Thème : Graine de café

Spécialité : Analyse de données exploratoire

Yasemin AKDAG - Melissa Zennaf - Amadou Diakhate DIOP

Contents

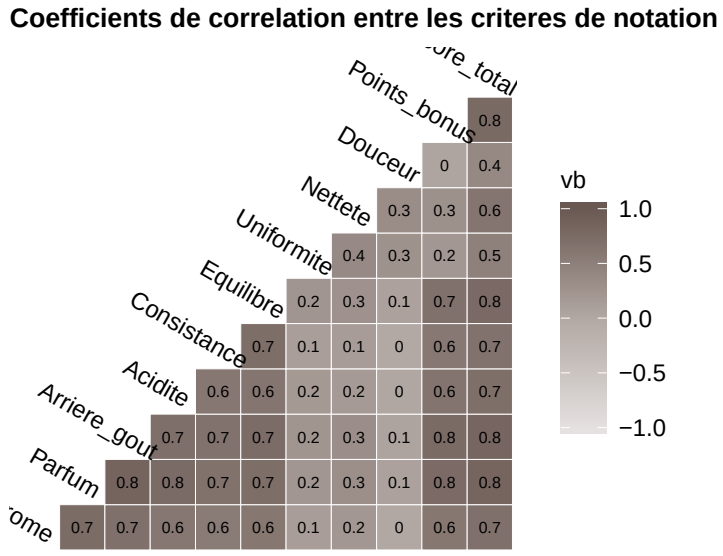
1 ANALYSE DES VARIABLES QUANTITATIVES

Dans cette partie, nous allons réaliser une analyse approfondie des variables quantitatives présentes dans la base de données "Coffee". Pour ce faire, nous allons commencer par étudier les liens apparents existants entre nos variables. Par la suite, nous utiliserons l'analyse factorielle de deux manières. Dans un premier temps, l'analyse par composantes principales nous permettra de faire apparaître des groupes d'individus. Dans un second temps, une analyse des correspondances multiples nous permettra de faire apparaître de nouveaux groupes. Ces deux classifications seront ensuite comparées.

1.1 ETUDES DES CORRELATIONS

Dans notre base de données, il y a des variables plus riches d'information que d'autres. Comme l'étude des corrélations le montre, on peut distinguer les variables de scoring des autres variables quantitatives. Il en existe 10 qui sont :

- Le critère arôme (Aroma)
- Le critère parfum (Flavor)
- Le critère arrière-goût (Aftertaste)
- Le critère acidité (Acidity)
- Le critère équilibre (Balance)
- Le critère uniformité (Uniformity)
- Le critère netteté (Clean cup)
- Le critère douceur (Sweetness)
- Le critère consistance (Body)
- Le critère bonus (Cupper points)



Les corrélations qui sont relativement importantes sont celles entre les variables de scoring. On voit aussi que les variables d'altitude sont logiquement corrélées entre elles. Dans les variables de scoring, il existe une grande corrélation entre les variables d'arôme, parfum, arrière-goût, acidité, consistance, équilibre. Nous avons également des coefficients de corrélation qui sont plus faibles entre les variables uniformité, netteté, douceur et les autres variables. Ces variables sont aussi peu corrélées entre elles. Il y a une corrélation assez importante entre les variables de notation et la variable score total. Cependant, les variables uniformité, netteté et douceur sont peu corrélées au score total.

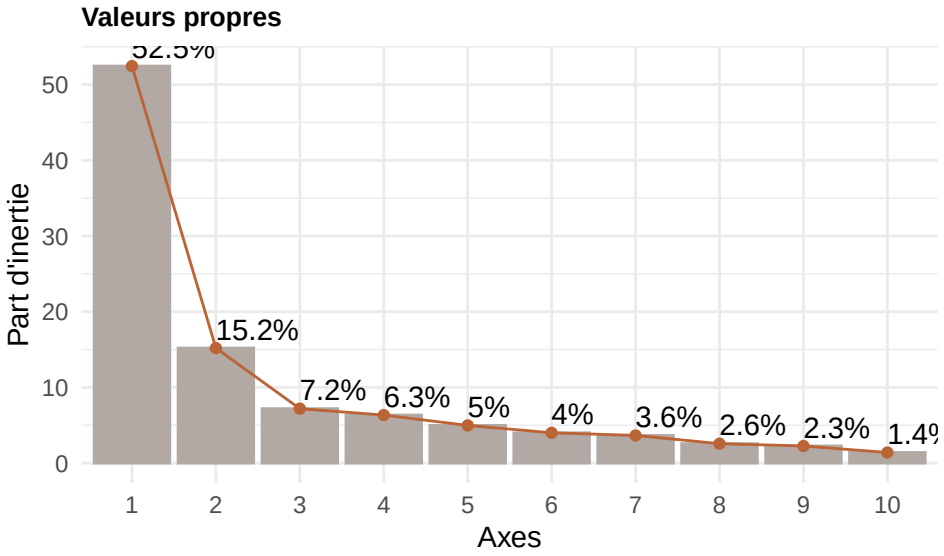
Nous pouvons donc distinguer deux types de variables de scoring:

- Les scores de goût : arôme, parfum, arrière-goût, acidité, consistance, équilibre
- Les scores d'apparence : uniformité, netteté et douceur

Le résultat du test du Chi2 est sans appel: les variables ne sont pas indépendantes.

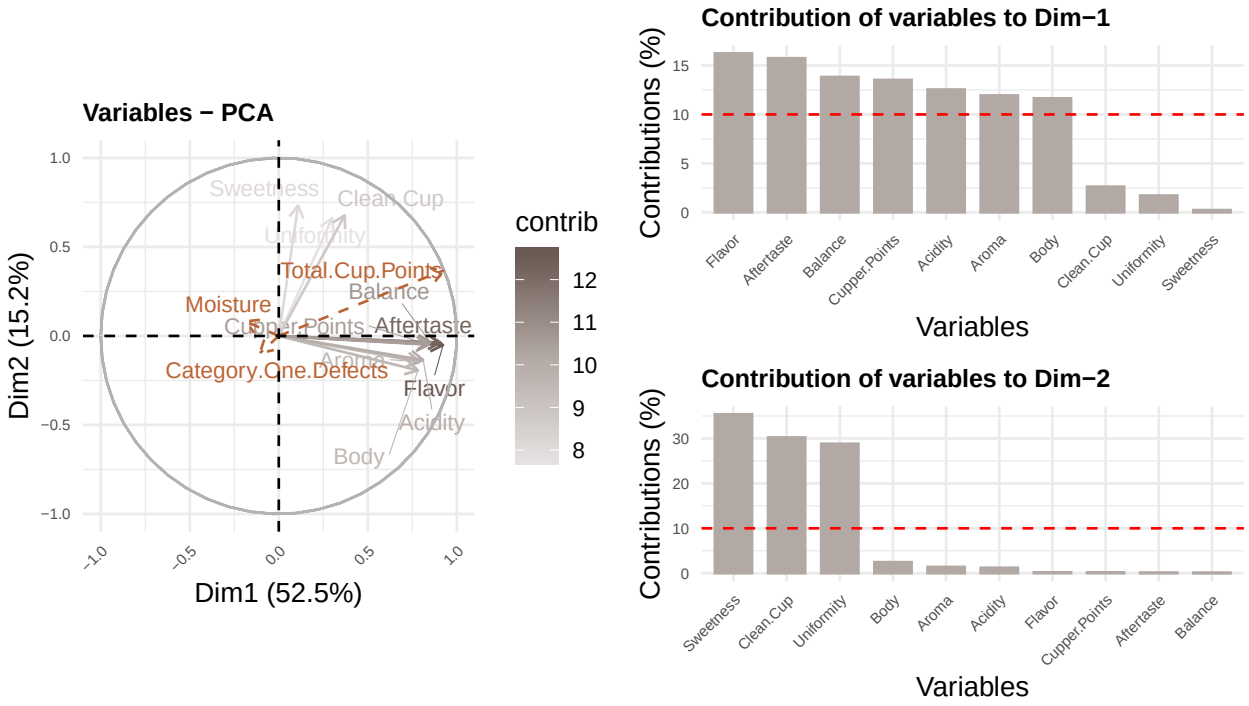
1.2 ANALYSE PAR COMPOSANTES PRINCIPALES

Nous effectuons donc notre ACP sur nos variables de scoring.



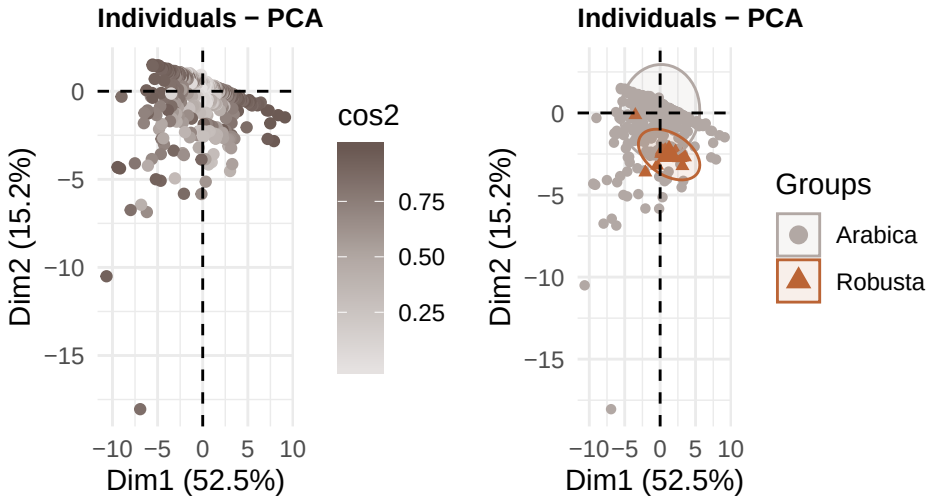
Sur ce graphique, nous voyons qu'avec les deux premiers axes, la variance expliquée est de 67.7% ce qui est acceptable. Nous allons donc conserver ces deux axes pour la suite de notre analyse.

Il y a deux groupes qui se distinguent et qui correspondent à nos observations de la matrice des corrélations.



Les scores de goût contribuent le plus à l'axe 1, les scores d'apparence contribuent le plus à l'axe 2 et les deux groupes se distinguent beaucoup. Concernant les variables quantitatives supplémentaires, on a une très bonne représentation du score total qui pointe dans une direction entre l'axe 1 et l'axe 2. De plus, bien qu'elle soit très mal représentée, la variable de défauts primaires va dans le sens inverse de la note totale. Cela signifie que un café avec beaucoup de défauts primaires aura tendance à avoir une note plus basse.

En observant le nuage des individus et en séparant les cafés en arabica et robusta, on observe que les deux espèces de café ont des notes différentes. Globalement, les cafés robusta ont des notes inférieures à celles des cafés arabica.



Le nuages des individus met en valeur le fait que les individus que l'on peut observer au croisement des deux axes sont mal représentés selon le critère du cosinus carré. Plus on

s'éloigne du barycentre, plus les individus sont bien représentés. Après plusieurs essais, on voit que représenter les individus en fonction d'autres variables qualitatives ne permet pas de faire apparaître un lien entre les notes obtenues par les cafés et leurs attributs modélisés par les variables qualitatives. L'analyse des composantes factorielles ne permet pas vraiment de mettre en avant des liens avec les variables qualitatives.

Nos analyse nous permettent de mettre en évidence les individus contribuant le plus aux axes. Les individus comme le numéro 1309 ou le 1310 ont des notes basses. En fait, les cafés contribuant le plus à la formation des axes sont ceux qui ont des mauvaises notes dans les deux types de variables. Les individus ayant des bonnes notes dans l'un ou l'autre des types de variables contribuent peu à la formation des axes 1 et 2.

Notre ACP nous permet donc de faire apparaître deux types de variables. Mais elle ne permet pas vraiment de mettre en lien nos variables de scoring avec les attributs qualitatifs des cafés de notre base de données.

1.3 ANALYSE DES CORRESPONDANCES MULTIPLES

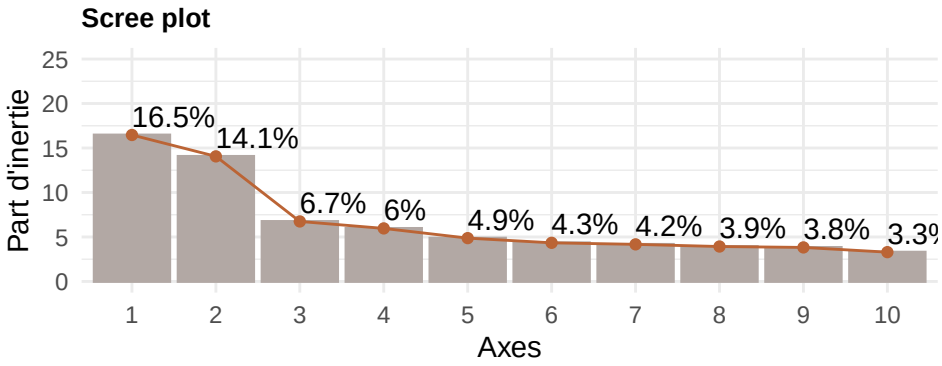
Nous allons à présent effectuer une analyse des correspondances multiples sur nos variables de scoring.

Ici, les variables sont devenues des variables qualitatives grâce à la transformation des notes en modalités. Les modifications sont les suivantes :

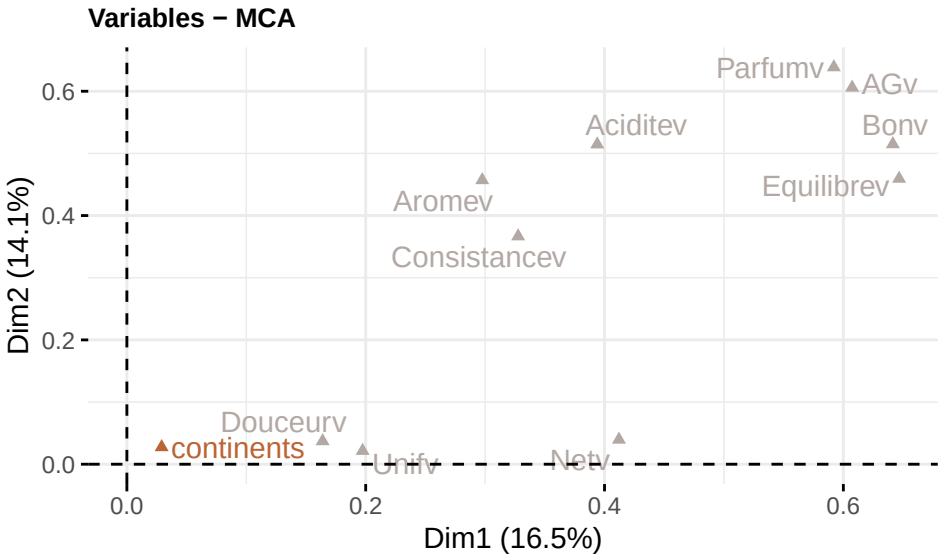
- les notes entre 0 et 5 portent la modalité "sans notation"
- les notes entre 5 et 7 portent la modalité "mauvais"
- les notes entre 7 et 8 portent la modalité "moyen"
- les notes entre 8 et 9 portent la modalité "bon"
- les notes entre 9 et 10 portent la modalité "excellent"

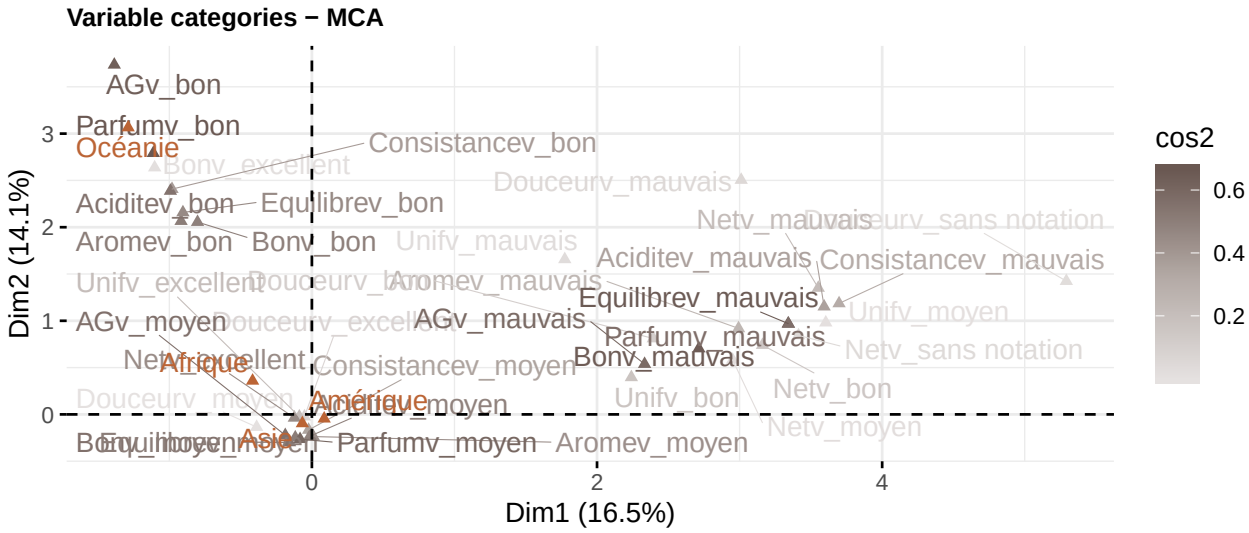
L'échelle de classification utilisée pour effectuer le changement de variable est basée sur les techniques de notation des professionnels. Il faut noter que la modalité "sans notation" ne doit pas être interprétée comme un "NA" mais comme une modalité signifiant que le café a une note tellement basse que cela ne vaut pas la peine de lui attribuer une note.

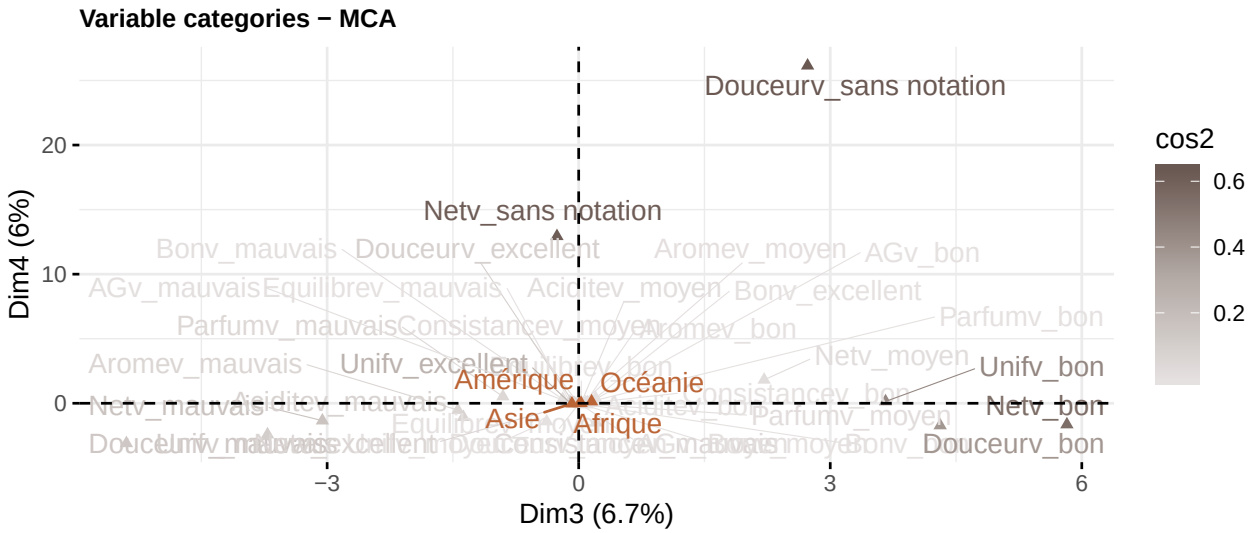
Garder l'axe 1 et l'axe 2 nous permet d'avoir une variance expliquée de 30.6%. Nous avons ajouté la variable continent car c'est la seule qui est informative dans notre ACM.



Y -t-il a un lien entre les groupes de variables observés avec un ACP et ceux observés avec une ACM? Le positionnement des variables sur les axes 1 et 2 fait apparaître deux groupes de variables : les variables de goût et les variables d'apparence , les mêmes que ceux trouvés précédemment grâce à notre ACP.







Dans le graphique représentant la dimension 1 et 2, il y a une disposition des modalités très caractéristique. Celles en haut de l'axe 1 sont les modalités avec l'attribut bon, celles près du barycentre sont celles avec l'attribut moyen et le groupe de modalités représenté dans le quart nord-est du graphique a l'attribut mauvais. Les attributs sans notation et excellent sont assez mal représentés sur ces deux axes.

Les attributs sans notation apparaissent sur les axes 3,4 et 5. Les attributs qui sont très bien représentés sont netteté sans notation et douceur sans notation. Ces modalités sont éloignées du barycentre. Cela est logique car la quasi totalité des individus ont l'attribut excellent pour les variables netteté et douceur. Ce sont donc des modalités rares.

La modalité mauvais contribue le plus à l'axe 1. Pour l'axe 2, c'est la modalité bon et pour l'axe 3, on a une opposition entre bon dans une direction et mauvais dans l'autre. Dans la dimension 4, c'est la modalité sans notation qui contribue le plus.

- Les individus 1062 et 1067 sont ceux qui contribuent le plus à la dimension 1 : ces deux individus ont des notes mauvaises dans la plupart des critères de notation.

- Les individus 5 et 10 sont ceux qui contribuent le plus à la dimension 2 : ces deux individus ont des notes bonnes et excellentes.
- Les individus 1067 et 983 sont ceux qui contribuent le plus à la dimension 3 : les individus ont des notes mauvaises et des notes bonnes et moyennes.
- L'individu 1068 est le seul qui contribue le plus significativement à l'axe 4 : l'individu a des notes mauvaises, moyennes ou même sans notation.

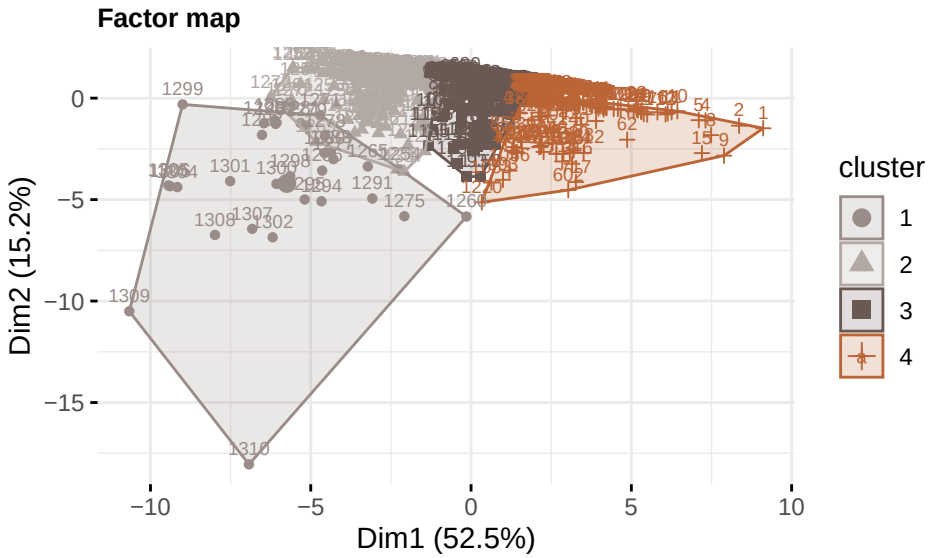
L'Analyse par composante multiple permet de voir que comme en ACP, des groupes de variables se forment. De plus, il y a certains individus comme l'individu 1067 qui n'ont pas des notes similaires aux autres individus. Cependant, les individus qui sont éloignés du barycentres ne sont pas les mêmes dans les deux types d'analyses factorielles.

1.4 CLASSIFICATION

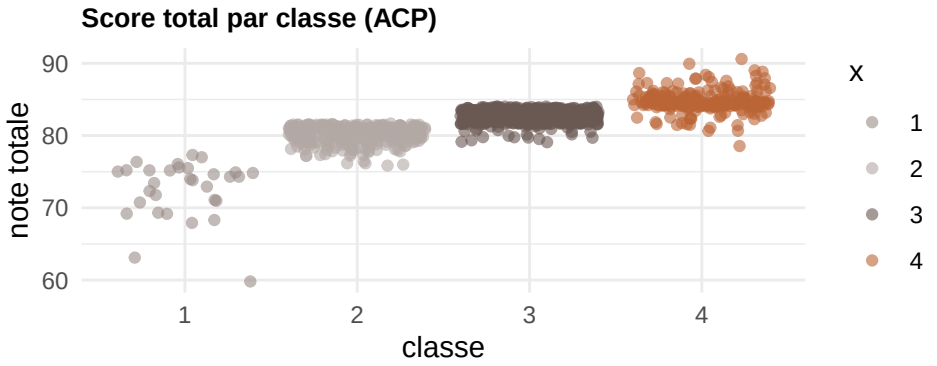
Les données étant observées, il faut classer les individus. On essaye de maximiser la variance intergroupe, et de minimiser la variance intragroupe. Ici, on essaye la méthode de classification hiérarchique ascendante avec consolidation.

1.4.1 CLASSIFICATION SUR L'ACP

Nous effectuons à présent une classification sur nos données, en se basant sur les résultats de l'ACP.



Les notes des 4 groupes sont représentées sur le graphique suivant. On voit bien qu'il y a une explication des classes par les notes totales. Le groupe 4 possède les meilleures notes. Les autres groupes ont des notes moyennes assez différentes. Le groupe 1 a les notes les plus basses.



Représenter les classes en fonction des autres variables qualitatives ne permet pas vraiment d'observer les caractéristiques des classes en fonction des variables quantitatives. Il y a certes des structures différentes pour chaque classe mais on ne peut pas vraiment émettre d'hypothèse sur les liens entre l'appartenance à une classe et les attributs quantitatifs de chaque individu.

L'individu représentatif de la classe 1 en ACP a un score total inférieur à 80/100 et des notes inférieures à 10/10 dans les variables douceur, uniformité, netteté. Dans les autres variables, il a des notes autour de 7/10 ou moins.

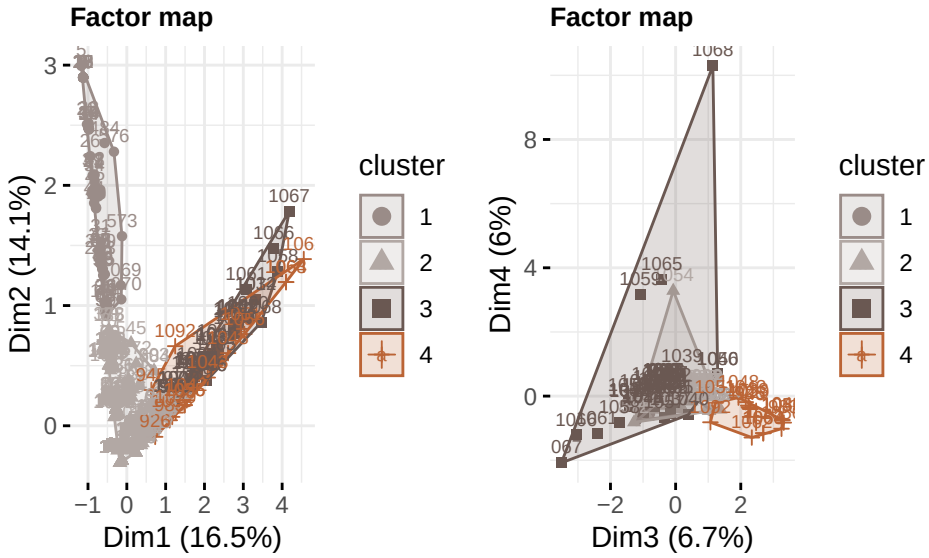
L'individu représentatif de la classe 2 en ACP a un score total inférieur à 82/100 et des notes de 10/10 dans les variables douceur, uniformité, netteté. Dans les autres variables, il a des notes autour de 7.1/10 ou plus.

L'individu représentatif de la classe 3 en ACP a un score total inférieur à 84/100 et des notes de 10/10 dans les variables douceur, uniformité, netteté. Dans les autres variables, il a des notes autour de 7.5/10 ou plus.

L'individu représentatif de la classe 4 en ACP a un score total de 85/100 ou plus et des notes de 10/10 dans les variables douceur, uniformité, netteté. Dans les autres variables, il a des notes autour de8/10 ou plus.

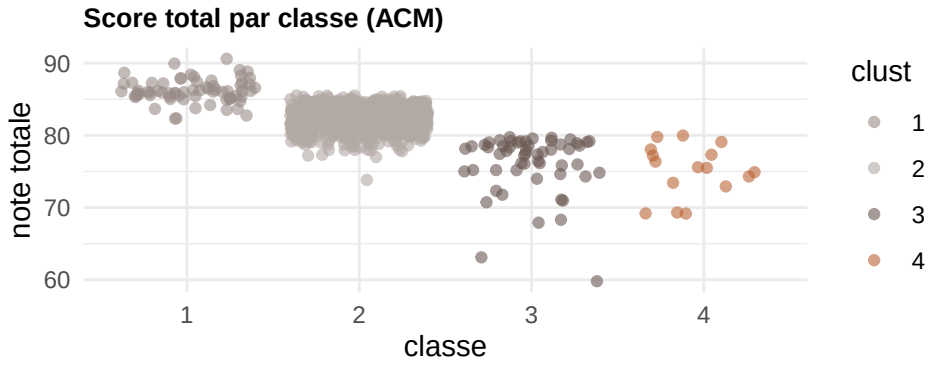
Notre classification sur les résultats de l'ACP est très pertinente sur les critères de notation et en particulier par rapport au score total (somme des notes à chaque critère /100).

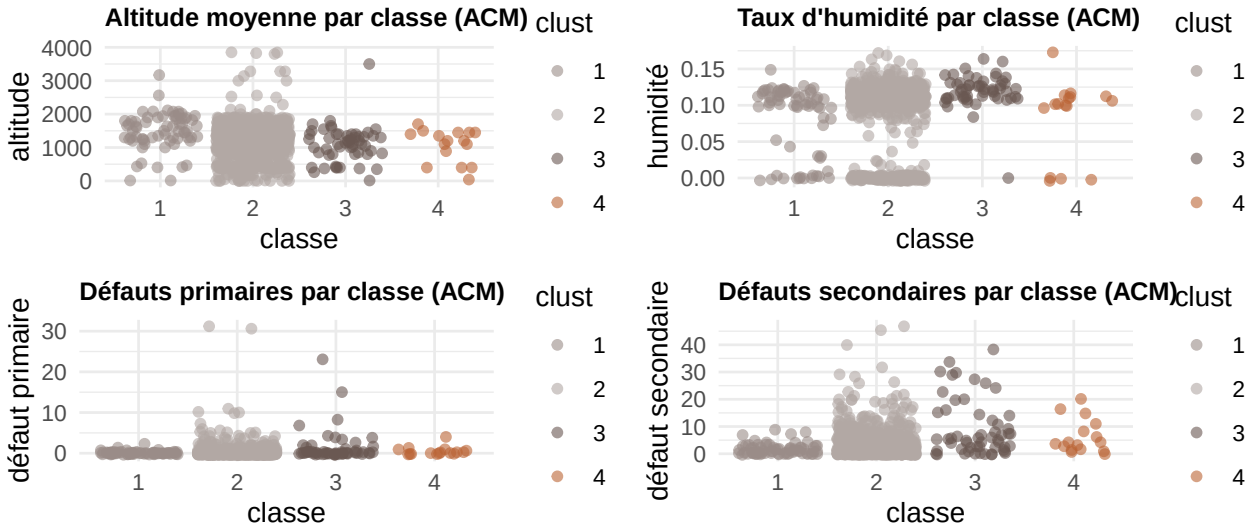
1.4.2 CLASSIFICATION SUR L'ACM



Comme on peut le voir, les groupes 3 et 4 se distinguent peu dans les dimensions 1 et 2 mais se distinguent dans les dimensions 3 et 4.

Les notes des 4 groupes sont représentées sur le graphique suivant. On voit bien qu'il y a une explication des classes par les notes totales. Le groupe 1 est celui avec les meilleures notes, le groupes 2 inclue la grande majorité des individus, les notes sont dans la moyenne. Les groupes 3 et 4 se ressemblent mais le groupe 3 est plus nombreux et plus dispersé. Cependant, par rapport aux classes découvertes grâce à l'ACP, les groupes sont moins différents en termes de score total.





Le graphique permet d'observer les caractéristiques des classes en fonction des variables quantitatives. Il y a une structure qui se distingue pour chaque classe, ce qui peut nous indiquer qu'il y a une lien entre le notes et les attributs (modélisés par les variables quantitatives) des cafés alors que cela n'apparaissait pas auparavant. Par exemple, les individus de la classe 1 avec des très bonnes notes ont toutes un nombre de défauts primaire inférieur à 5. Donc peut-être qu'avoir peu de défauts primaires influence la note du café.

En ce qui concerne le lien entre les classes et les variables qualitatives, il est également possible d'en faire. Les continents d'origine pour chaque classe sont différents : un café avec une bonne note peut provenir de tous les continents, par contre, les cafés appartenant à la classe 4 (qui contient les plus mauvaises notes) proviennent uniquement d'Amérique.

Les individus les plus représentatifs des classes découvertes sur les résultats de l'ACM sont les suivants :

L'individu représentatif de la classe 1 en ACM a des notes excellentes dans les critères de douceur, uniformité et netteté, des notes bonnes dans les autres critères. Il peut provenir de n'importe quel continent.

L'individu représentatif de la classe 2 en ACM a des notes excellentes dans les critères de douceur, uniformité et netteté, des notes moyennes dans les autres critères. Il peut provenir de n'importe quel continent sauf d'Océanie.

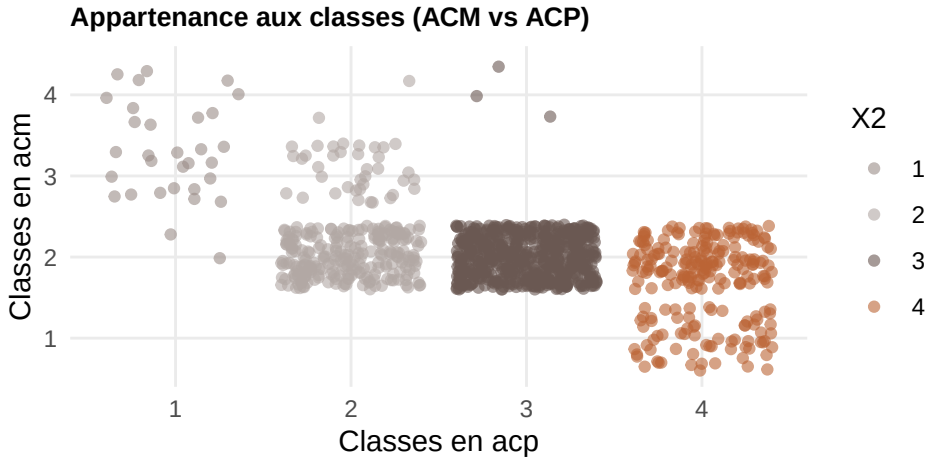
L'individu représentatif de la classe 3 en ACM a des notes excellentes dans les critères de douceur, uniformité et netteté, des notes mauvaises dans les autres critères. Il peut provenir d'Amérique ou d'Asie.

L'individu représentatif de la classe 4 en ACM a des notes bonnes dans les critères d'uniformité et netteté, excellente en douceur, des notes mauvaises ou moyennes dans les autres critères. Il peut provenir uniquement d'Amérique.

Comme on a pu le deviner avec nos analyses précédentes, les groupes 3 et 4 ne se distinguent pas dans toutes les variables. En effet, ils ont à peu près le même score moyen mais c'est pour les variables uniformité et netteté que les modalités sont différentes (excellent dans le

groupe 3 et bon dans le groupe 4).

1.4.3 COMPARAISON



X2obs	-21.92
df	1091.00
pvalue	0.00

Les classes sont corrélées, cela signifie que les deux partitions effectuées grâce à la classification sont toutes les deux pertinentes. Cependant, la classification effectuée sur les résultats de l'ACP met en avant des différences de note totale qui ne sont pas présentes dans la classification effectuée grâce à l'ACM. En effet, certains groupes de la classification en ACM ont à peu près le même score total moyen mais ont des modalités différentes dans certaines variables. Pour la classification en ACP, les moyennes de notes totales pour chaque classe sont différentes, de même, les notes sont différentes dans chaque critère de notation.

Au contraire, on a grâce à la classification sur les résultats de l'ACM des liens qui apparaissent avec les autres variables de la base de données par exemple avec la variable défaut primaire. A l'inverse, la classification effectuée grâce à l'ACP met moins en avant les liens avec les autres variables de la base de données mais s'appuie sur les différences de notes uniquement.

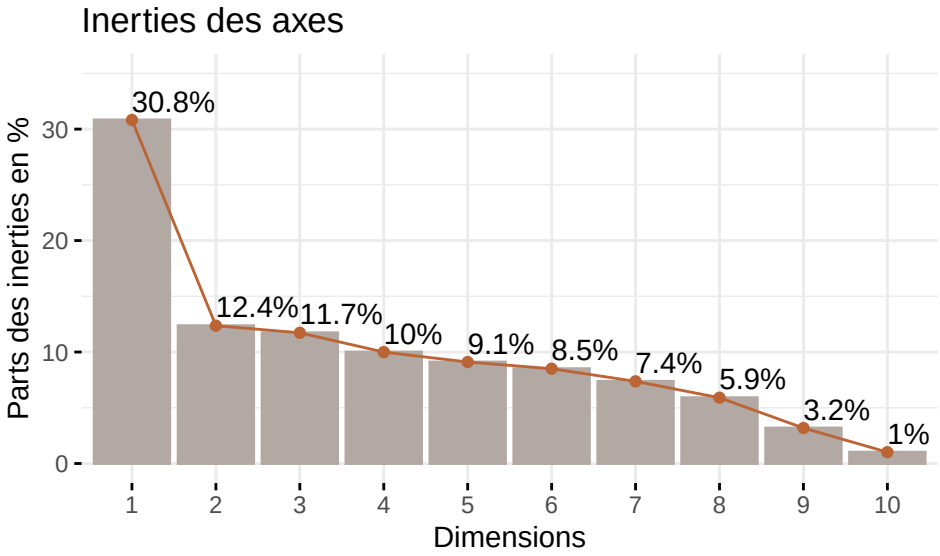
Nos deux analyses factorielles nous ont donc appris plusieurs choses :

- Nos variables de scoring peuvent se distinguer en deux groupes.
- Les individus possédant la modalité "sans notation" dans au moins un critère ne sont pas ceux qui ont les plus mauvaises notes.
- L'ACP nous offre une classification très pertinente en termes de note alors que la classification effectuée grâce à l'ACM est plus pertinente concernant d'autres variables que les notes.
- Notre base de données permet de mettre en valeur 4 groupes d'individus relativement homogènes.
- Le continent d'origine a une influence sur la note du café.
- Le nombre de défauts (primaires ou secondaires) influence négativement la note du café.

2 ANALYSE DES VARIABLES QUALITATIVES

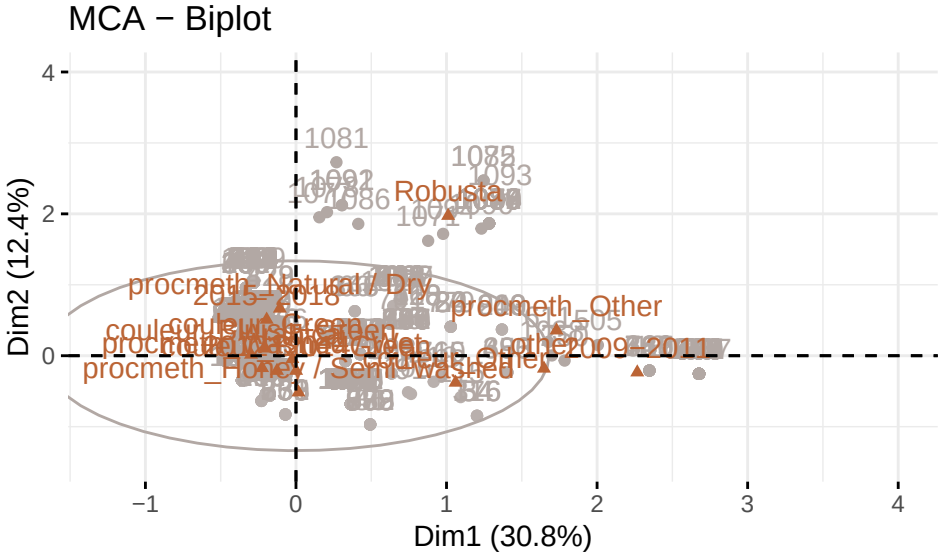
2.1 ANALYSE DES CORRESPONDANCES MULTIPLES - ACM

2.1.1 VALEURS PROPRES - INERTIES DES AXES

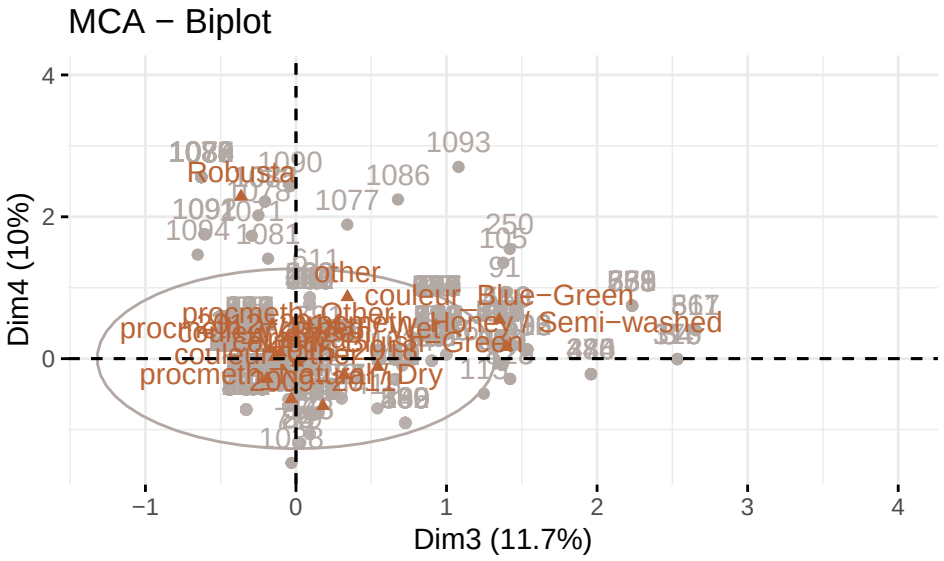


Nos deux premiers axes regroupent 43% de l'inertie totale des axes, ce pourcentage est assez élevé pour une analyse effectuée sur un ensemble de variables. Notre étude sera donc effectué plus particulièrement sur les dimensions 1 et 2 mais il serait tout de même intéressant d'observer la dispersion des individus et variables selon les axes 3 et 4.

2.1.1.1 Nuages de points Dim 1 - Dim 2



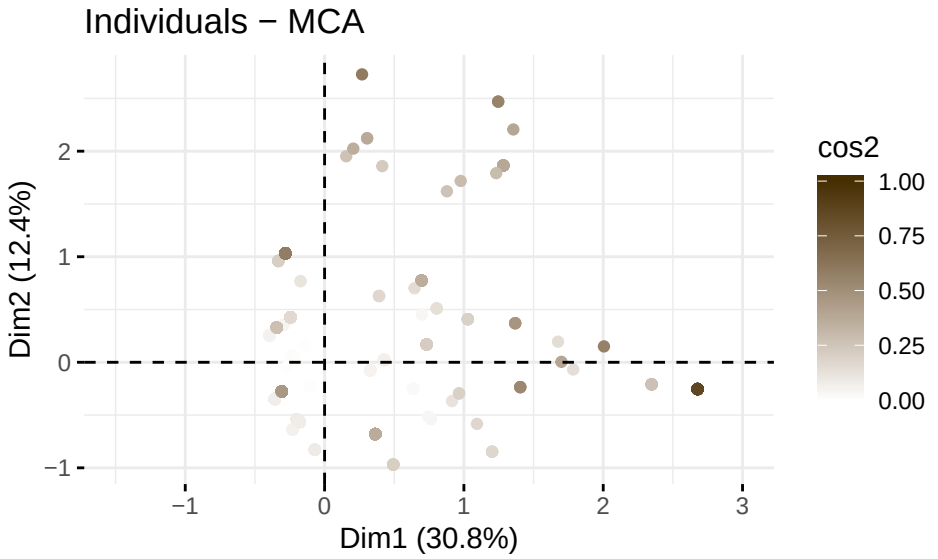
2.1.1.2 Nuages de points Dim 3 - Dim 4



Nous remarquons deux groupes bien distinct dans notre nuage de points qui correpondent aux deux espèces de graines de café : Arabica et Robusta.

2.1.2 QUALITÉ DE REPRÉSENTATION

2.1.2.1 Selon les individus



- Top 5 des individus les mieux représentés selon l'axe F1

	species	country	variete	procmeth	couleur	Mois	years1
5	Arabica	Ethiopia	NA	Other	Other	Septembre	2009-2011
30	Arabica	Brazil	Yellow Bourbon	Other	Other	Janvier	2009-2011
31	Arabica	Brazil	Yellow Bourbon	Other	Other	Janvier	2009-2011
55	Arabica	El Salvador	NA	Other	Other	Mai	2009-2011
75	Arabica	Mexico	Pacamara	Other	Other	Avril	2009-2011

Selon l'axe 1, l'individu le mieux représenté correspond à l'individu 5 faisant partie de la catégorie Arabica provenant d'Éthiopie produit entre 2009 et 2011.

- Top 5 des individus les mieux représentés selon l'axe F2

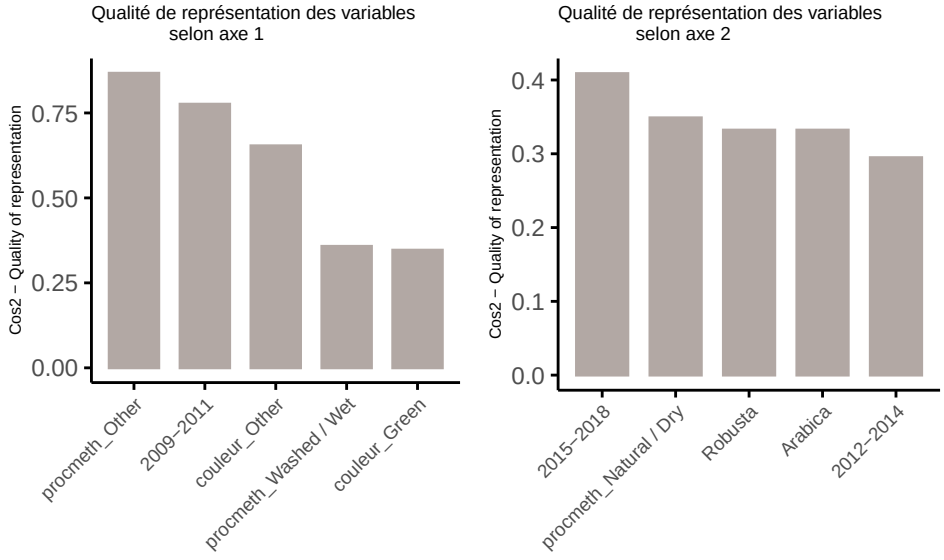
	species	country	variete	procmeth	couleur	Mois	years1
1081	Robusta	India	Other	Natural / Dry	Green	Aout	2015-2018
49	Arabica	Ethiopia	Other	Natural / Dry	Green	Mars	2015-2018
56	Arabica	China	Catimor	Natural / Dry	Green	Avril	2015-2018
84	Arabica	Taiwan	Typica	Natural / Dry	Green	Mai	2015-2018
135	Arabica	Colombia	Caturra	Natural / Dry	Green	Avril	2015-2018

Selon l'axe 2, l'individu 1081 est le mieux représenté. Cet individu correspond à des graines de café de type Robusta provenant d'Inde, récolté entre 2015 et 2018.

Ces deux individus nous informent sur les éléments formants nos axes F1 et F2. L'axe F1 représente la répartition des individus selon leur processus de lavage des différentes graines de café et pour ce qui est l'axe F2, elle représente les deux groupes d'espèces de graines de café.

2.1.2.2 Selon les variables

Les variables ayant une meilleure représentation selon l'axe 1 et l'axe 2 sont



2.1.3 CONTRIBUTION

2.1.3.1 Selon les individus

- Individus contribuant à l'axe F1

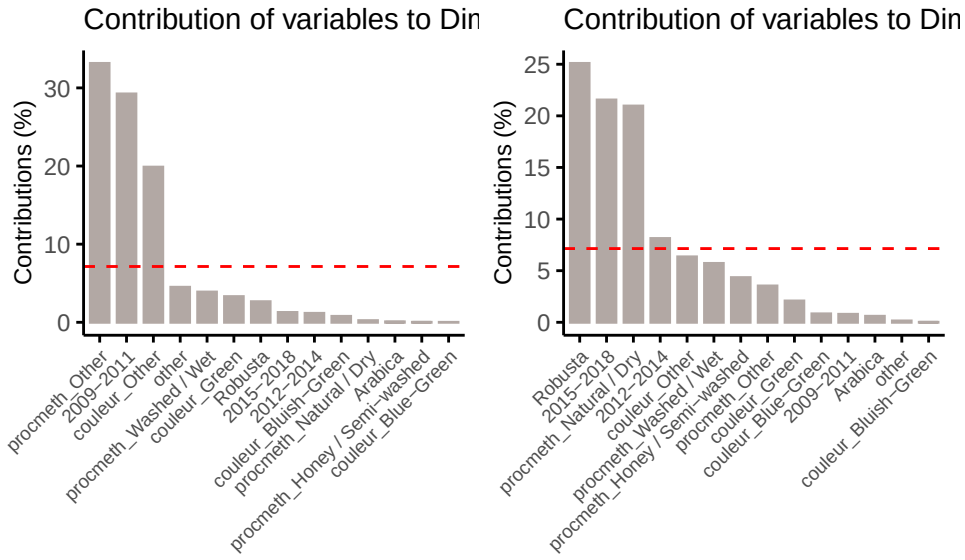
	species	country	variete	procmeth	couleur	Mois	years1
5	Arabica	Ethiopia	NA	Other	Other	Septembre	2009-2011
30	Arabica	Brazil	Yellow Bourbon	Other	Other	Janvier	2009-2011
31	Arabica	Brazil	Yellow Bourbon	Other	Other	Janvier	2009-2011

- Individus contribuant à l'axe F2

	species	country	variete	procmeth	couleur	Mois	years1
1081	Robusta	India	Other	Natural / Dry	Green	Aout	2015-2018
1072	Robusta	India	NA	Other	Green	Avril	2015-2018
1085	Robusta	India	NA	Other	Green	Avril	2015-2018

Nous remarquons une forte similitude des résultats entre les individus ayant une meilleure représentation et une forte contribution aux axes.

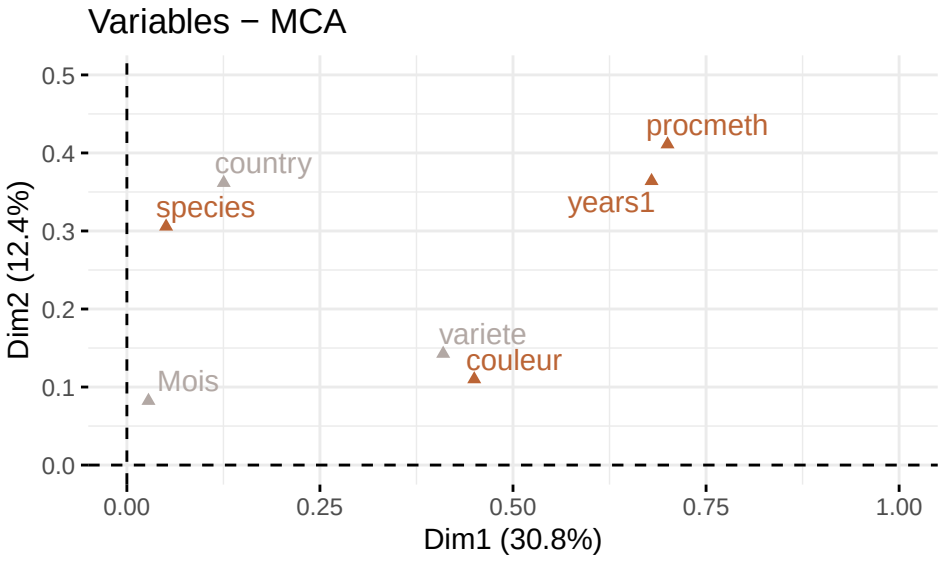
2.1.3.2 Selon les variables



De même pour les variables, "procmeth_other" et "Robusta" sont les deux variables contribuant le plus aux axes 1 et 2 respectivement.

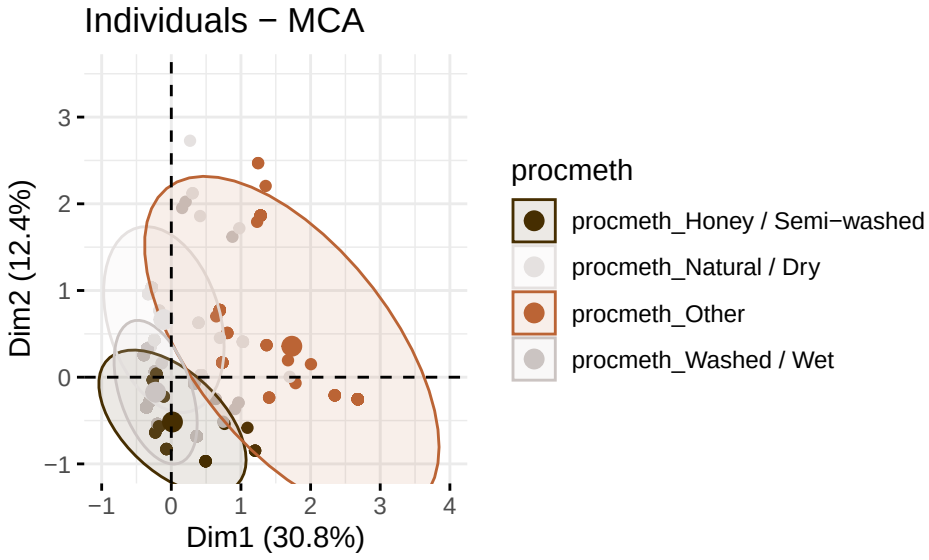
2.1.4 VISUALISATION DES CORRÉLATIONS

Dans le graphique suivant, nous avons représenté les corrélations des différentes variables selon les différents axes.



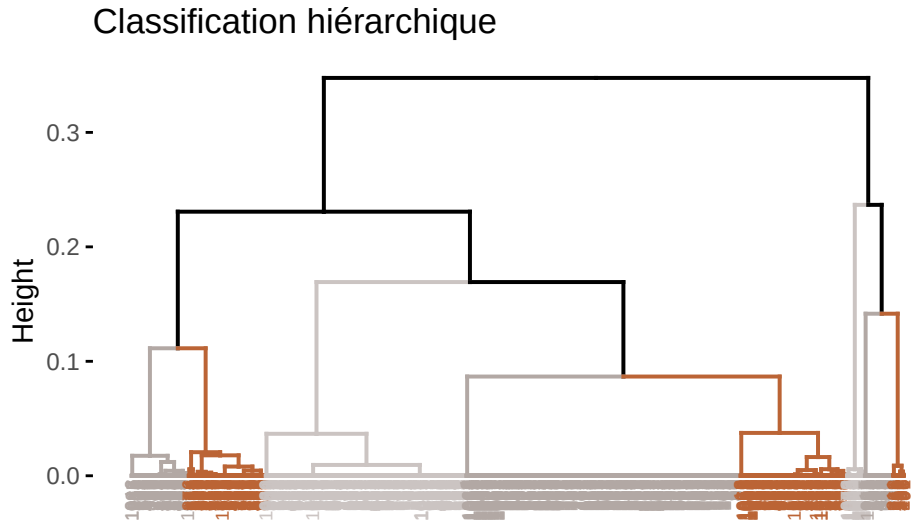
Parmis les variables actives représentées par la couleur oranges, procmeth et years ont une forte corrélation avec les 4 axes représentés. La variable species a une corrélation d'environ de 0.3 avec les axes F2 et F4. Et la variable couleur a une corrélation plus élevé pour les axes F1 et F3.

Selon la corrélation des variables avec les axes, nous pouvons observer que la variable procmeth est celle qui a une part d'inertie expliquée la plus élevée dans notre ACM. Il serait donc intéressant de visualiser le nuage de point selon un habillage représentant les différentes catégories de la variable de méthode de lavage des graines.



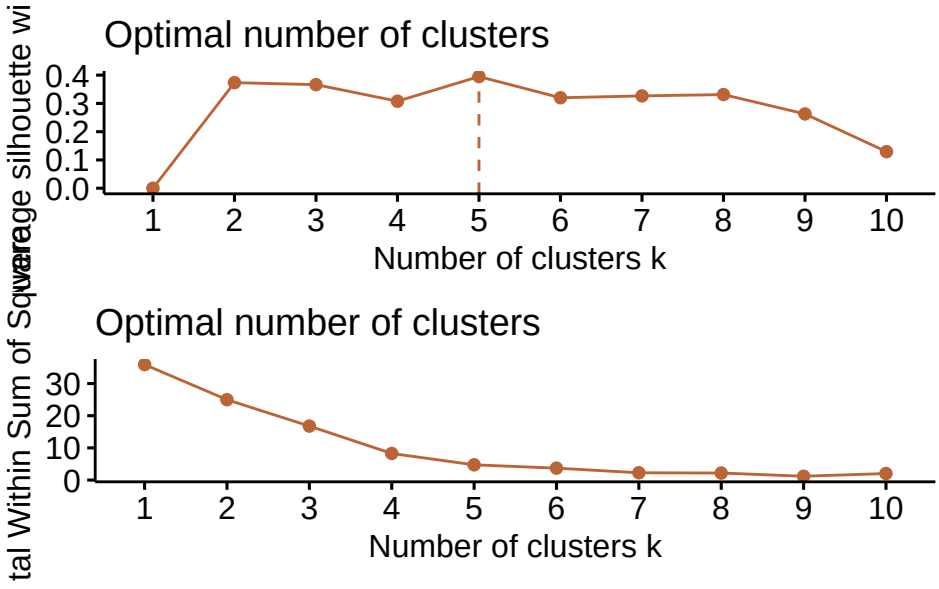
Nous observons donc quatre groupes grâce à ce graphique. Il faut maintenant certifier ces regroupements avec la méthode des K-means.

2.1.5 CLASSIFICATION

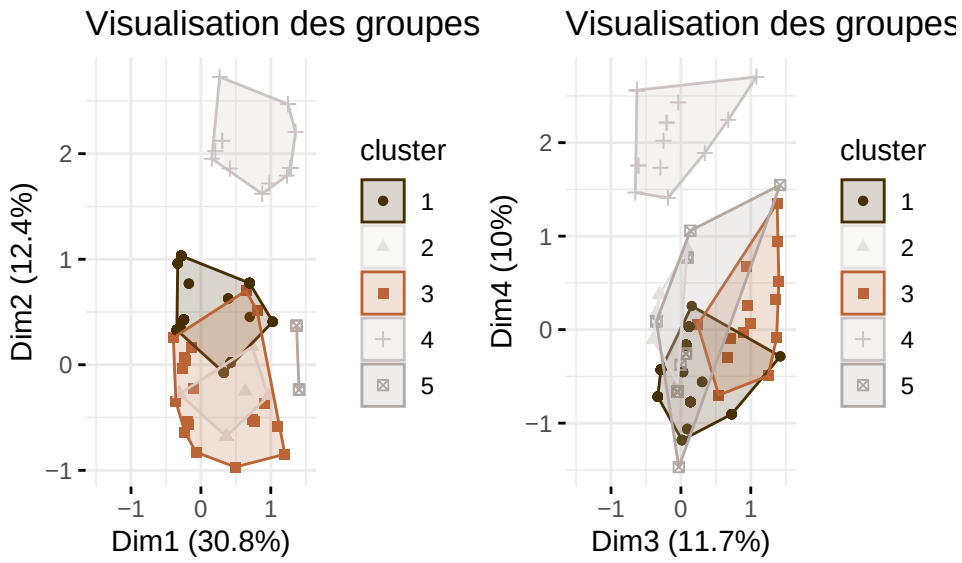


Le dendogramme nous permet de distinguer 4 ou 5 groupes. Mais cette méthode de classification hiérarchique ascendante ne se focalise pas seulement sur ce dendogramme. Il est donc nécessaire d'observer d'autres classifications comme les k-means. Cette méthode sera donc étudiée par la suite.

Pour la classification de données par la méthode des k-means, il nous faut choisir en combien de groupe nous voulons diviser nos observations. Pour cela il est intéressant d'analyser les graphiques du nombre optimal de clusters.



D'après le graphique précédent, pour la suite de notre étude il serait plutôt utile de visualiser le découpage en 5 groupes.



Nous remarquons bien les 5 groupes dans le premier graphique qui se base dans un premier temps sur les modalités des espèces et parmis les espèces Arabica il y a apparition de 4 groupes.

Le groupe 1 décrit les individus de catégorie Arabica (espèces) qui ont eu un processus de nettoyage dit naturel/sec où la graine est de couleur verte.

Le groupe 2 décrit les individus de catégorie Arabica qui ont eu un processus de nettoyage dit lavé ou humide où la graine est de couleur verte aussi.

Le groupe 3 décrit lui aussi les graines de la catégorie Arabica mais qui ont subit un processus de nettoyage semi-lavé où la graine est de couleur verte aussi.

Le groupe 4 décrit les graines de la catégorie Robusta qui eux n'ont pas de processus de nettoyage spécifique.

Le groupe 5 décrit les graines de type Arabica qui n'ont pas eu d'attribution de processus de nettoyage spécifique et leurs récoltes ont été particulièrement entre 2009 et 2011.

2.1.6 CONCLUSION ACM (VARIABLES QUALITATIVES)

Suite à notre étude, notre base de données peut être divisée en 5 groupes. Lorsque l'on considère que l'on peut avoir une information complète, il serait idéal de faire 4 groupes en enlevant la formation d'une classe par les "other". Les différentes classes seront donc dépendant du type d'espèces puis ,s'il est du type Arabica, elles devraient dépendrent des modalités du processus de nettoyage des graines de café.