes entre les denrées alimentaires et les émissions des différentes étapes du cycle de vie. Pour réaliser l'ACP, deux Packages disponibles dans *R* seront utilisés :

- `FactMineR`, pour l'analyse ;
- `factoextra`, pour la visualisation des données.

Nous avons réalisé l'ACP en excluant la variable `Total_emissons` car il s'agit de la somme de toutes les autres variables.

3.1 Choix du nombre de composantes principales

Afin de retenir le nombre adéquat de composantes principales pour l'ACP, une analyse de la matrice des corrélations entre variables a été réalisée (Table 3).

Table 3: Valeurs propres de la matrice de corrélation

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	3.006	42.95	42.95
comp 2	1.178	16.83	59.78
comp 3	1.01	14.43	74.21
comp 4	0.8396	11.99	86.2
comp 5	0.4531	6.473	92.67
comp 6	0.3292	4.703	97.38
comp 7	0.1837	2.625	100

Ainsi, en utilisant le critère de Kaiser, les trois premières composantes principales seront retenues pour l'analyse. Ces composantes expliquent un total de 74.21% de la variance totale du jeu de données.

3.2 Résultats de l'ACP - Variables

Nous allons maintenant passer à l'analyse des résultats de l'ACP, en commençant par les variables. Les tableaux dont nous discuterons ci-après se trouvent dans l'Annexe B.1.

3.2.1 Coordonnées factorielles

Le premier tableau montre les coordonnées factorielles des variables. L'ACP étant faite sur la matrice de corrélation, il s'agit des corrélations entre chacune des variable de départ et les composantes principales.

On remarque une assez bonne corrélation pour la majorité des variables avec la première composante, à l'exception des variables `Transport` et `Packaging`. Bien que la corrélation soit faible, elle est tout de même positive. Concernant les autres composantes, les résultats sont plus chaotiques. Les deux variables précédemment mal représentées dans la première composante sont cette fois largement majoritaires, sur la seconde composante pour la variable `Packaging`, et la troisième pour la variable `Transport`. Quant aux autres variables, certaines présentent une corrélation négative, comme `Animal Feed` et `Retail` pour respectivement la seconde et la troisième composante ; d'autres sont très proches d'une absence de corrélation. À ce stade-ci, nous pouvons imaginer l'allure des cercles des corrélations. Par exemple, nous nous attendons à des variables assez dispersées dans le demi-cecle de droite sur le cercle des corrélations pour les composantes 1 et 2, à l'exception des variables `Farm`, `Processing` et dans une moindre mesure `Land use change` probablement regroupées.

3.2.2 Qualité de représentation des variables

La seconde propriété analysée est la qualité de représentation des variables par les compostantes. Cette dernière correspond pour une variable i et une composante k au $cos^2\theta_{ik}$. Le tableau 5 montre la qualité de représentation de chaque variable sur les trois premiers axes factoriels. Nous considérons qu'une variable est mal représentée dans un plan factoriel quand elle a un cos^2 inférieur à 0.5. En effet, cela signifie que le résidu (la partie verticale au plan de projection) est plus long que la projection elle-même.

Selon ce critère, on remarque que les variables `Transport` et `Retail` ne sont pas bien représentées dans le plan factoriel déterminé par les composantes principales 1 et 2. D'autre part, les variables `Land use change`, `Animal Feed` et `Packaging` ne sont pas bien représentées dans le plan factoriel déterminé par les composantes principales 1 et 3. Finalement, seule les variables `Transport` et `Packaging` sont bien représentées dans le plan factoriel déterminé par les composantes principales 2 et 3.

3.2.3 Contribution des variables

Le tableau 6 montre les contributions à l'explication de l'inertie (de la variance) par chaque variable dans chacune des trois premières composantes principales (en pourcentages).

On voit que pour la composante principale 1, la plupart des variables ont une contribution similaire à l'explication de l'inertie, à l'exception de `Transport` et `Packging`, qui ont une contribution plus faible. Pour la composante principale 2, la variables `Packging` a la contribution la plus importante (46.55%) et est suivie des variables `Animal Feed`, `Land use change` et `Retail` qui expliquent respectivement à 20.46%, 12.8% et 11.93% de l'inertie. Finalement, seules les variables `Transport` et `Retail` contribuent de manière importante à l'explication de l'inertie dans la troisième composante principale (73.82% pour `Transport` et 25.30% pour `Retail`).

3.2.4 Cercles des corrélations

Les cercles des corrélations permettent de visualiser les résultats présentés ci-dessus. Plus une flèche est longue, plus la variable correspondante est bien représentée dans le plan factoriel considéré. La couleur indique également la qualité de représentation de chaque variable.

Comme escompté, on remarque aisément que la variable `Transport` est mal représentée dans le premier plan factoriel (Figure 2). Au contraire, les variables `Farm`, `Processing`, `Animal Feed` et `Packaging` sont représentées par des flèches qui se rapprochent du cercle de corrélation, indiquant une bonne qualité de représentation. Les qualités de représentation des variables `Land use change` et `Retail` ne sont pas excellentes mais sont considérées comme acceptables. Ce résultat n'est pas inattendu étant donné que la somme de la variance expliquée par ces deux axes n'est que de 59.7%. Il est donc normal d'avoir des variables dont la qualité de représentation est faible.

Le cercle des corrélations permet aussi et surtout de représenter la corrélation entre les variables. Lorsqu'on observe un angle de 90° entre deux flèches, les deux variables correspondantes sont indépendantes. Un angle de 0°C indique quant à lui une corrélation parfaite entre deux variables. Dans notre cas, on observe que la variable `Farm` est positivement corrélée avec les variables `Processing`, `Land use change` et, dans une moindre mesure, avec la variable `Animal Feed`. La corrélation entre les variables `Processing` et `Farm` est confirmée par le cercle des corrélations pour le plan factoriel déterminé par les composantes principales 1 et 3 (Annexe B.1 - figures 7). On observe finalement que les variables `Animal Feed` et `Packaging` semblent indépendantes. Nous ne pouvons rien affirmer par rapport aux variables qui ne sont pas bien représentées dans le premier plan factoriel. Mais sur le cercle des corrélations pour les composantes 1 et 3, on remarque que les variables `Transport` et `Retail` semblent indépendantes. Finalement, le cercle des corrélations pour les composantes 2 et 3 indique que les variables `Transport`et `Packaging` présentent une très faible corrélation négative (Annexe B.1 - figure 8).

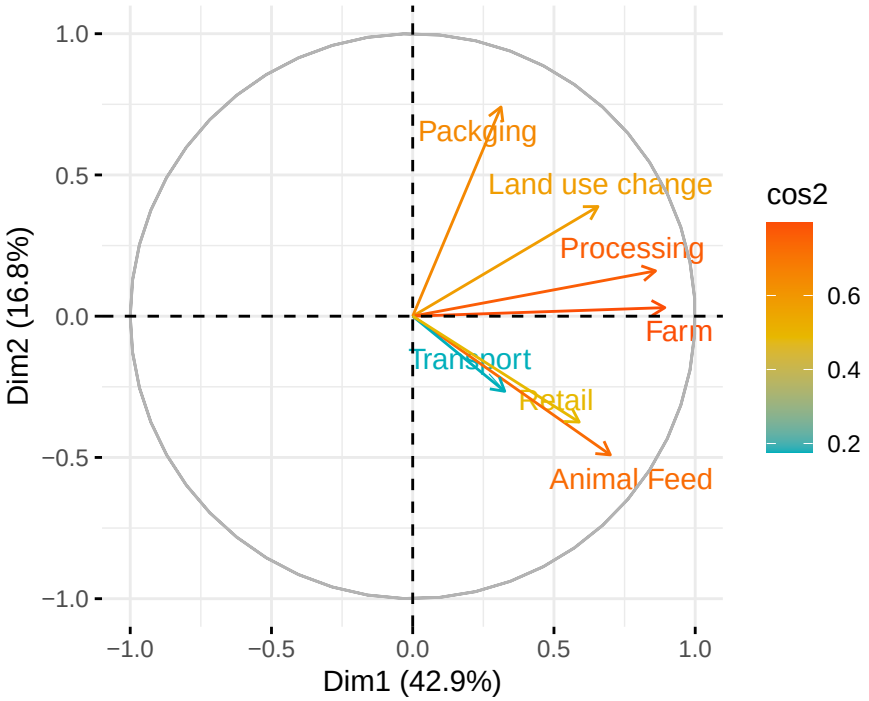

Figure 2: Cercle des corrélations pour les composantes principales 1 et 2

3.3 Résultats de l'ACP - Observations

L'analyse en composantes principales permet également d'analyser la représentation des observations selon les différents plans factoriels. Les tableaux discutés ci-dessous se retrouvent dans l'Annexe B.2.

3.3.1 Coordonnées factorielles des observations

Le tableau 7 montre les coordonnées factorielles des différents aliments pour les différentes composantes principales. On remarque une grande diversité de résultats avec certaines observations particulièrement éloignées sur les axes factoriels. C'est en général le cas des produits d'origine animale, comme les `Beef` ou encore `Lamb & Mutton` sur la première composante principale.

3.3.2 Qualité de représentation des observations

Le tableau 8 montre la qualité de représentation de chaque observation sur les trois premiers axes factoriels. On remarque une grosse disparité entre les observations. Certaines sont très bien représentés sur la première composante principale, à l'image des produits `Berries & Grapes`, `Apples` ou encore `Tomatoes` qui excède 80% de qualité de représentation ; d'autres ne le sont quasiment pas, et les deux autres composantes sont nécessaires pour offrir une qualité suffisante. On peut citer les produits `Milk`, `Rapeseed Oil` et `Beet Sugar`.

3.3.3 Contribution des observations

La contribution de chaque aliment aux différentes composantes principales a également été mesurée (Tableau 9). On s'attend à ce que les productions possédant de grandes valeurs pour une variables contribuent largement aux composantes fortement corrélées avec cette variable. De fait, les types de produits `Beef (beef et dairy herd)` ou encore `Lamb & Mutton`, possédant de grandes valeurs pour la variable `Farm`, sont en tête dans les contributions à la première composante. Concernant la seconde composante, ce sont les produits `Coffee` et `Dark chocolate` qui arrivent en tête. Enfin `Beet Sugar` et `Cane Sugar` contribuent principalement à la troisième composante.

3.3.4 Cartes des individus

Pour visualiser les différents aliments et tenter de déceler des liens entre ceux-ci, nous pouvons représenter les cartes des individus pour les différents plans factoriels (Annexe B.2 - Figures 9 à 11). Les individus sont colorés selon leur qualité de représentation. Il est important de rappeler que les individus mal représentés sur un plan factoriel ne peuvent pas faire l'objet d'interprétations sur la carte des individus correspondante.

Sur la carte des individus obtenue pour les composantes principales 1 et 2, on remarque par exemple que les produits d'origine animale se retrouvent en majeure partie regroupés dans le coin inférieur droit, traduisant des coordonnées factorielles positives sur la composante 1 et négatives sur la composante 2. Ces observations sont pour la plupart assez bien représentées, avec un cos^2 supérieur ou égal à 0.5. On peut en déduire, via l'analyse des variables, que les émissions de gaz à effet de serre de ces aliments sont majoritairement expliquées par les variables `Animal Feed` et `Farm`.

Un autre groupe semble se détacher, comprenant notamment les produits `Coffee`, `Palm Oil` ou `Dark Chocolate`. Ce sont tous des produits qui subissent une transformation après récolte, au contraire des fruits, légumes et céréales principalement retrouvés proche de l'origine. Ces produits prennent une valeur positive sur la seconde composante. Les variables `Packaging` et `Land use change` sont donc responsables de la majorité de leurs émissions. Il convient néanmoins de rester prudent avec ces interprétations, puisque leur qualité de représentation n'est pas toujours bonne (`Dark Chocolate` notamment).

Concernant la troisième composante, il semble plus difficile d'y apercevoir une tendance claire. Pour rappel, la plupart des produits y sont mal représentés (Annexe B.2 - Table 8). Il apparait tout de même que les sucres (`Beet Sugar` et `Cane Sugar`) sont très bien représentés sur cette composante et qu'ils prennent des valeurs positives sur celle-ci. En reliant cette information avec celle fournie par les cercles des corrélations, on déduit que la variable `Transport` est à l'origine d'une grande part des émissions liées à ces denrées.

Afin d'y voir plus clair, tentons maintenant de visualiser la distinction entre les aliments d'origine animale et les aliments d'origine végétale (Figure 3). Nous nous limiterons pour ce cas précis aux composantes 1 et 2, qui expliquent 59,78% de la variance.

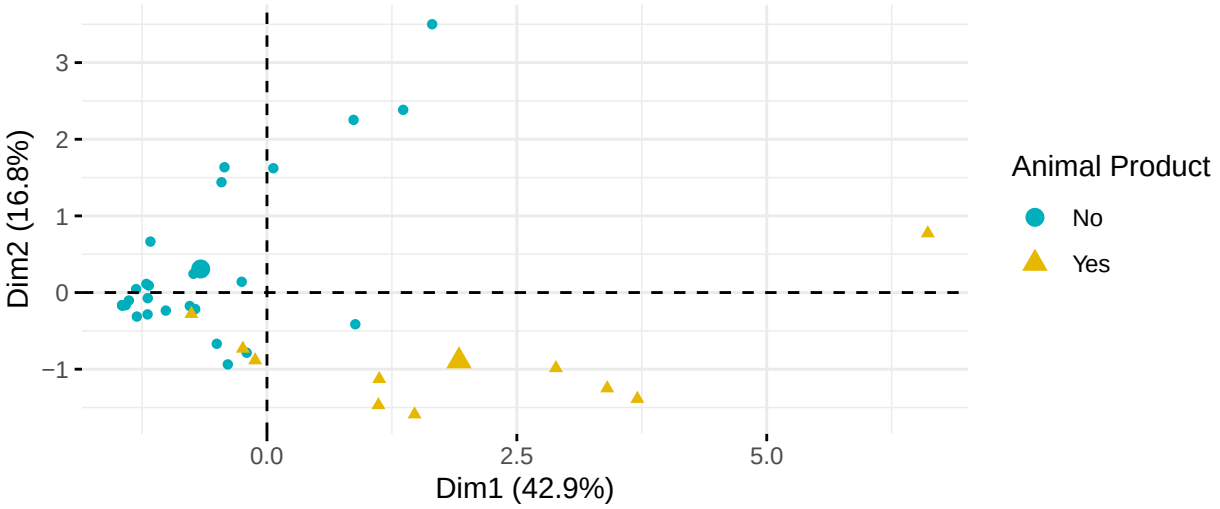

Figure 3: Carte des individus dans le premier plan factoriel

Comme attendu, le premier plan factoriel semble distinguer les deux types d'aliments. Cependant, cette distinction est assez grossière. Le recoupement assez large au centre du plan factoriel entre des aliments d'origine animale et des aliments d'origine végétale semble indiquer que ce choix de groupes n'est pas idéal. Dès lors, il serait approprié d'implémenter une méthode plus précise afin d'effectuer une séparation en groupes. C'est ce que nous allons voir dans la section suivante.

4 Clustering

4.1 Choix de la méthode

Une méthode de clustering va maintenant être implémentée pour répondre au second objectif de ce projet, visant à classer les différentes denrées alimentaires en groupes les plus homogènes possibles selon les émissions de gaz à effet de serre des différentes étapes de leur cycle de vie. En effet, la classification vise à créer des groupes au sein desquels les observations sont similaires, alors que les différences entre les groupes doivent être importantes. Il existe deux grandes sortes de classification: le k-means clustering et la classification hiérarchique. N'ayant pas d'idée sur le nombre de classes à fixer, nous avons décidé d'implémenter une méthode de classification hiérarchique, qui aidera à déterminer ce nombre optimal. Plus précisément, un algorithme hiérarchique ascendant (de la partition la plus fine jusqu'à la plus large) est utilisé ici.

Il existe plusieurs types de classification hiérarchique. L'algorithme de Ward est utilisé pour cette étude. À chaque étape, cet algorithme cherche le regroupement qui minimise l'augmentation de l'inertie au sein des groupes (I_W) et qui, par conséquent, maximise l'augmentation de l'inertie entre les groupes (I_B). En effet, l'inertie totale (I_T) est invariante (ne dépend pas de la partition) et $I_T = I_B + I_W$. Par ailleurs, la méthode pour calculer la distance entre deux points utilisée ici est la distance euclidienne.

Bien que toutes les variables soient exprimées dans les mêmes unités (kg d'équivalent CO_2 par kg de produit), le clustering est appliqué sur les données standardisées pour éviter que les variables avec de grandes valeurs ne dominent.

4.2 Nombre de classes et interprétation

Nous avons réalisé le clustering en excluant la variable `Total_emissons` car elle est calculée comme la somme de toutes les autres variables.

Pour choisir le nombre de classes, nous nous basons sur le dendrogramme (Figure 4) et sur un diagramme en barre qui reprend les niveaux d'agrégation associés à chaque partition (Figure 5). Ces figures indiquent un saut marqué dans les niveaux d'agrégation entre les répartitions en 1 et 2 classes. De manière moins évidente, nous remarquons des sauts dans les niveaux d'agrégation entre les répartitions en 2 et 3 classes ainsi qu'entre les répartitions en 5 et 6 classes. On peut visualiser ces groupes sur le dendrogramme, où les cadres bleus, verts et rouges correspondent respectivement aux partitions en 2, 3 et 6 classes.

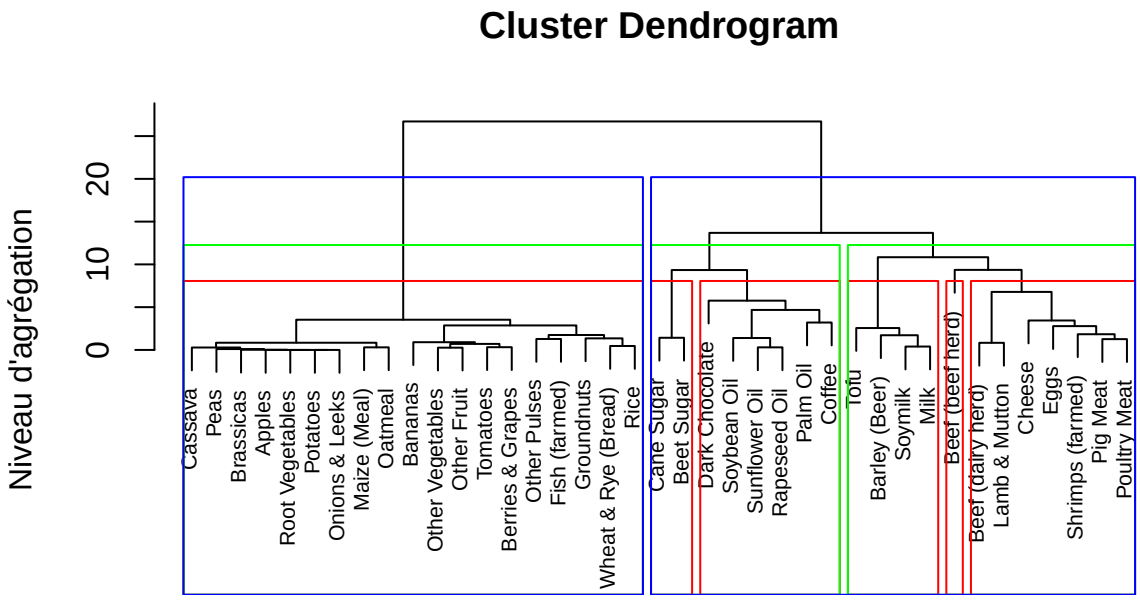

Figure 4: Classification hiérarchique ascendante avec l´algorithme de Ward - dendrogramme

La répartition en 2 classes semble opposer des aliments impliquant d'importantes émissions de gaz à effet de serre (cadre bleu de droite sur le dendrogramme) avec des produits dont la production implique des émissions de gaz à effet de serre plus faibles (cadre bleu de gauche sur le dendrogramme). Cela se vérifie quand on regarde les valeurs prises par la variable `Total_emissions` dans la base de données. Notons une exception pour le riz et le poisson d'élevage dont les émissions totales sont pourtant plus élevées que celles de produits contenus dans la classe "à fortes émissions" tels que le sucre de canne, le tofu et le lait. Cela indique que, au-delà de l'émission totale, la part de chaque étape du cycle de vie est un facteur intervenant dans la répartition. Dès lors, il apparait intéressant de se pencher sur les répartitions en 3 et 6 classes pour voir si des classes plus spécifiques mais pertinentes peuvent être obtenues.

Dans la répartition en 3 classes (cadres verts sur le dendrogramme), la classe contenant les produits "à faibles émissions" obtenue précedemment avec la répartition en 2 classes est préservée (cadre de gauche). La classe de produits "à fortes émissions" est elle splitée en 2. D'une part, nous retrouvons principalement des produits d'origine animale qui sont, en toute logique, caractérisés par de grandes valeurs pour la variable `Animal Feed` (cadre de droite). D'autre part, nous ne retrouvons que des produits d'origine végétale (cadre

vert du milieu). Notons que le tofu, l'orge et le lait de soja se retrouvent dans la même classe que les produits d'origine animale. Au vu du jeu de données, le point commun entre ces 3 produits et les produits d'orine animale est qu'ils présentent une valeur supérieure ou égale à 0.2 pour la variable `Retail`. La majorité des autres aliments présentent une valeur nulle pour cette variable.

Finalement, dans la répartition en 6 classes (cadres rouges sur le dendrogramme), plusieurs points intéressants sont à noter:

1) La classe contenant les produits "à faibles émissions" obtenue pour les répartitions à 2 et 3 classes est préservée (premier cadre en partant de la gauche).

2) La classe contenant les produits à "fortes émissions et d'origine végétale" obtenue avec la partition en 3 classes est splitée en 2 classes. La première (deuxième cadre en partant de la gauche sur le dendrogramme) contient le sucre de canne et le sucre de betterave, caractérisés par des valeurs importantes pour la variable `Transport` par rapport aux autres aliments. La deuxième classe (troisième cadre en partant de la gauche) contient les huiles végétales ainsi que le café et le chocolat. Ces produits présentent des valeurs relativement élevées pour les variables `Packaging` et `Farm` par rapport aux autres produits d'origine végétale.

3) La classe contenant principalement des produits d'origine animale obtenue avec la partion en 3 classes est splitée en 3 classes. Parmi celles-ci, 2 classes contiennent exclusivement des produits d'orignie animale (premier et deuxième cadres en partant de la droite). Notons qu'une de ces deux classes ne contient qu'un produit (Beef (beef herd)), qui est caractérisé par des valeurs particulièrement élevées pour les variables `Farm` et `Land use change`. La troisième classe (troisième cadre) contient, en plus du lait, les produit d'origine végétale qui étaient, pour la partition en 3 classes, mélangés aux produits d'origine animale. Comme mentionné plus haut, ces produits sont caractérisés par des valeurs relativement importantes pour la variable `Retail`.

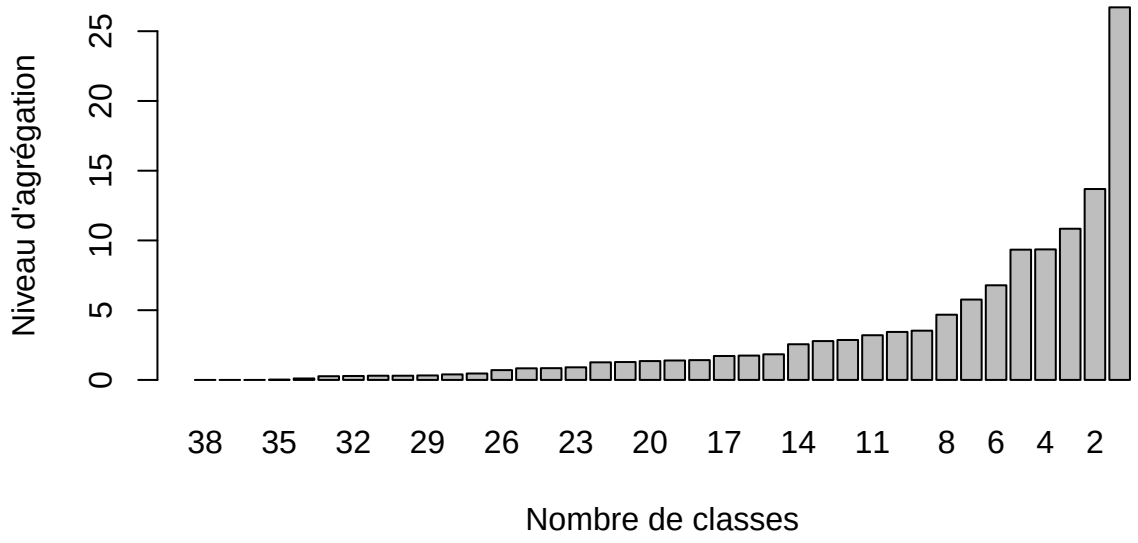

Figure 5: Niveaux d´agrégation des partitions obtenues par classification hiérarchique ascendante (algorithme de Ward)

4.3 Qualité des partitions

La qualité d'une partition définit le pourcentage de l'inertie totale (I_T) expliquée par l'inertie entre les groupes (I_B). Pus le nombre de classes est faible, plus la qualité de la partition est faible car il y a un transert de I_B vers I_W. Les qualités des partitions en 6, 3 et 2 classes valent respectivement 56.72%, 32.77% et 21.67%.

5 Conclusion

Ce travail visait à étudier l'emprunte carbone des différentes étapes du cycle de vie de 39 denrées alimentaires largement consommées dans le monde.

La première partie regroupe quelques statistiques descriptives qui ont permis de dégager certaines tendances. Nous avons notamment remarqué que la majeure partie des aliments étudiés émettent relativement peu de gaz à effet de serre par rapport à quelques aliments qui présentent des valeurs plus élevées. Des corrélations entre certaines variables ont également pu être établies. De manière intéressante, la variable `Farm` présente une corrélation de 0.97 avec la variable `Total_emissions`, indiquant l'impact très important de l'étape "Ferme" sur les émissions de gaz à effet de serre associées à un aliment.

Deux méthodes d'analyse de données multivariées ont ensuite été implémentées pour analyser les données plus en profondeur: une ACP et une classification hiérarchique ascendante.

L'ACP a notamment permis de visualiser les corrélations remarquées lors de l'analyse descriptive. Nous avons par exemple remarqué que a variable `Farm` est corrélée positivement avec les variables `Processing`, `Land use change` et, dans une moindre mesure, avec la variable `Animal Feed`. La variable `Transport` est quant à elle probablement indépendante des autres étape du cycle de vie des produits. De plus, l'analyse en composantes principales a permis d'établir des liens entre les variables et les aliments mais également entre les aliments. Par exemple, de manière logique, nous avons remarqué que les aliments d'origine animale présentent des valeurs élevées pour les variables `Animal Feed` et `Farm`. Notons toutefois que l'interprétation des résultats de l'ACP n'est pas évidente. Cela s'explique par le fait que les deux premières composantes principales ne capturent que 59.78% de la variabilité. La troisième composante est nécessaire pour atteindre 74.21%, ce qui reste encore plutôt modeste. Cela fait que la qualité de représentation des variables et des observations dans les trois plans factoriels étudiés laisse parfois à désirer. Finalement, l'élaboration de groupes d'individus sur unique base de cette ACP est laborieuse, indiquant la nécessité d'implémenter une méthode de clustering. Nous soupçonnons tout de même certaines tendances comme un regroupement des produits d'origine animale.

Le clustering hiérarchique ascendant (avec l'algorithme de Ward) a permis l'identification de deux groupes principaux : les denrées "à fortes émissions" et les denrées "à faibles émissions". La composition du groupe "à faibles émissions" est préservée (identique) pour les répartitions en 3 ou 6 classes. Il est intéressant de noter que cette classe contient, à l'exception du poisson d'élevage, uniquement des produit d'origine végétale. Ensuite, nous avons vu qu'il était intéressant de décomposer la classe contenant les denrées "à fortes émissions". Par exemple, nous avons pu regrouper la majorité des aliments d'origine animale et faire des distinctions entre les aliments d'origine végétale présents dans cette classe. Toutefois, il est important de rappeler que les qualités des partitions sont assez faibles, surtout pour les répartitions en 2 et 3 classes.

Finalement, pour consolider notre analyse, il serait intéressant de réaliser des classifications avec la méthode des k-means pour 2, 3 et 6 classes. Les centres des classes, choisis à priori, seraient les centres des classes que nous avons obtenues par classification hiérarchique.

6 Bibliographie

- https://www.kaggle.com/selfvivek/environment-impact-of-food-production
- https://ourworldindata.org/environmental-impacts-of-food
- https://www.science.org/doi/pdf/10.1126/science.aaq0216
- http://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/73-acp-analyse-en-composantes-principales-avec-r-l-essentiel/#packages-r
- http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Nb_Components_PCA.pdf

7 Annexes

Annexe A : Statistiques descriptives

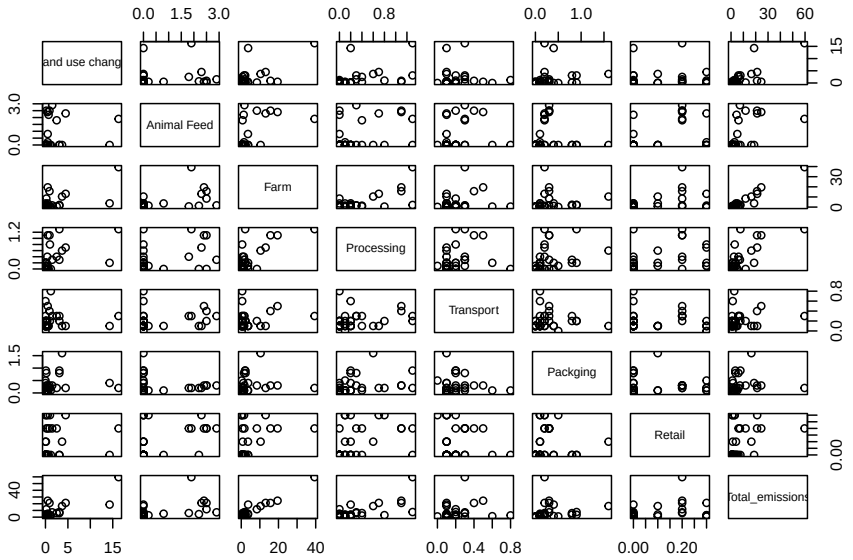

Figure 6: Scatter plot matrix

Annexe B.1 : Analyse composantes principales - Variables

Table 4: Coordonnées factorielles des variables

	Dim.1	Dim.2	Dim.3
Land use change	0.656	0.388	-0.018
Animal Feed	0.7	-0.491	-0.064
Farm	0.893	0.03	0.019
Processing	0.86	0.16	0.063
Transport	0.326	-0.266	0.863
Packging	0.313	0.741	0.007
Retail	0.589	-0.375	-0.506

Table 5: Qualité de représentation des variables (Cos^2)

	Dim.1	Dim.2	Dim.3
Land use change	0.43	0.151	0
Animal Feed	0.49	0.241	0.004
Farm	0.797	0.001	0
Processing	0.739	0.026	0.004
Transport	0.106	0.071	0.746
Packging	0.098	0.548	0
Retail	0.347	0.14	0.256

Table 6: Contribution des variables à l'élaboration des composantes principales

	Dim.1	Dim.2	Dim.3
Land use change	14.3	12.8	0.032
Animal Feed	16.28	20.46	0.409
Farm	26.5	0.078	0.038
Processing	24.58	2.183	0.392
Transport	3.529	6.001	73.82
Packging	3.25	46.55	0.004
Retail	11.56	11.93	25.3

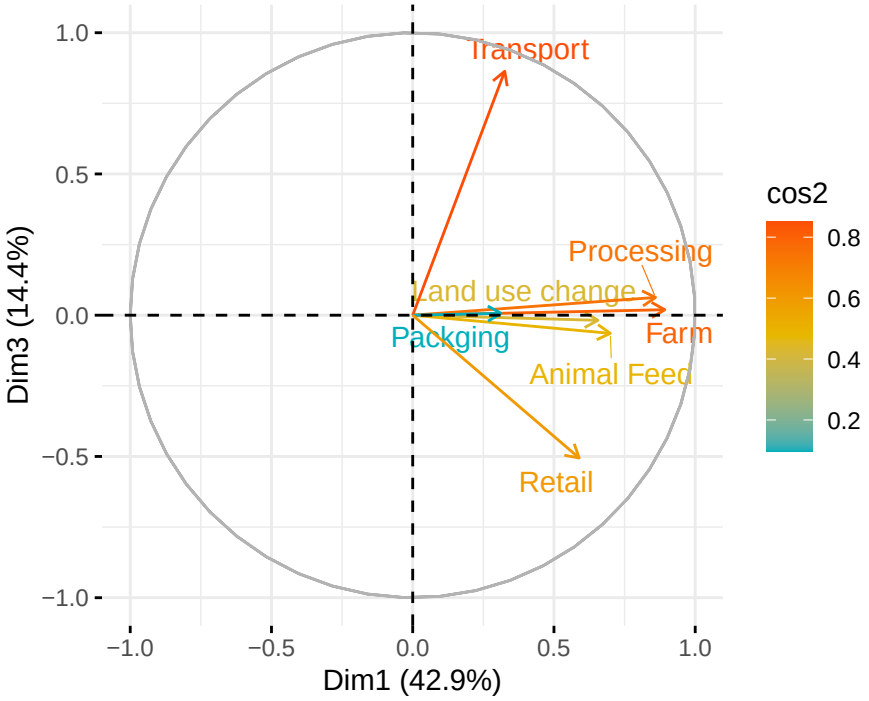

Figure 7: Cercle des corrélations pour les composantes principales 1 et 3

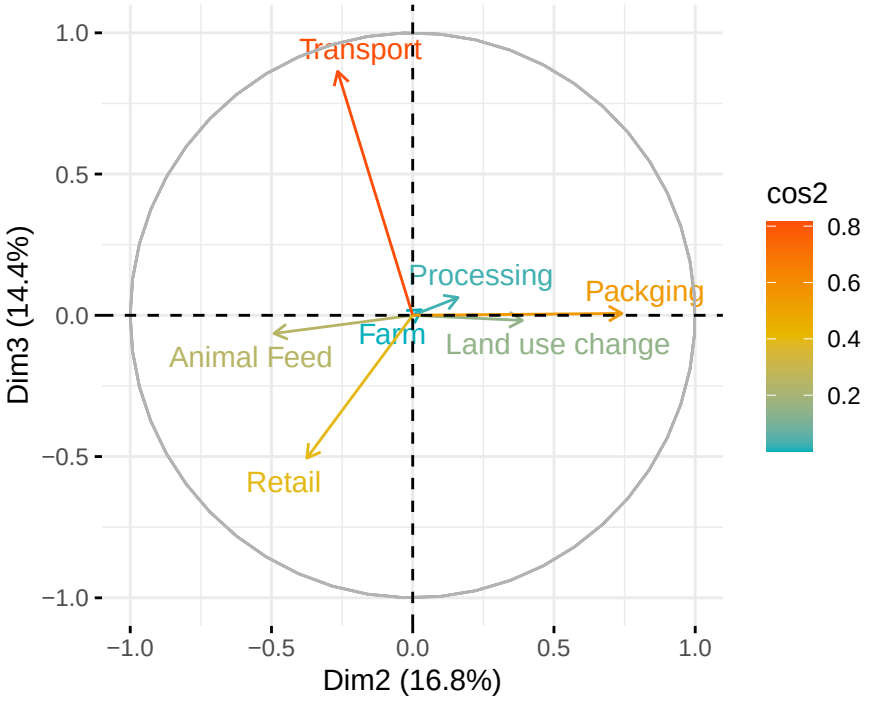

Figure 8: Cercle des corrélations pour les composantes principales 2 et 3

Annexe B.2 : Analyse composantes principales - Observations

Table 7: Coordonnées factorielles des observations

	Dim.1	Dim.2	Dim.3
Wheat & Rye (Bread)	-0.772	-0.176	-0.608
Maize (Meal)	-1.206	0.113	-0.178
Barley (Beer)	-0.253	0.141	-2.07
Oatmeal	-1.309	0.045	-0.19
Rice	-0.718	-0.215	-0.617
Potatoes	-1.448	-0.167	-0.196
Cassava	-1.381	-0.104	-0.199
Cane Sugar	-0.392	-0.938	3.677
Beet Sugar	-0.502	-0.669	2.609
Other Pulses	-1.166	0.665	-0.185
Peas	-1.413	-0.165	-0.194
Groundnuts	-0.735	0.245	-0.126
Soymilk	-0.203	-0.785	-1.509
Tofu	0.884	-0.414	-0.856
Soybean Oil	0.064	1.622	0.965
Palm Oil	1.363	2.384	0.582
Sunflower Oil	-0.425	1.635	0.414
Rapeseed Oil	-0.455	1.439	0.412
Tomatoes	-1.193	-0.074	0.359
Onions & Leeks	-1.448	-0.167	-0.196
Root Vegetables	-1.448	-0.167	-0.196
Brassicas	-1.441	-0.166	-0.195

	Dim.1	Dim.2	Dim.3
Other Vegetables	-1.194	-0.285	0.375
Bananas	-1.012	-0.235	0.931
Apples	-1.448	-0.167	-0.196
Berries & Grapes	-1.182	0.092	0.363
Other Fruit	-1.302	-0.313	0.358
Coffee	1.653	3.501	-0.505
Dark Chocolate	0.867	2.253	-0.22
Beef (beef herd)	6.611	0.774	0.124
Beef (dairy herd)	3.404	-1.249	0.624
Lamb & Mutton	3.706	-1.387	1.196
Pig Meat	1.477	-1.59	-0.131
Poultry Meat	1.124	-1.128	-0.05
Milk	-0.118	-0.884	-1.537
Cheese	2.891	-0.985	-1.568
Eggs	-0.241	-0.732	-0.342
Fish (farmed)	-0.756	-0.278	-0.242
Shrimps (farmed)	1.115	-1.469	-0.683

Table 8: Qualité de représentation des observations

	Dim.1	Dim.2	Dim.3
Wheat & Rye (Bread)	0.476	0.025	0.296
Maize (Meal)	0.798	0.007	0.017
Barley (Beer)	0.009	0.003	0.615
Oatmeal	0.816	0.001	0.017
Rice	0.403	0.036	0.298
Potatoes	0.806	0.011	0.015
Cassava	0.769	0.004	0.016
Cane Sugar	0.009	0.052	0.802
Beet Sugar	0.031	0.055	0.832
Other Pulses	0.646	0.21	0.016
Peas	0.785	0.011	0.015
Groundnuts	0.322	0.036	0.009
Soymilk	0.008	0.118	0.435
Tofu	0.119	0.026	0.112
Soybean Oil	0.001	0.614	0.217
Palm Oil	0.148	0.451	0.027
Sunflower Oil	0.037	0.554	0.035
Rapeseed Oil	0.056	0.561	0.046
Tomatoes	0.833	0.003	0.076
Onions & Leeks	0.806	0.011	0.015
Root Vegetables	0.806	0.011	0.015
Brassicas	0.802	0.011	0.015
Other Vegetables	0.739	0.042	0.073
Bananas	0.511	0.028	0.433
Apples	0.806	0.011	0.015
Berries & Grapes	0.872	0.005	0.082
Other Fruit	0.777	0.045	0.059
Coffee	0.14	0.629	0.013
Dark Chocolate	0.048	0.327	0.003

	Dim.1	Dim.2	Dim.3
Beef (beef herd)	0.803	0.011	0
Beef (dairy herd)	0.759	0.102	0.025
Lamb & Mutton	0.723	0.101	0.075
Pig Meat	0.264	0.306	0.002
Poultry Meat	0.316	0.318	0.001
Milk	0.003	0.156	0.472
Cheese	0.704	0.082	0.207
Eggs	0.012	0.112	0.025
Fish (farmed)	0.332	0.045	0.034
Shrimps (farmed)	0.183	0.318	0.069

Table 9: Contribution des observations

	Dim.1	Dim.2	Dim.3
Wheat & Rye (Bread)	0.508	0.068	0.939
Maize (Meal)	1.241	0.028	0.08
Barley (Beer)	0.054	0.043	10.88
Oatmeal	1.461	0.004	0.092
Rice	0.44	0.101	0.967
Potatoes	1.788	0.06	0.097
Cassava	1.628	0.023	0.1
Cane Sugar	0.131	1.914	34.32
Beet Sugar	0.215	0.974	17.28
Other Pulses	1.159	0.962	0.087
Peas	1.702	0.059	0.096
Groundnuts	0.461	0.13	0.04
Soymilk	0.035	1.342	5.781
Tofu	0.666	0.373	1.861
Soybean Oil	0.003	5.726	2.364
Palm Oil	1.585	12.37	0.86
Sunflower Oil	0.154	5.82	0.435
Rapeseed Oil	0.176	4.51	0.431
Tomatoes	1.213	0.012	0.328
Onions & Leeks	1.788	0.06	0.097
Root Vegetables	1.788	0.06	0.097
Brassicas	1.771	0.06	0.097
Other Vegetables	1.217	0.177	0.357
Bananas	0.873	0.121	2.199
Apples	1.788	0.06	0.097
Berries & Grapes	1.192	0.018	0.335
Other Fruit	1.445	0.213	0.326
Coffee	2.33	26.67	0.648
Dark Chocolate	0.641	11.05	0.123
Beef (beef herd)	37.28	1.303	0.039
Beef (dairy herd)	9.885	3.398	0.987
Lamb & Mutton	11.71	4.185	3.633
Pig Meat	1.86	5.501	0.044
Poultry Meat	1.077	2.769	0.006
Milk	0.012	1.7	5.998
Cheese	7.128	2.113	6.24
Eggs	0.049	1.166	0.297

	Dim.1	Dim.2	Dim.3
Fish (farmed)	0.487	0.169	0.148
Shrimps (farmed)	1.06	4.695	1.185

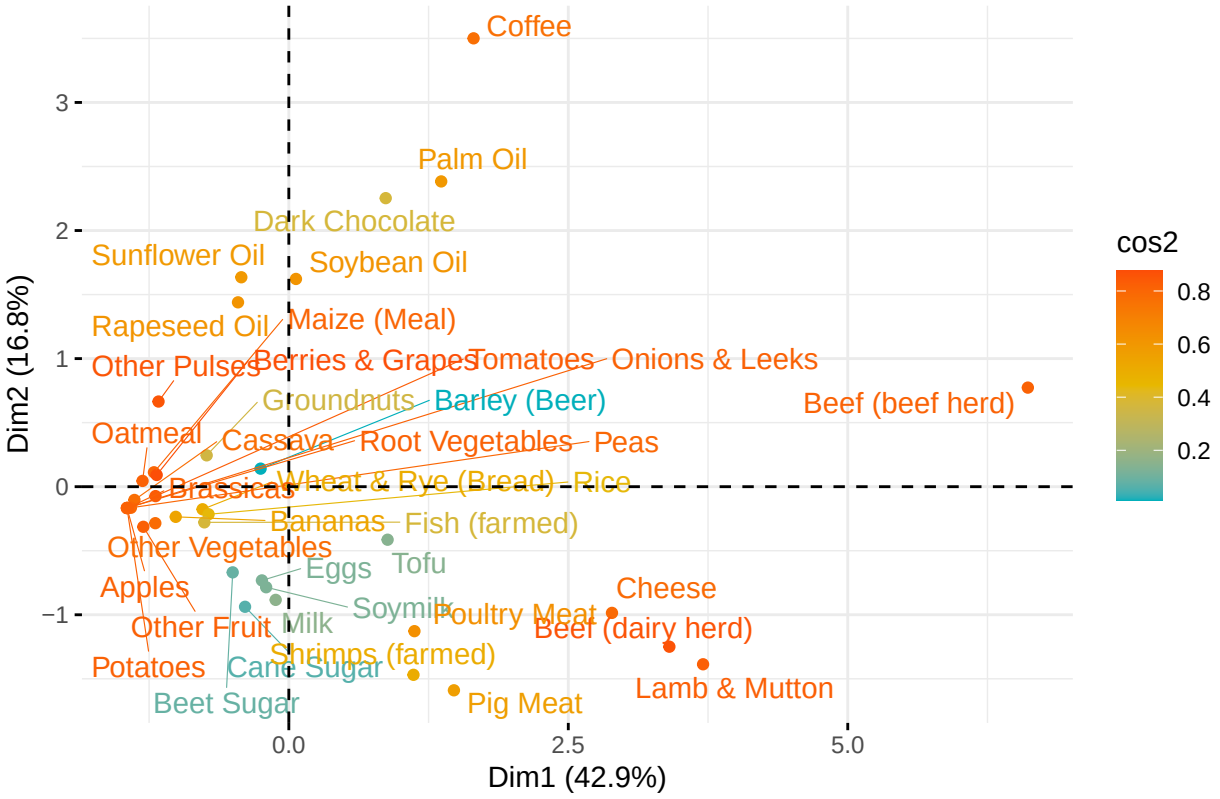

Figure 9: Carte des individus selon les composantes principales 1 et 2

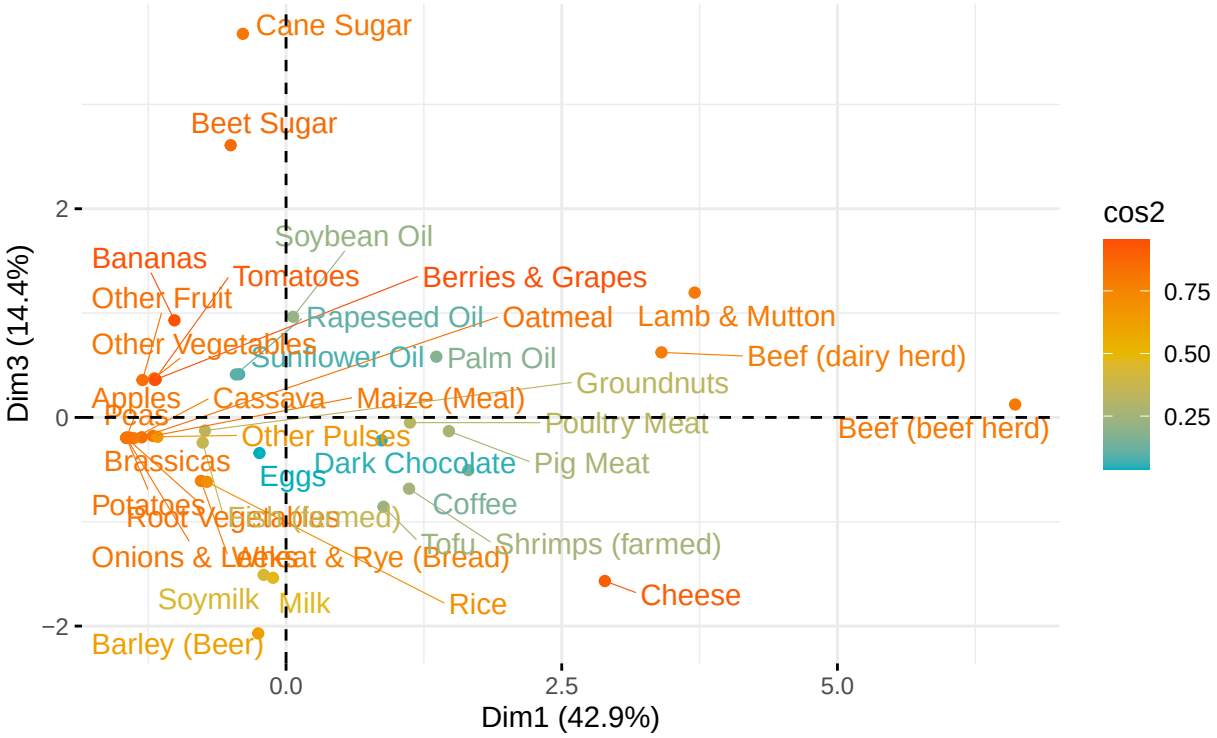

Figure 10: Carte des individus selon les composantes principales 1 et 3

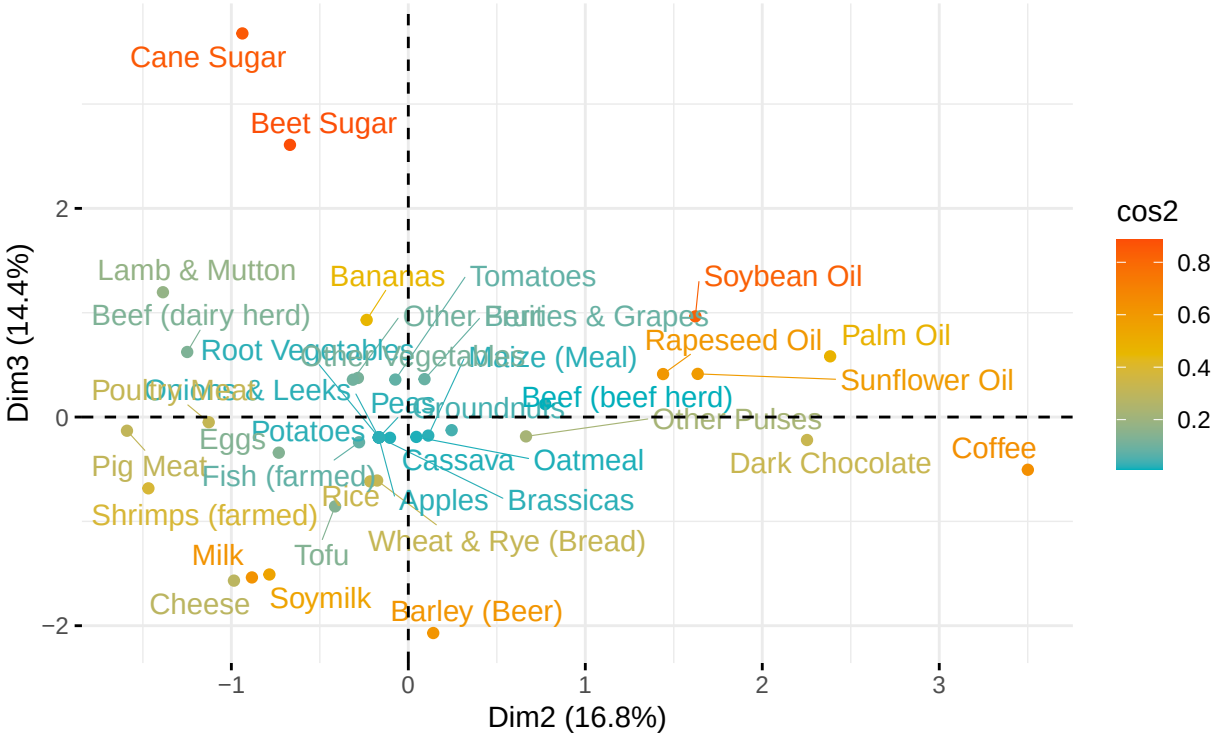

Figure 11: Carte des individus selon les composantes principales 2 et 3

Annexe C : Script du rapport entier

```
# Initialisation ----

# Chargemeent des données
donnees <- fread("Food_Production.csv")
donnees <- as.data.frame(donnees)
# Nomme les colonnes par le nom du produit alimentaire correspondant
names <- donnees$`Food product`
rownames(donnees) <- names
# Elimination des produits qui ont une valeur négative pour la variable Lande use change
donnees <- donnees %>% filter(`Land use change`>=0)
donnees <- donnees[,c("Land use change", "Animal Feed","Farm","Processing", "Transport","Packging","Reta

# Satistiques descriptives ----
set.caption('Statistiques descriptives') #Légende du tableau
panderOptions('table.split.table', 300) # Pour éviter de spliter les tableaux
pander(apply(donnees,2,stat.desc), digits = 2, cex = 0.7)

boxplot(donnees[,1:7], ylab = "Kg d'équivalent CO2 par kg")

pairs(donnees)

# Corrélations entre les variables
set.caption('Matrice des corrélations')
pander(cor(donnees), digits = 2, cex = 0.1)
panderOptions('table.split.table', 300)

Data_ACP <- subset(donnees, select = -c(Total_emissions))

# ACP ----

# Ajout d'une variable catégorielle
x=c()
for (i in 1:29) {
  x[i] <- "No"

}
y= c()
for (i in 1:10) {
  y[i] <- "Yes"
}



Product_type <- c(x,y)

Data_ACP <- cbind(Data_ACP,Product_type)
res <- PCA(Data_ACP[,-8], graph=FALSE)

# Choix composantes principales
set.caption('Valeurs propres de la matrice de corrélation')
eig <- res$eig
pander(eig)
```

```r
## Variables ----

# Coordonnées factorielles des variables
coord <- round(res$var$coord[,1:3],3)
set.caption('Coordonnées factorielles des variables')
pander(coord)

# Qualité de représentation des variables
cos2 <- round(res$var$cos2[,1:3],3)
set.caption('Qualité de représentation des variables (Cos^2)')
pander(cos2)

# Contribution des variables
contrib <- round(res$var$contrib[,1:3],3)
set.caption("Contribution des variables à l'élaboration des composantes principales")
pander(contrib)


# Cercles des corrélations
options(ggrepel.max.overlaps = Inf)
fviz_pca_var(res,axes = c(1,2), col.var = "cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), r


options(ggrepel.max.overlaps = Inf)
fviz_pca_var(res, axes = c(1,3), col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE )

options(ggrepel.max.overlaps = Inf)
fviz_pca_var(res, axes = c(2,3), col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE )

## Individus ----

# Coordonnées factorielles des individus
coord_ind <- round(res$ind$coord[,1:3],3)
set.caption('Coordonnées factorielles des individus')
pander(coord_ind)

# Qualité de représentation des observations
coord_ind <- round(res$ind$cos2[,1:3],3)
set.caption('Qualité de représentation des observations')
pander(coord_ind)

# Contributions des observations
contrib_ind <- round(res$ind$contrib[,1:3],3)
set.caption('Contribution des observations')
pander(contrib_ind)

# Contributions des observations
dist_ind <- round(res$ind$dist,3)
set.caption("Distance à l'origine des observations")
pander(dist_ind)
```

```r
# Distance à l'origine
dist_ind <- round(res$ind$dist,3)
set.caption("Distance à l'origine des observations")
pander(dist_ind)

# Cartes des individus

options(ggrepel.max.overlaps = Inf)
fviz_pca_ind(res, axes = c(1,2), col.ind = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE # Évite le chevauchement de texte
             )

options(ggrepel.max.overlaps = Inf)
fviz_pca_ind(res, axes = c(1,3), col.ind = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE # Évite le chevauchement de texte
             )

options(ggrepel.max.overlaps = Inf)
fviz_pca_ind(res, axes = c(2,3), col.ind = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE # Évite le chevauchement de texte
             )

options(ggrepel.max.overlaps = Inf)
fviz_pca_ind(res,
             geom.ind = "point",
             col.ind = Data_ACP$Product_type, # colorer by groups
             palette = c("#00AFBB", "#E7B800", "#FC4E07"),
             addEllipses = TRUE,
             mean.point = TRUE,
             legend.title = "Product type"
             )

# Clustering ----

# Standardisation des données
donnees <- scale(donnees)
donnees <- donnees[,1:7]

# Distances euclidiennes
dist_donnees <- dist(donnees, method = "euclidean")

# Classification hiérarchique algorithme de Ward
res_clust <- hclust(dist_donnees, method = "ward.D")

plot(res_clust,
     xlab = "",
     ylab = "Niveau d'agrégation",
     cex = 0.7)

rect.hclust(res_clust, k=6, border="red")
```

```r
rect.hclust(res_clust, k=3, border="green")
rect.hclust(res_clust, k=2, border="blue")

# Barplot
barplot(res_clust$height,
        xlab = "Nombre de classes",
        names.arg = (nrow(donnees)-1):1,
        ylab = "Niveau d'agrégation")

#qualité partition - 6 classes
BSS_W <- sum(tail(res_clust$height,n=(6-1)))
TSS_W <- sum(res_clust$height)
BSS_W/TSS_W*100

#qualité partition - 3 classes
BSS_W <- sum(tail(res_clust$height,n=(3-1)))
TSS_W <- sum(res_clust$height)
BSS_W/TSS_W*100

#qualité partition - 2 classes
BSS_W <- sum(tail(res_clust$height,n=(2-1)))
TSS_W <- sum(res_clust$height)
BSS_W/TSS_W*100
```