STATISTIQUE
SCIENCE DES DONNÉES BIOSTATS
UNIVERSITÉ DE MONTPELLIER

TP1 - Analyse et modélisation multivariée

Master 2 biostatistiques

Régression pénalisée

Présenté par :
Sarah Matoub
Pauline Dusfour–Castan

05-10-2023

Table des matières

1 Introduction

Nous disposons d'un jeu de données, disponible dans le fichier "genus.csv", porté sur l'abondance de 27 espèces d'arbres dans le bassin du Congo.

Ce dataframe contient :

1. Les différents niveaux d'abondance des espèces (gen1 à gen27)
2. Des variables géographiques quantitatives (latitude, longitude, altitude, pluviométries annuelles et mensuelles)
3. Des variables géographiques qualitatives (type de sol)
4. Des variables relatives à la photosynthèse (indices EVI)

Toutes ces informations sont collectées à partir de parcelles forestières dont la surface figure dans la dernière colonne.

On note que la variable "forest" (type de forêt) ne sera pas utilisée dans cette analyse, car elle ne contribue pas à notre objectif spécifique de modélisation de la densité globale de peuplement arboré.

L'objectif de ce TP est de modéliser au mieux la densité globale de peuplement en utilisant les autres variables disponibles. Pour ce faire, nous allons explorer plusieurs méthodes successives et évaluer leurs performances en les comparant.

2 Calculs de variables

2.1 Calcul de la variable dépendante Y

Dans un premier temps, nous allons calculer la variable dépendante Y, qui représente la densité de peuplement arboré sur la parcelle, pour ce faire, nous utilisons la fonction `mutate` de la librairie **dplyr** [HRL23] pour créer une nouvelle colonne nommée "density". Cette colonne contient la somme des abondances de 27 espèces d'arbres, obtenue en utilisant la fonction `rowSums` sur un sous-ensemble de données sélectionné avec la fonction `select(., starts_with("gen"))`. Enfin, cette somme est divisée par la surface de chaque parcelle, ce qui donne la densité de peuplement arboré par unité de surface.

Vous trouverez un extrait des résultats dans le tableau 9 à l'annexe 8.1.

2.2 Calcul du tableau des variables explicatives

Dans un second temps, nous procéderons au calcul du tableau des variables explicatives, noté X. Celui-ci est composé des variables géographiques quantitatives, de leurs interactions avec la variable géologie, de la variable géologie elle-même, ainsi que des indices EVI.

Pour ce faire, nous allons commencer par calculer le produit d'interaction entre les variables géographiques quantitatives et les indicatrices de la variable geology. L'objectif est de créer de nouvelles variables qui capturent les interactions potentielles entre ces deux types de variables.
Voici comment nous procédons :

— Nous allons recoder la variable geology en indicatrices à l'aide de la fonction `acm.disjonctif()` de la librairie **ade4** [DDT].

— Une fois que nous avons les indicatrices de geology, nous allons effectuer le calcul d'interaction. On note $G = ((G_{ij}))_{i,j}$ avec $i = 1 \dots 1000,\ j = 1 \dots 16$ la matrice contenant les variables géographiques quantitatives et m la matrice de taille 1000×5 contenant les indicatrices de la variable geology, le produit d'interactions est donné alors par la formule :

$$G * m = P_{i,j} = G_i * m_j$$

où G_i représente la ligne i de la matrice G, et m_j représente la colonne j de la matrice m.
En d'autres termes, chaque élément de la matrice résultante P est obtenu en multipliant la ligne correspondante de la matrice G par la colonne correspondante de la matrice m.
Une fois que nous avons obtenu notre tableau X de variables explicatives, nous allons le centrer et réduire en utilisant la fonction `scale` qui utilise l'estimateur sans biais de la variance, on corrigera cela en multipliant par $\sqrt{\frac{1000}{999}}$.
Vous pouvez retrouver un extrait des résultats de ce calcul dans le tableau 10 de l'annexe 8.1.

Faisons à présent l'inventaire des multicolinéarités présentes dans l'ensemble des variables explicatives :

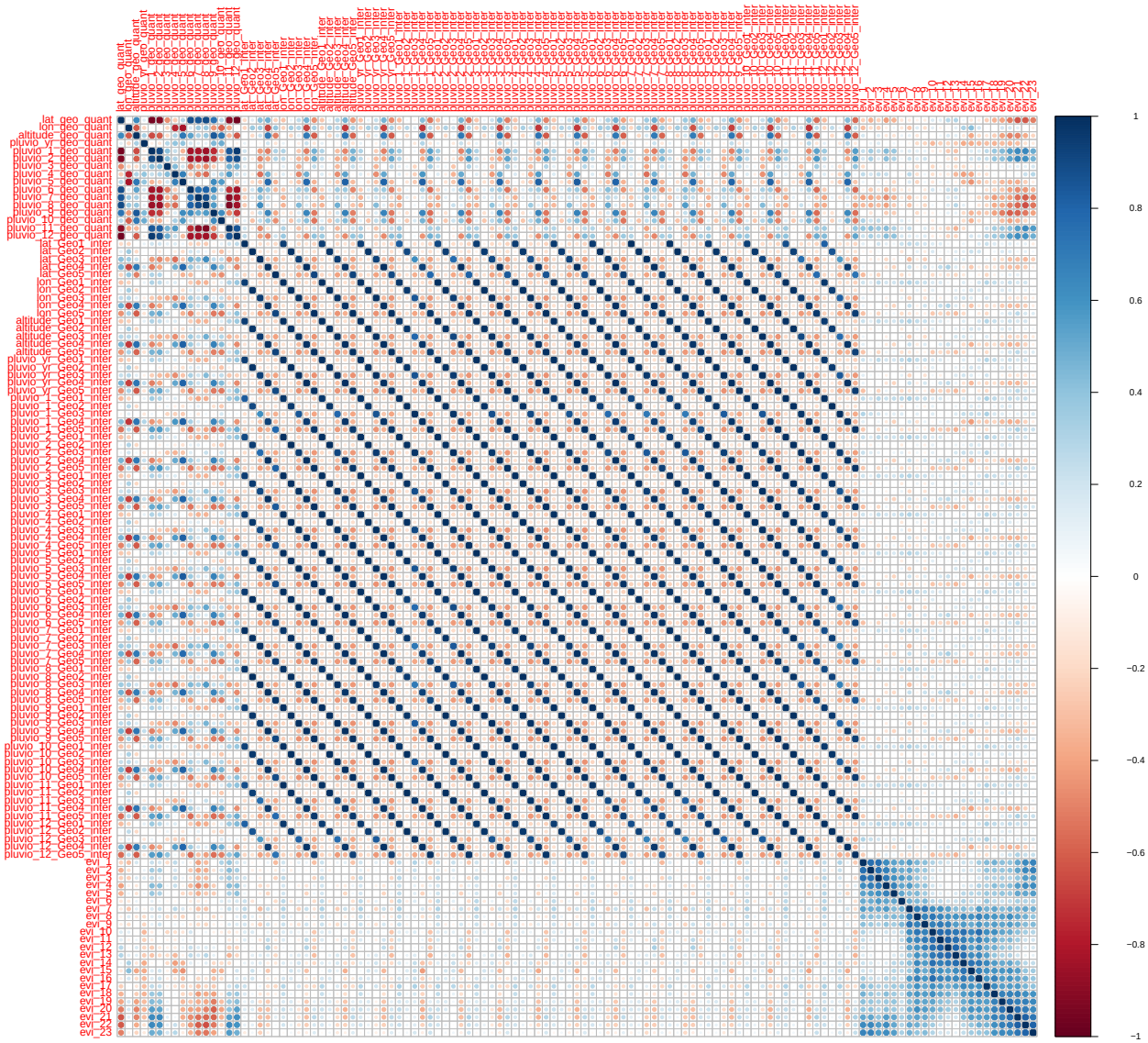


FIGURE 1 – Matrice des corrélations entres les variables de X

La figure 1 présente la matrice de corrélation des variables de X. Les variables avec le suffixe "quant" correspondent aux variables géographiques quantitatives, tandis que celles avec le suffixe "inter" représentent les produits d'interaction entre les variables géographiques et les indicatrices de la variable "geology".

Par exemple, `lat_Geo1_inter` désigne le produit d'interaction entre la latitude et l'indicatrice de "Geology1".

En regardant de plus près la figure 1, on voit que dans le coin inférieur droit du graphique, les indices EVI sont corrélés positivement entre eux. Cependant, si l'on examine la partie supérieure du graphique, on constate qu'ils présentent très peu de corrélations avec les autres variables, notamment avec les variables quantitatives et les produits d'interaction, voire même négativement corrélés comme c'est le cas, par exemple de la variable de latitude et `evi_21`.

Lorsqu'on regarde la section centrale du graphique, il est évident qu'il existe des multicolinéarités parfaites entre plusieurs variables d'interaction. Notamment, une parfaite colinéarité est observée entre la variable `lat_geo1_inter` et un groupe de variables comprenant `"lon_geo5_inter", "altitude_geo5_inter," "pluvio_4_geo1_inter," "pluvio_5_geo1_inter," "pluvio_7_geo1_inter," "pluvio_6_geo1_inter"` et ainsi de suite.

3 Première régression sur composantes principales

3.1 ACP globale des variables explicatives

3.1.1 Distribution de l'inertie

Dans cette section, nous allons réaliser une ACP globale du tableau des variables explicatives X qui contient 1000 individus et 124 variables.
Pour ce faire, nous allons calculer les valeurs propres de la matrice d'inertie afin de savoir le nombre d'axes que nous allons retenir pour faire notre ACP.
L'histogramme ci-dessous donne le pourcentage d'inertie apporté par chaque dimension :

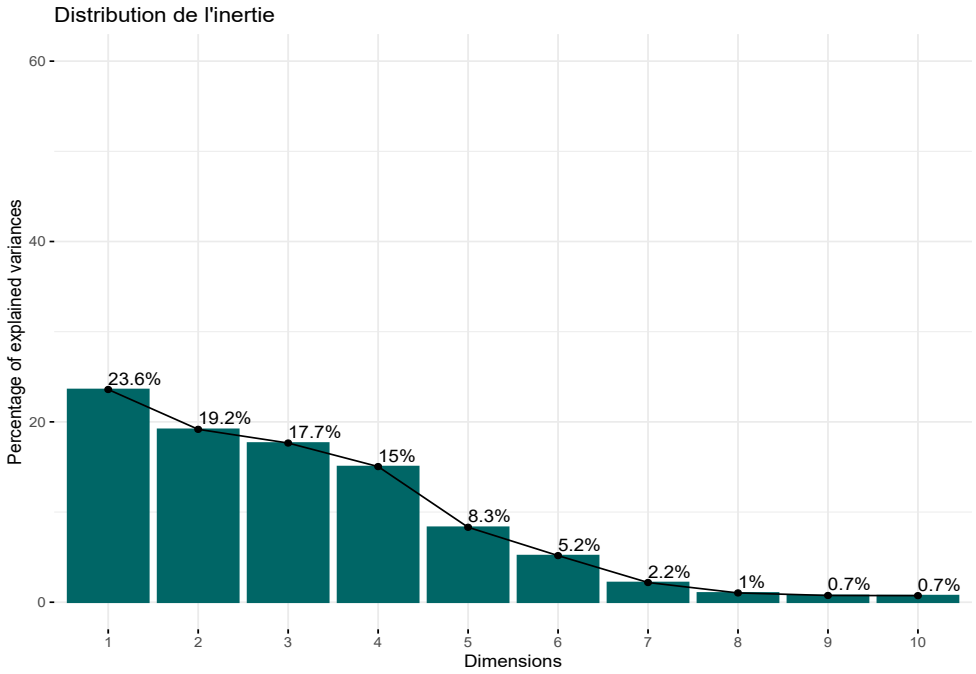


FIGURE 2 – Pourcentage d'inertie apporté par chaque dimension

A la lecture de la figure 2, on constate que les écarts entre les premières valeurs propres sont faibles, que ce soit entre la première et la deuxième valeur propre, entre la deuxième et la troisième, ou entre la troisième et la quatrième.
De plus, en utilisant la commande `res.pca$eig`, nous allons afficher les valeurs propres :

	Valeurs propres	Pourcentage d'inertie	Pourcentage d'inertie cumulé
comp 1	29.25	23.59	23.59
comp 2	23.77	19.17	42.76
comp 3	21.89	17.65	60.41
comp 4	18.65	15.04	75.46
comp 5	10.31	8.31	83.77
comp 6	6.41	5.17	88.94
comp 7	2.71	2.19	91.12
comp 8	1.28	1.03	92.15
comp 9	0.93	0.75	92.9
comp 10	0.91	0.73	93.63

TABLE 1 – Description des dimensions

Les deux premières composantes représentent 42,76% de l'inertie totale de notre jeu de données, cela signifie que 42,76% de la variabilité totale du nuage est représentée dans le plan (1,2) ,ce qui est un pourcentage relativement moyen, donc le premier plan ne représente qu'une partie de la variation contenue dans l'ensemble du jeu de données actif, nous allons alors nous intéresser aux dimensions supérieures.

On remarque que 92.15% de l'inertie est capturé par les 8 premières composantes et 7.85% de l'inertie y est perdue, les composantes 9 à 124 possédant des valeurs propres strictement inférieures à 1 ne seront pas conservées dans la suite de nos analyses car elles apportent moins d'information qu'une seule variable toute seule et peuvent être considérées comme du bruit.

3.1.2 Description du plan (1,2)

Dans cette section nous allons nous intéresser à la structure de corrélations entre les variables , pour cela nous avons fait le choix d'afficher le cosinus carré de chaque variable afin de voir leur qualité de représentation sur le plan(1,2).
En effet, un cosinus carré élevé traduit une bonne représentation de la variable sur les composantes tandis qu'un cosinus carré faible indique que la variable n'est pas très bien représentée.

Premièrement , nous allons visualiser le cos2 de chaque variable pour chaque dimension, pour cela nous allons afficher la qualité de représentation de chaque variable sur le premier plan, il est important de noter que nous avons limité l'affichage à 60 variables, car un nombre plus élevé rendrait le graphique illisible.

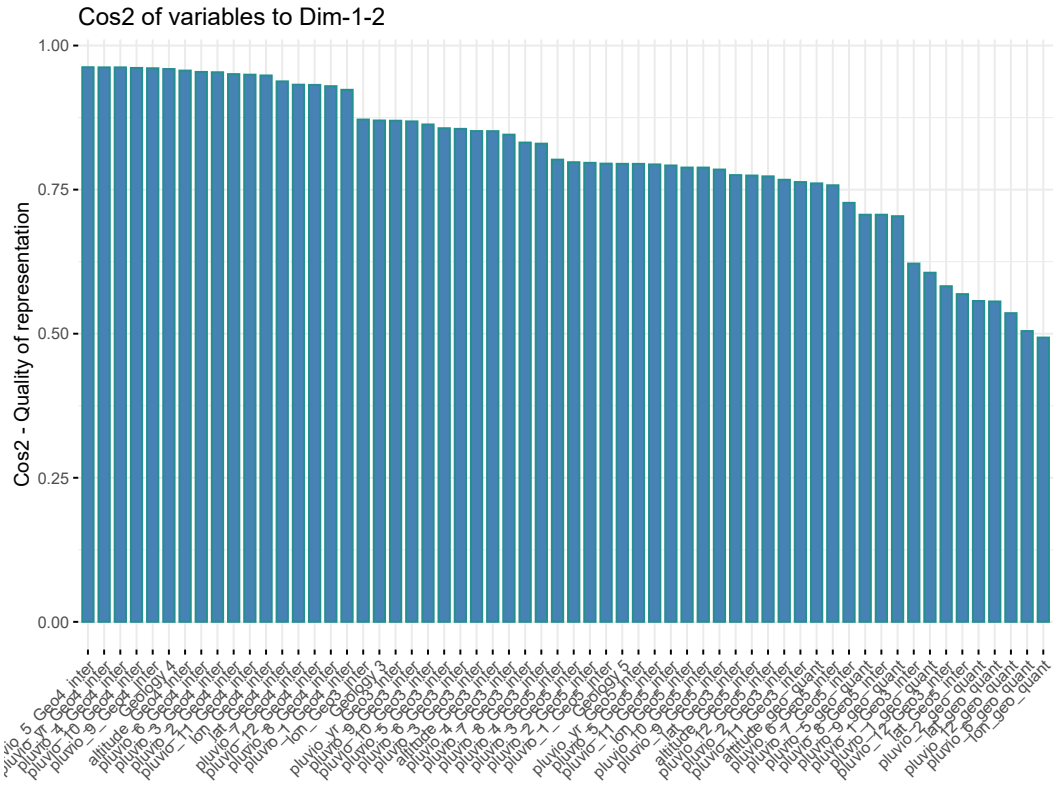


FIGURE 3 – Cosinus carré de chaque variable sur le plan (1,2)

À l'analyse de la figure 3, il est clair que les variables d'interaction, marquées par le suffixe "inter," et qui représentent des combinaisons entre les variables géographiques quantitatives et les indicateurs de la géologie, affichent des cosinus carrés remarquablement élevés, s'approchant de la valeur maximale de 1. Cela indique une corrélation significative et une excellente représentation de ces variables sur le plan (1,2). En outre, les variables géographiques quantitatives, reconnaissables par le suffixe "quant," montrent également des cosinus carrés élevés, reflétant une corrélation substantielle avec le plan (1,2).

Plus spécifiquement, on peut observer que le cos2 de la variable `Pluvio_5_Geo4_inter` atteint 0,96 sur le plan (1,2), ce qui signifie que 96 % de l'originalité de cette dernière est représentée sur ce plan, illustrant une forte corrélation.

Ces constatations mettent en évidence l'importance de ces variables d'interaction dans la structure des données.

Nous obtenons la projection sur le premier plan dual, représenté par la figure 4 :

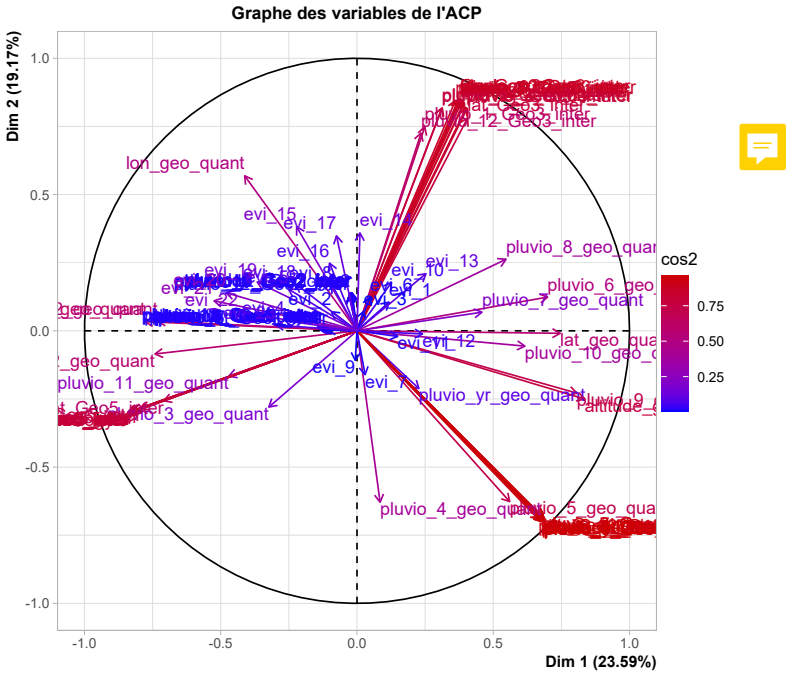


FIGURE 4 – Photographie du plan (1,2)

En examinant cette projection, nous pouvons observer plusieurs tendances importantes. Tout d'abord, les variables d'interaction, identifiées par le suffixe 'inter', sont celles dont le cosinus carré est le plus élevé, ce qui confirme ce qui a déjà été mis en évidence dans la figure 3 précédente, soulignant la forte corrélation entre ces variables. Elles se regroupent étroitement sur le plan de projection, formant une sorte de "vaisseau" en raison de leur forte corrélation mutuelle.

De plus, ces variables d'interaction sont également fortement corrélées avec les indicatrices de la variable geology, ce qui est tout à fait cohérent puisque ces dernières ont été construites en combinant les variables géographiques quantitatives avec les indicatrices de géologie.

À l'inverse, les indices EVI ont des cos2 très faibles, ce qui indique qu'elles sont mal représentées sur ces composantes principales.

Pour conclure, cette ACP met en évidence particulièrement en évidence l'importance des variables d'interaction dans l'explication de la variance des données.

3.2 Modélisation de la densité

Nous allons à présent modéliser la densité Y en utilisant les 8 composantes principales retenues de X. Pour ce faire, nous avons réalisé une ACP sur les données de X, extrayant ainsi les 8 premières composantes principales. Ensuite, nous avons employé ces composantes principales (stockées dans la matrice nommée `X_reg`) pour construire un modèle de régression linéaire multiple avec la variable de réponse Y, dont les résultats sont affichés dans la table 2 :

	Coefficients estimés	**Pr(>|t|)**
`Intercept`	17.62022	$< 2 \times 10^{-16}$
`X_regDim.1`	0.28074	1.06×10^{-10}
`X_regDim.2`	0.30555	2.31×10^{-10}
`X_regDim.3`	-0.19368	0.000104
`X_regDim.4`	0.14589	0.006862
`X_regDim.5`	-0.25682	0.000411
`X_regDim.6`	1.11101	$< 2 \times 10^{-16}$
`X_regDim.7`	-0.13299	0.346770
`X_regDim.8`	0.28420	0.167609

TABLE 2 – Résumé de la régression linéaire multiple entre Y et les 8 CP de X

On obtient un $R^2 = 0.2127$ ce qui signifie que 21.27 % de la variation dans la densité peut être expliquée par l'ensemble de ces composantes principales, ce qui est un pourcentage relativement faible.

Par ailleurs, en examinant la table 2, nous pouvons analyser les coefficients estimés pour chacune des composantes principales, accompagnés des p-values associées au test de Student, répertoriées dans la colonne **'Pr(>|t|)'**.

Ces p-values évaluent la signification statistique de chaque coefficient estimé. Lorsque la p-value est très faible, comme par exemple $< 2 \times 10^{-16}$, cela indique que le coefficient est statistiquement significatif. En revanche, si la p-value est élevée, comme c'est le cas pour les composantes `'X_regDim.7'` et `'X_regDim.8'`, cela veut dire que ces coefficients ne contribuent pas de manière statistiquement significative à l'explication de la variation de la variable dépendante.

Il est donc justifié d'envisager l'élimination des deux dernières composantes, car leur contribution à l'analyse est statistiquement non significative.

Vérifions la pertinence de l'application du test de Student dans notre analyse, on rappelle que ce test est approprié sous les hypothèses suivantes :

- **Linéarité** : tout d'abord, il est essentiel de vérifier que la relation entre la variable dépendante Y et les composantes principales X_{reg} est bien linéaire. Pour vérifier cette hypothèse nous allons représenter les résidus studentisés en fonction des valeurs prédites par le modèle $\hat{Y}$.

- **Normalité des résidus** : ce qui signifie que les résidus (les écarts entre les valeurs observées et les valeurs prédites par le modèle) suivent une distribution normale. Pour évaluer cette hypothèse, nous allons utiliser le QQ-plot entre quantiles normaux théoriques et résidus standardisés.

- **Homoscédasticité** : il est important que la variance des résidus reste relativement constante à tous les niveaux des valeurs prédites. En d'autres termes, l'homoscédasticité signifie que la dispersion des résidus ne varie pas en fonction du niveau de la variable dépendante. Nous allons vérifier cette condition à l'aide du graphique représentant le carré des résidus studentisés en fonction de $\hat{Y}$.

Premièrement, vérifions l'hypothèse de linéarité en traçant les résidus studentisés en fonction des valeurs prédites par le modèle $\hat{Y}$, comme sur le graphique 5 :

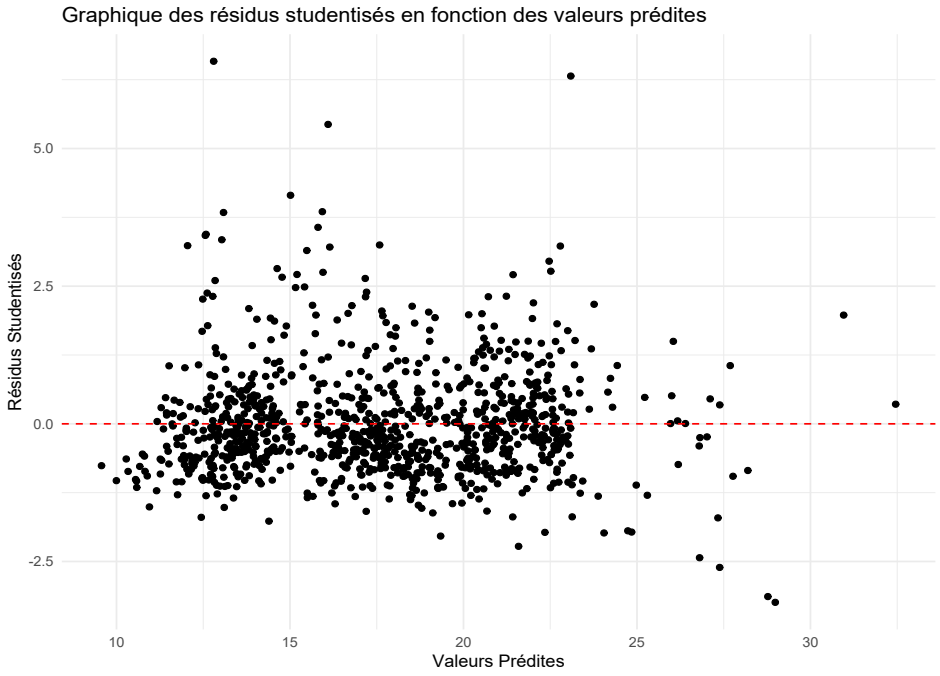


FIGURE 5 – Graphique des résidus studentisés en fonction de $\hat{Y}$

L'analyse du graphique 5 montre que notre modèle n'est pas linéaire. En effet, les points de données ne forment pas une ligne droite et il y a beaucoup de variation autour de la ligne rouge $y = 0$. L'hypothèse de linéarité n'est donc pas vérifiée.

Deuxièmement, vérifions l'hypothèse de normalité des résidus en faisant un QQ-plot entre quantiles normaux théoriques et résidus standardisés :

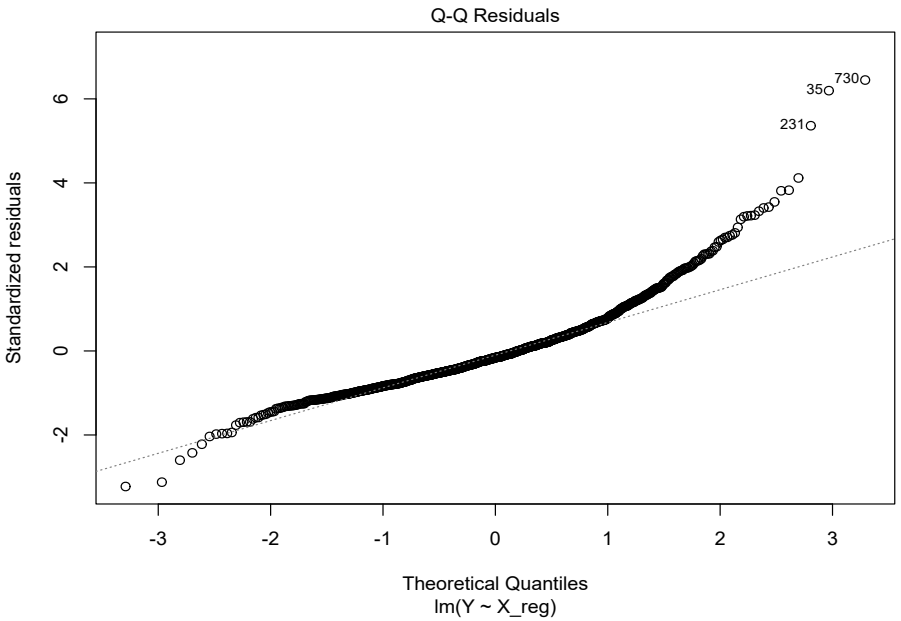


FIGURE 6 – QQ-plot entre quantiles normaux théoriques et résidus standardisés

D'après le graphique 6, les données ne semblent pas suivre une distribution normale car les points ne sont pas alignés sur une ligne droite, l'hypothèse de normalité n'est donc pas vérifiée.

Finalement, vérifions l'hypothèse d'homoscédasticité à l'aide du graphique représentant le carré des résidus studentisés en fonction de $\hat{Y}$:

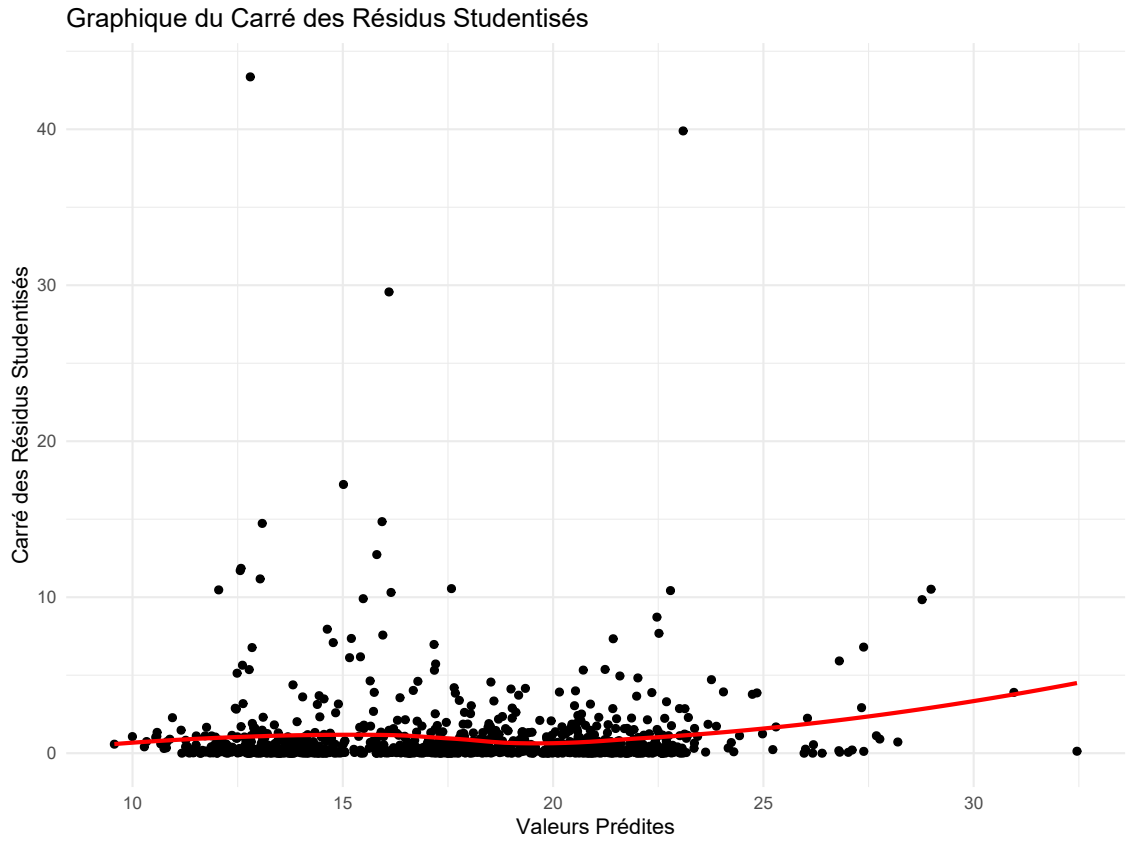


FIGURE 7 – Graphique du carré des résidus studentisés en fonction de $\hat{Y}$

On note que ce graphique est utilisé pour diagnostiquer l'hétéroscédasticité. Si les points sur le graphique sont dispersés de manière aléatoire, cela indique que les résidus sont homoscédastiques et que la variance est constante. Si les points sur le graphique sont dispersés de manière non aléatoire, cela indique que les résidus sont hétéroscédastiques et que la variance n'est pas constante [Lef23].

A la lecture du graphique 7, on voit que les points sont dispersés de manière aléatoire autour de la courbe rouge, ce qui peut indiquer que les résidus sont homoscédastiques et que la variance est constante. On conserve donc l'hypothèse d'homoscédasticité.

Malgré les doutes concernant la validité du test de Student en raison de la non-normalité des résidus et la non-linéarité, nous avons décidé de procéder à une régression sur six composantes principales sélectionnées, on obtient un $R^2 = 0.21$, ce qui signifie que le modèle ne peut expliquer que 21% de la variation de la densité de peuplement, ce qui est un pourcentage relativement faible.

Pour évaluer la performance de notre modèle, nous allons tracer le graphique des valeurs observées Y en fonction des valeurs prédites $\hat{Y}$:

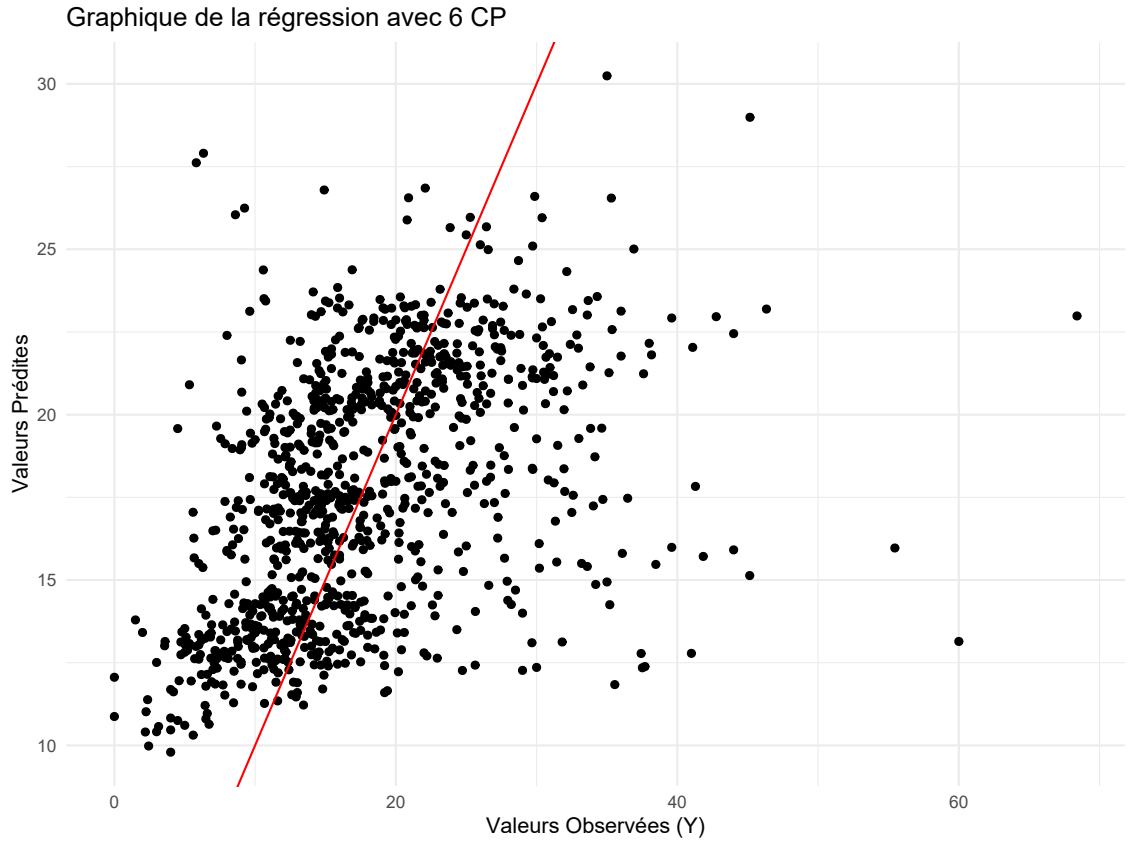


FIGURE 8 – Graphique des valeurs observées Y en fonction des valeurs prédites $\hat{Y}$

En observant le graphique 8, on remarque que les valeurs de Y faibles sont systématiquement surestimées par le modèle, tandis que les valeurs de Y élevées sont sous-estimées. Cette tendance se traduit par un nuage de points incliné de manière plus horizontale que la première bissectrice ($y = x$).

Ce qui signifie que notre modèle de régression linéaire n'arrive pas à capturer efficacement la variation des données, en particulier aux deux extrémités du spectre des valeurs de Y. Plusieurs facteurs pourraient expliquer ces difficultés, notamment la non-linéarité des relations entre les variables.

3.3 Coefficients des variables originelles

Dans cette section, nous allons retrouver les coefficients des variables originelles dans le prédicteur linéaire. Dans un premier temps, nous construirons la matrice U à l'aide des six premiers vecteurs propres u_k de l'ACP. Ces vecteurs propres seront nécessaires pour la régression sur les six premières composantes principales f_k. Pour les obtenir, nous normaliserons les composantes principales duales par la racine carrée de la valeur propre $\sqrt{\lambda_k}$ correspondante :

$$u_k = \frac{1}{\sqrt{\lambda_k}} f_k$$

On note $C = [c^1, \ldots, c^6]$ le vecteur des 6 premières composantes principales.
Comme notre tableau X est centré réduit, la métrique de l'ACP $M = I_{124}$, on écrit alors :

$$\hat{Y} = C\beta$$
$$Or,\ C = XMU = XU$$

On obtient alors : $\hat{Y} = XU\beta$, ce qui signifie que $U\beta$ correspond aux coefficients des variables X dans le prédicteur linéaire $\hat{Y}$.

Vous pouvez retrouver les résultats dans la colonne "RCP sur 6 CP" da la table 11 dans l'annexe 8.1, ainsi qu'une interprétation un peu plus détaillée de ces derniers.

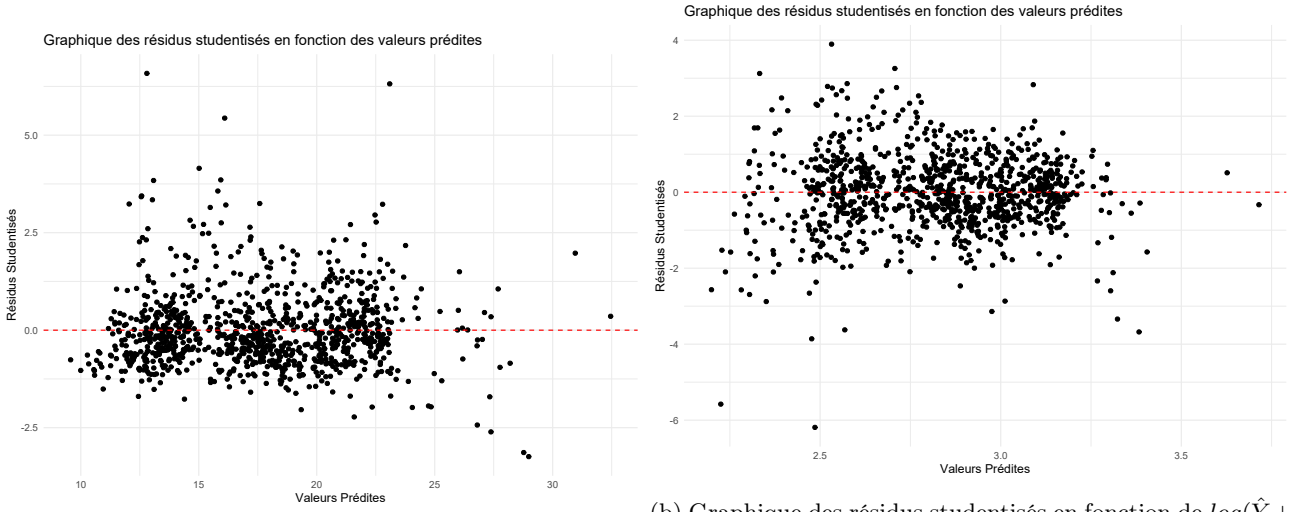
3.4 Correction de la linéarité

Dans cette section, nous allons utiliser une transformation logarithmique $log(Y+1)$ pour rectifier la non-linéarité.
Ensuite, nous allons effectuer une régression de $log(Y+1)$ sur les 8 composantes principales de X, dont les résultats sont affichés dans table 3 :

	Coefficients estimés	**Pr(>|t|)**
`Intercept`	2.821583	$< 2 \times 10^{-16}$
`X_regDim.1`	0.016452	1.21×10^{-11}
`X_regDim.2`	0.018424	7.79×10^{-12}
`X_regDim.3`	-0.013529	1.24×10^{-06}
`X_regDim.4`	0.009816	0.00112
`X_regDim.5`	-0.028612	2.69×10^{-12}
`X_regDim.6`	0.068774	$< 2 \times 10^{-16}$
`X_regDim.7`	0.006575	0.40418
`X_regDim.8`	0.019817	0.08455

TABLE 3 – Résumé de la régression linéaire multiple entre $log(Y+1)$ et les 8 CP de X

On obtient un $R^2 = 0.2684$, ce qui signifie que 26.84% de la variabilité de $log(Y+1)$ peut être expliqué par les 8 composantes principales de X. Comparé au premier modèle où on avait un $R^2 = 0.2127$, on observe une légère amélioration dans l'explication de la variance.
Pour évaluer la linéarité de la liaison, nous allons comparer les graphiques des résidus studentisés en fonction des valeurs prédites par les deux modèles $\hat{Y}$. Cela nous permettra de déterminer si l'application de la transformation log(Y+1) a amélioré la linéarité par rapport au modèle précédent :



(a) Graphique des résidus studentisés en fonction de $\hat{Y}$

(b) Graphique des résidus studentisés en fonction de $log(\hat{Y}+1)$

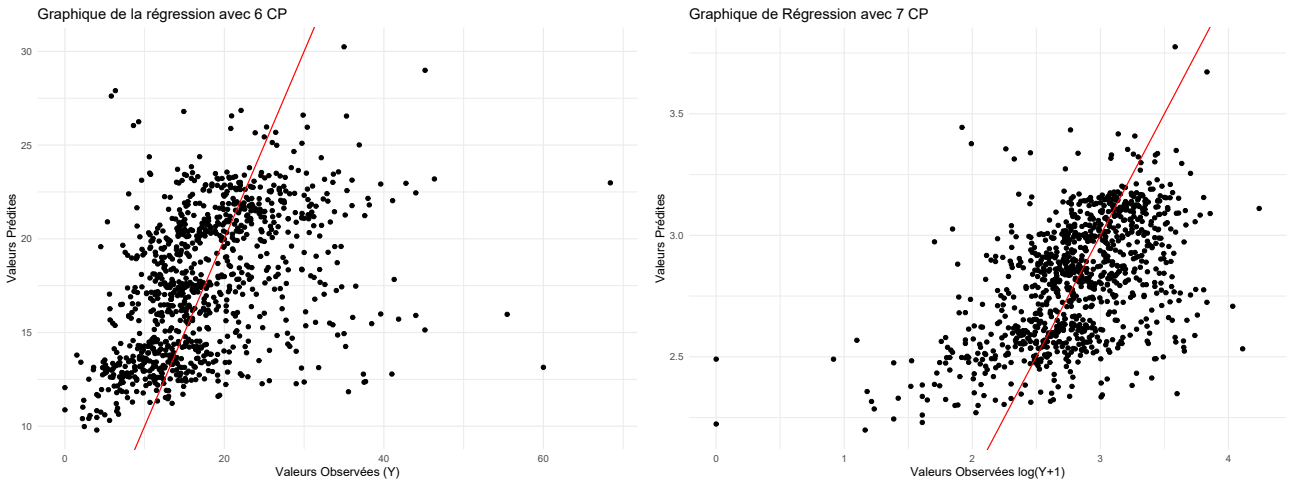
FIGURE 9 – Comparaison des résidus studentisés en fonction de $\hat{Y}$ des deux modèles

En examinant attentivement le graphique 9b des résidus studentisés en fonction des valeurs prédites après l'application de la transformation logarithmique, on peut observer que les points sont dispersés autour de ligne rouge $y = 0$. Ce qui veut dire que les données sont linéaires comparé au modèle précédent dans la figure 9a.

Nous allons désormais appliquer le test de Student pour évaluer la signification des composantes principales, en particulier sur la base des résultats présentés dans la table 3. On remarque que la p-value associée à la composante `X_reg.Dim7` est de 0.4, qui est supérieure à 0.05. Par conséquent, nous avons décidé de l'éliminer de notre modèle.

Ensuite, nous avons effectué une régression sur les composantes principales restantes, ce qui a permis d'obtenir un R^2 de 0.2679, on voit qu'il n'y pas d'amélioration comparé au modèle initial qui incluait les 8 composantes principales. Par conséquent, la suppression de la composante `X_reg.Dim7` n'a pas eu énormément d'impact sur l'ajustement global du modèle.

Pour comparer la performance des deux modèles, nous allons tracer le graphique des valeurs observées en fonction des valeurs prédites pour chaque modèle :



(a) Graphique des valeurs observées Y en fonction des valeurs prédites $\hat{Y}$

(b) Graphique des valeurs observées $log(Y+1)$ en fonction des valeurs prédites $log(\hat{Y}+1)$

FIGURE 10 – Comparaison des performances des deux modèles

A la lecture du graphique 10, on voit que la plupart des points de la figure 10b sont légèrement plus proches de la droite de référence $y = x$ en comparaison avec ceux de la figure 10a. Cette observation suggère que les valeurs observées sont légèrement plus proches des valeurs prédites dans le second modèle, qui incorpore la transformation logarithmique. Cependant, cette amélioration est relativement modeste.

4 Seconde régression sur composantes principales

Dans cette section nous allons effectuer une ACP sur les deux thèmes suivants :
— Le thème photosynthèse qui contient les variables associées aux indices EVI.
— Le thème Géographie qui contient le reste des variables.
Ensuite, nous allons reprendre ce que nous avons fait dans la section précédente avec la réusnion des composantes principales issue de chaque ACP afin de mieux modéliser Y.

4.1 ACP sur le thème Photosynthèse

4.1.1 Distribution de l'inertie

Dans cette partie, nous allons réaliser une ACP globale du tableau des variables qui représentent les indices EVI contenant 1000 individus et 23 variables.
Pour ce faire, nous allons calculer les valeurs propres de la matrice d'inertie afin de savoir le nombre d'axes que nous allons retenir pour faire notre ACP.
L'histogramme de la figure 11 donne le pourcentage d'inertie apporté par chaque dimension :

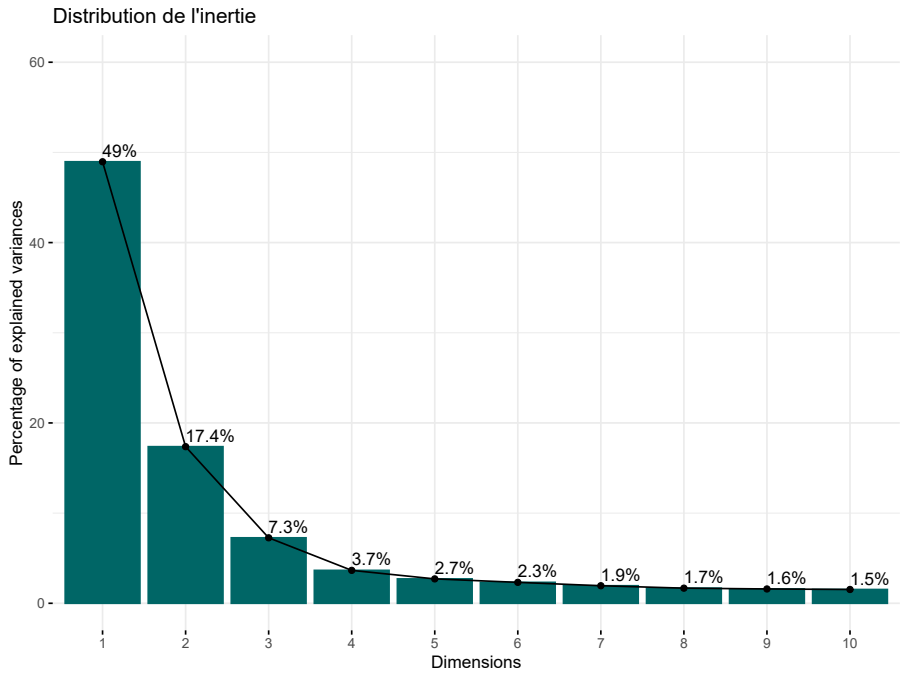


FIGURE 11 – Pourcentage d'inertie apporté par chaque dimension

A la lecture du graphique 11, on constate que les 2 premiers axes de l'analyse expriment 66.33% de l'inertie totale du jeu de données, cela signifie que 66.33% de la variabilité totale du nuage est représentée dans ce plan. C'est un pourcentage assez important, et le premier plan représente donc convenablement la variabilité contenue dans une grande part du jeu de données actif.
De plus, on remarque que 73.59% de l'inertie est capturé par les 3 premières composantes et 26.41% y est perdue, les composantes 4 à 23 possédant des valeurs propres strictement inférieures à 1 ne seront pas conservées dans la suite de nos analyses car elles apportent moins d'information qu'une seule variable toute seule.

Une estimation du nombre pertinent d'axes à interpréter suggère de restreindre l'analyse à la description des 3 premiers axes.

4.1.2 Description du plan (1,2)

La projection sur le premier plan est illustré par la figure 12 :

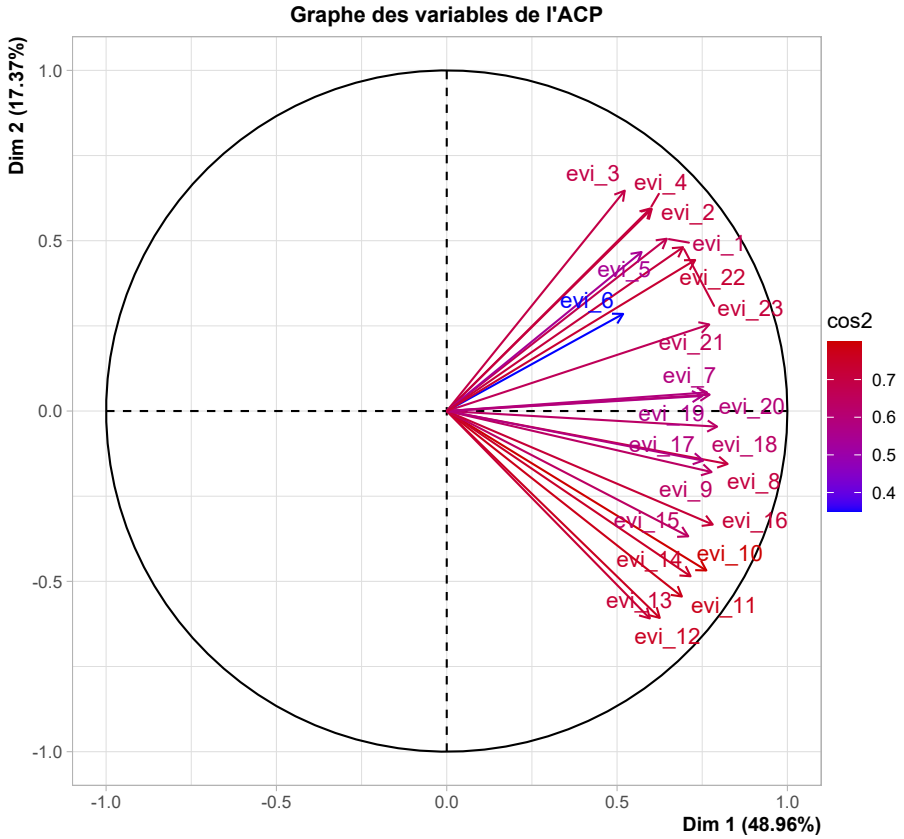


FIGURE 12 – Photographie du plan (1,2)

En terme de qualité de représentation, on voit que les variables `evi_8, evi_18, evi_16, evi_9, evi_20, evi_21, evi_10 et evi_22` sont les mieux représentées sur l'axe 1 car leur cosinus carré est le plus élevé, en revanche, les variables `evi_3` et `evi_6` présentent des cosinus carrés plus faibles et sont donc moins bien représentées sur cet axe.

En terme de corrélations, on voit que toutes les corrélations sont positives, de plus, les variables `evi_1, evi_2 et evi_22` sont fortement corrélées, de même pour les variables `evi_2, evi_4`.

4.2 ACP sur le thème Géographie

4.2.1 Distribution de l'inertie

De manière analogue, nous allons effectuer l'ACP du thème géographie qui contient les variables restantes (101 variables).

Nous allons afficher le tableau des valeurs propres en utilisant la commande `pca_geo$eig` :

Valeur propre	Pourcentage d'inertie	Pourcentage d'inertie cumulé
28.06	27.78	27.78
23.24	23.01	50.79
20.01	19.81	70.61
18.44	18.26	88.86
5.31	5.26	94.12
1.66	1.64	95.76
0.82	0.82	96.58
0.68	0.67	97.25
0.60	0.60	97.85

TABLE 4 – Description des dimensions

En observant la table 4, on constate que les 2 premiers axes de l' analyse expriment 50.79% de l'inertie totale du jeu de données, cela signifie que 50.79% de la variabilité totale du nuage est représentée dans ce plan, c'est un pourcentage assez moyen, il serait tout de même probablement préférable de considérer également dans l'analyse les dimensions supérieures ou égales à la troisième.
De plus, on remarque que 95,76% de l'inertie est représentée par les 6 premières composantes et 4.24% y est perdue, les composantes 7 à 101 possédant des valeurs propres strictement inférieures à 1 ne seront donc pas conservées.

Une estimation du nombre pertinent d'axes à interpréter suggère de restreindre l'analyse à la description des 6 premiers axes.

4.2.2 Description du plan (1,2)

Nous allons afficher la projection sur le premier plan dans la figure 13 :

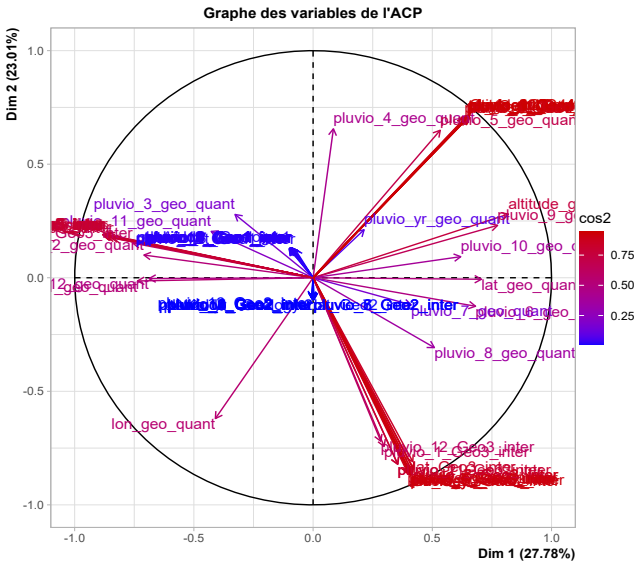


FIGURE 13 – Photographie du plan (1,2)

En terme de corrélations, nous pouvons observer une structure intéressante, où certaines variables présentent des corrélations très fortes entre elles, formant un motif en forme de trois "vaisseaux". Cela se produit particulièrement parmi les variables d'interaction. On voit également que les variables `altitude, et pluvio_9` affichent une forte corrélation entre elles. En terme de qualité de représentation, on voit que les variables quantitatives liées aux pluviométries annuelles, ainsi que certaines variables relatives aux pluviométries mensuelles, ne sont pas correctement représentées sur le plan.

4.3 Modélisation de la densité Y

Dans cette section, nous allons modéliser la densité Y à l'aide de la réunion des composantes principales des deux ACP séparées.
Pour ce faire, nous avons réalisé une ACP sur le thème photosynthèse (respectivement géographie), extrayant ainsi les 3 premières composantes principales (respectivement les 6 composantes principales). Ensuite, nous avons employé ces 9 composantes principales (stockées dans la matrice nommée `union_comp`) pour construire un modèle de régression linéaire multiple avec la variable de réponse Y, dont les résultats sont affichés dans la table 5 :

	Coefficients estimés	**Pr(>|t|)**
`Intercept`	17.620225	$< 2 \times 10^{-16}$
`union_compgeo_dim1`	0.207981	1.54×10^{-3}
`union_compgeo_dim2`	-0.298551	4.16×10^{-9}
`union_compgeo_dim3`	-0.112299	4.405×10^{-2}
`union_compgeo_dim4`	0.093756	0.09798
`union_compgeo_dim5`	-0.899434	4.72×10^{-11}
`union_compgeo_dim6`	-0.272416	0.21088
`union_compevi_dim1`	0.000291	0.99738
`union_compevi_dim2`	-0.699608	5.44×10^{-6}
`union_compevi_dim3`	0.010834	0.96905

TABLE 5 – Résumé de la régression linéaire multiple entre Y et les 9 composantes principales

On obtient un $R^2 = 0.2073$, ce qui signifie que 20.73 % de la variation dans la densité peut être expliquée par l'ensemble de ces 9 composantes principales. Ce pourcentage est relativement faible, bien inférieur à celui du premier modèle qui s'élevait à $R^2 = 0.2127$.

En analysant la table 5, on remarque que les p-values associées aux composantes `geo_dim4, geo_dim6, evi_dim1 et evi_dim3` sont toutes inférieures à 0.05, ce qui indique qu'elles ne sont pas statistiquement significatives.

Par conséquent, nous décidons de les retirer du modèle et de régresser Y sur les composantes restantes. Cependant, le modèle obtenu présente un $R^2 = 0.2035$, soit un ajustement encore plus faible que le précédent.

Pour évaluer les performances de notre modèle, nous allons tracer le graphique des valeurs observées Y en fonction des valeurs prédites $\hat{Y}$:

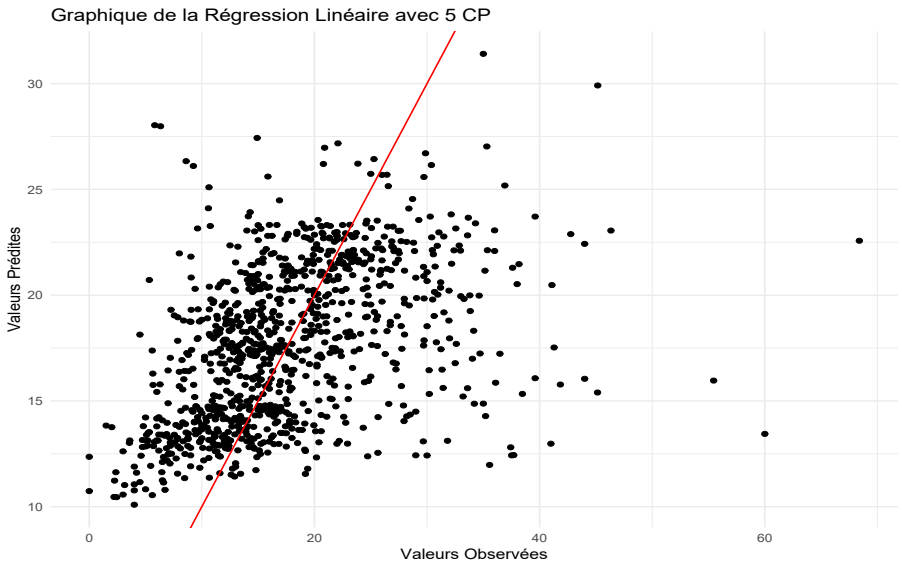


FIGURE 14 – Graphique des valeurs observées Y en fonction des valeurs prédites $\hat{Y}$

En observant le graphique 14, on constate que le nuage de points ne se rapproche toujours pas de la droite $y = x$, cela signifie que les données réelles (valeurs observées) ne sont pas très bien estimées par notre modèle.

4.4 Coefficients de la variable originelle

De manière analogue à ce que nous avons fait dans la section 3.3, nous allons retrouver les coefficients des variables originelles dans le prédicteur linéaire :

On note :

— X_{geo} le tableau des variables du thème géographie.

— X_{evi} le tableau des variables associées aux indices EVI.

— $C^{geo} = [c^1_{geo}, \ldots, c^6_{geo}]$ le vecteur des 6 composantes principales de l'ACP de X_{geo}.

— $C^{evi} = [c^1_{evi}, \ldots, c^3_{evi}]$ les composantes principales du tableau de l'ACP de X_{evi}.

— U_{geo}, U_{evi} la matrice des vecteurs propres de l'ACP de X_{geo}, X_{evi} respectivement.

On a :

$$\hat{Y} = [C^{geo}, C^{evi}]\beta$$
$$Or,\ C^{geo} = X_{geo}U_{geo}\ et\ C^{evi} = X_{evi}U_{evi}$$

On obtient alors :

$$\hat{Y} = [X_{geo}U_{geo},\ X_{evi}U_{evi}]\beta = [X_{geo}U_{geo}\beta_{geo},\ X_{evi}U_{evi}\beta_{evi}]$$

Vous pouvez retrouver les résultats dans la colonne "RCP par thèmes" de la table 11 dans l'annexe 8.1, ainsi qu'une interprétation un peu plus détaillée de ces derniers.

4.5 Correction de la linéarité

Dans cette section, nous allons utiliser une transformation logarithmique $log(Y + 1)$ pour rectifier la non-linéarité.
Ensuite, nous allons effectuer une régression de $log(Y + 1)$ sur les 9 composantes principales de X, dont les résultats sont affichés dans table 6 :

	Coefficients estimés	**Pr(>|t|)**
`Intercept`	2.821583	$< 2 \times 10^{-16}$
`union_compgeo_dim1`	0.007103	0.05186
`union_compgeo_dim2`	-0.020586	4.48×10^{-13}
`union_compgeo_dim3`	-0.004927	0.11270
`union_compgeo_dim4`	0.009456	0.00279
`union_compgeo_dim5`	-0.061489	9.83×10^{-16}
`union_compgeo_dim6`	-0.020747	0.08737
`union_compevi_dim1`	-0.011640	0.01877
`union_compevi_dim2`	-0.037276	1.36×10^{-05}
`union_compevi_dim3`	0.029638	0.05701

TABLE 6 – Résumé de la régression linéaire multiple entre $log(Y + 1)$ et les 9 CP de X

Si l'on regarde de plus près les coefficients de la table 6, on constate que certaines composantes principales de X ont un impact significatif sur la log densité du peuplement arboré ($log(Y+1)$). Par exemple, une augmentation d'une unité dans la deuxième composante principale (`union_compgeo_dim2`) est associée à une réduction de la densité du peuplement arboré. En revanche, la cinquième composante principale (`union_compgeo_dim5`) est fortement associée à une diminution significative de la densité. D'autre part, la première composante principale (`union_compgeo_dim1`) et la neuvième composante principale (`union_compevi_dim3`) présentent des effets plus faibles, mais restent significatifs.

On obtient un $R^2 = 0.2645$, ce qui signifie que 26.45% de la variabilité de $log(Y + 1)$ peut être expliqué par les 8 composantes principales de X. Comparé au premier modèle où on avait un $R^2 = 0.2073$, on observe une amélioration dans l'explication de la variance.
De plus, si l'on retire les composantes 1, 3 et 6 associées aux variables géographiques ainsi que la dernière composante associé aux evi on obtient un $R^2 = 0.079$, ce qui est encore plus faible.

5 Régression PLS

La régression PLS (Partial Least ~~Square)~~ est une technique de régularisation basée sur la réduction de la dimension des prédicteurs.

Nous allons à présent appliquer cette méthode à nos données afin de modéliser au mieux la densité. Nous utiliserons la variable $log(Y + 1)$ comme variable dépendante, compte tenu de la non-normalité des résidus et de la non-linéarité du modèle initial, ce qui renforce la qualité de notre analyse.
Afin de déterminer le meilleur nombre de composantes, nous avons réalisé une validation croisée K-fold.

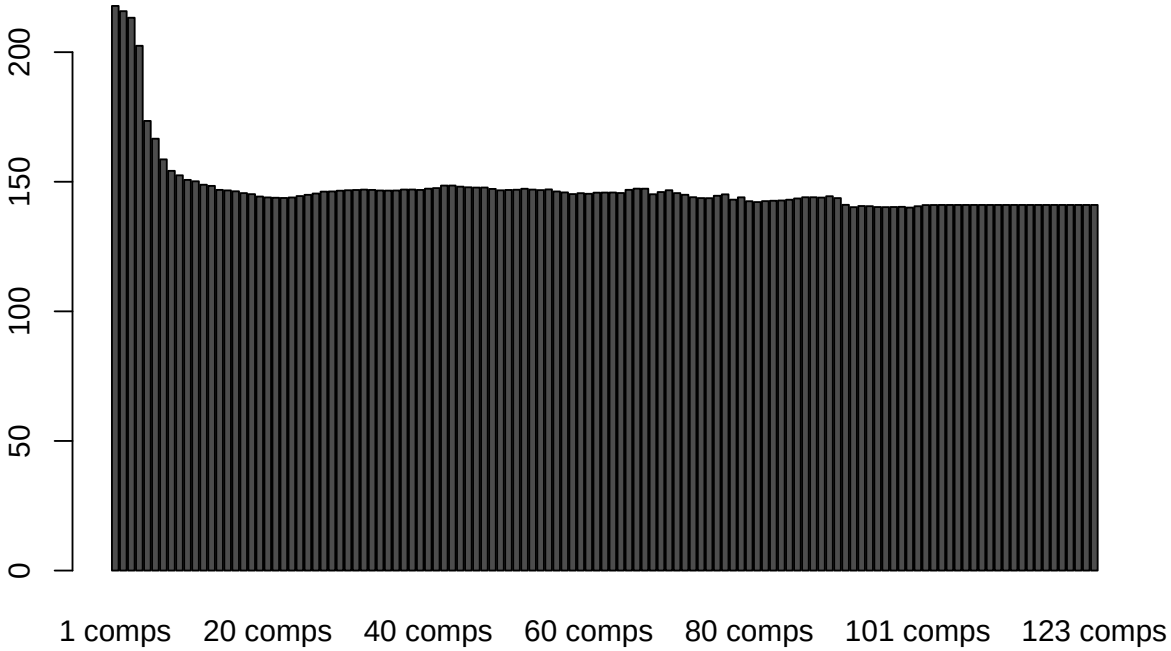


FIGURE 15 – Barplot des composantes

D'après le barplot ci-dessus, le nombre de composantes optimal à retenir semble être 22. Pour valider notre première impression, nous allons donc tracer le MSEP (Mean Square Error of Prediction) en fonction du nombre de composantes retenues.

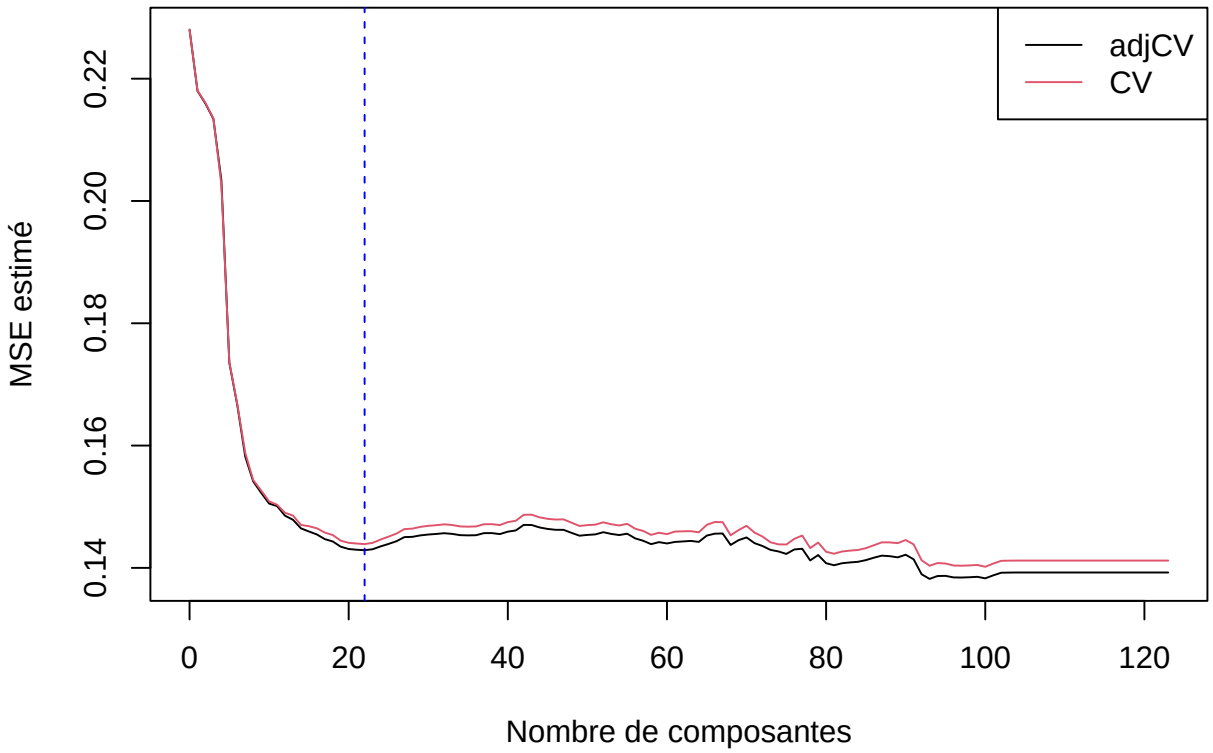


FIGURE 16 – Evolution du MSE

Le MSEP est minimal, et vaut 0.358, pour un nombre de composantes égal à 22 ce qui confirme notre première impression de choix du nombre de composantes. Dans la suite, nous retiendrons donc 22 composantes.

A présent, nous allons effectuer notre régression PLS avec ce nombre de composantes retenues choisi.

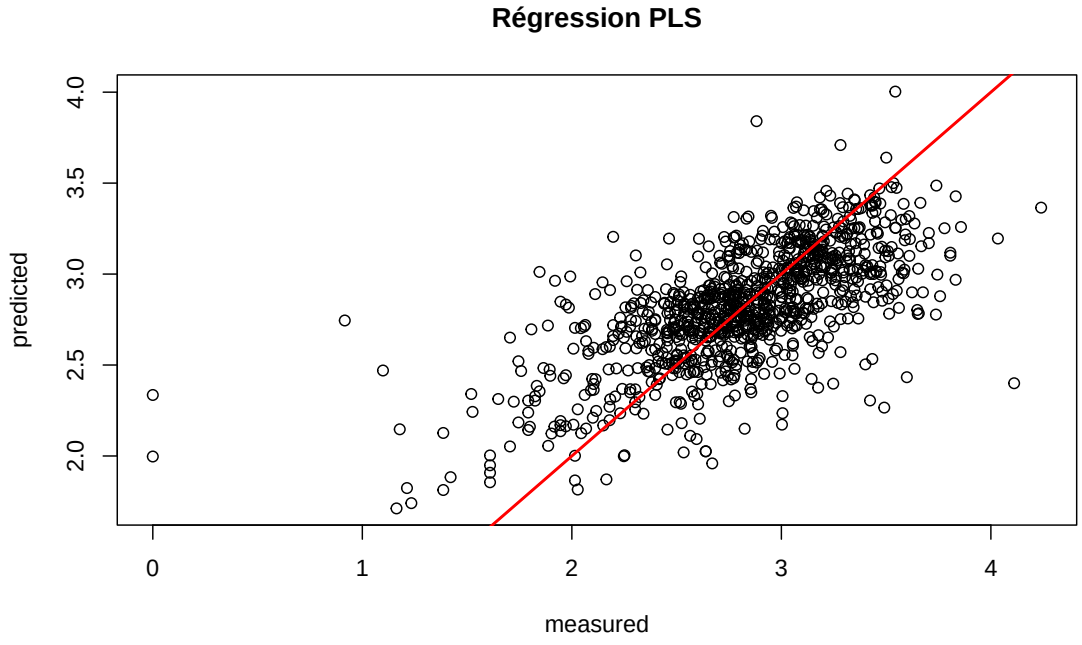


FIGURE 17 – Evolution du MSE

Pour ce modèle de régression, on trouve un R^2 égal à 0.4511, le modèle explique donc presque la moitié de la variance présente dans les données.

A présent, nous allons regarder les corrélations de la variable density avec les autres variables explicatives pour les 22 premières composantes.
Nous avons fait le choix de sélectionner seulement les corrélations supérieures à 0.7 en valeur absolue afin de ne sélectionner que les variables présentant une forte corrélation avec la variable density.
En effectuant ce premier tri, on se rend alors compte que les plus fortes corrélations sont sur les première et seconde composantes. Cela semble logique car ce sont les deux premières composantes trouvées par la PLS.

Table 7 – Extrait corrélations composante 1

lat :Geo3	lon :Geo3	lon :Geo5	altitude :Geo3	altitude :Geo5	pluvioyr :Geo3	pluvioyr :Geo5
0.86	0.92	-0.74	0.91	-0.73	0.92	-0.74
pluvio1 :Geo5	pluvio2 :Geo3	pluvio2 :Geo5	pluvio3 :Geo3	pluvio3 :Geo5	pluvio4 :Geo3	pluvio4 :Geo5
-0.72	0.88	-0.73	0.92	-0.73	0.92	-0.74

Table 8 – Extrait corrélations composante 2

lon :Geo1	altitude :Geo1	pluvioyr :Geo1	pluvio2 :Geo1	pluvio3 :Geo1	pluvio4 :Geo1	pluvio5 :Geo1
-0.72	-071	-0.71	-0.71	-0.71	-0.72	-0.72
pluvio6 :Geo1	pluvio8 :Geo1	pluvio9 :Geo1	pluvio10 :Geo1	pluvio11 :Geo5	pluvio1 :Geo1	pluvio12 :Geo1
-0.71	-0.70	-0.72	-0.72	-0.71	-0.71	-0.70

Les variables en lien avec la pluviométrie sont prédominantes dans l'analyse des corrélations sur ces deux premières composantes. Cela laisse donc à penser que ces composantes peuvent être des directions liées à la pluie.
Les 20 autres composantes semblent donc être des composantes "correctives", c'est à dire qu'elles font des corrections d'effets sur les deux premières composantes.

Voici quelques coefficients obtenus :

Variables	Valeur coefficient	Variables	Valeurs coefficient
latitude_geo	0.39	lat_Geo1	-0.108
longitude_geo	0.14	lon_Geo3	0.92
pluvio1_geo	-0.40	pluvio1_Geo5	-0.733
pluvio5_geo	-0.11	pluvio4_Geo1	-0.11
evi_1	0.18	evi_4	-0.109
evi_2	0.026	evi_9	-0.11

Intercept : 2.822

Regardons maintenant si nous retrouvons les coefficients des variables originelles dans ce prédicteur linéaire :
Pour la variable latitude-geo, le coefficient de la variable originelle est de 0.27 contre 0.39 dans notre prédicteur.
Pour la variable longitude-geo, le coefficient de la variable originelle est de 0.22 contre 0.14 dans notre prédicteur.
Pour la variable pluvio1-geo, le coefficient de la variable originelle est de -0.25 contre -0.40 dans notre prédicteur.
Pour la variable evi-1, le coefficient de la variable originelle est de -0.19 contre 0.18 dans notre prédicteur, les signes ont donc été inversés.
Si l'on regarde dans la globalité, les coefficients obtenus par la régression PLS ne sont pas très proches de ceux de la variables originelles.

6 Régressions pénalisées

Il est possible d'améliorer les propriétés numériques et la variance des estimations en considérant un estimateur biaisé des paramètres par procédure de pénalisation. Les techniques Ridge et Lasso sont basées sur la pénalisation des moindres carrés par des pénalités de type L2 et L1.

6.1 Régression Ridge

Dans le cas de la régression RIDGE, on cherche à minimiser

$$||y - X\beta||_W^2$$

en β sous la contrainte :

$$\sum_{j=1}^{p} \beta_j^2 < \gamma$$

ce qui revient à minimiser :

$$||y - X\beta||_W^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

où $\lambda > 0$ depend de γ.
La solution est alors de la forme :

$$\tilde{\beta} = (X'WX + \lambda I)^{-1} X'Wy$$

avec :

1. $W = \frac{1}{1000} I_{1000}$: la matrice de poids des individus.
2. λ : le paramètre de pénalité.
3. X : la matrice des variables explicatives.
4. y = log(Y+1) : la variable que l'on cherche à modéliser.

Pour trouver le paramètre de régularisation optimal, nous avons effectué une méthode de validation croisée à 10 ensembles. Nous nous sommes alors rendues compte que ce dernier devait se trouver dans une fenêtre entre 0 et 0.1, nous avons donc réitéré la validation croisé dans cet intervalle et nous avons trouvé $\lambda_{opt} = 0.02342595$.

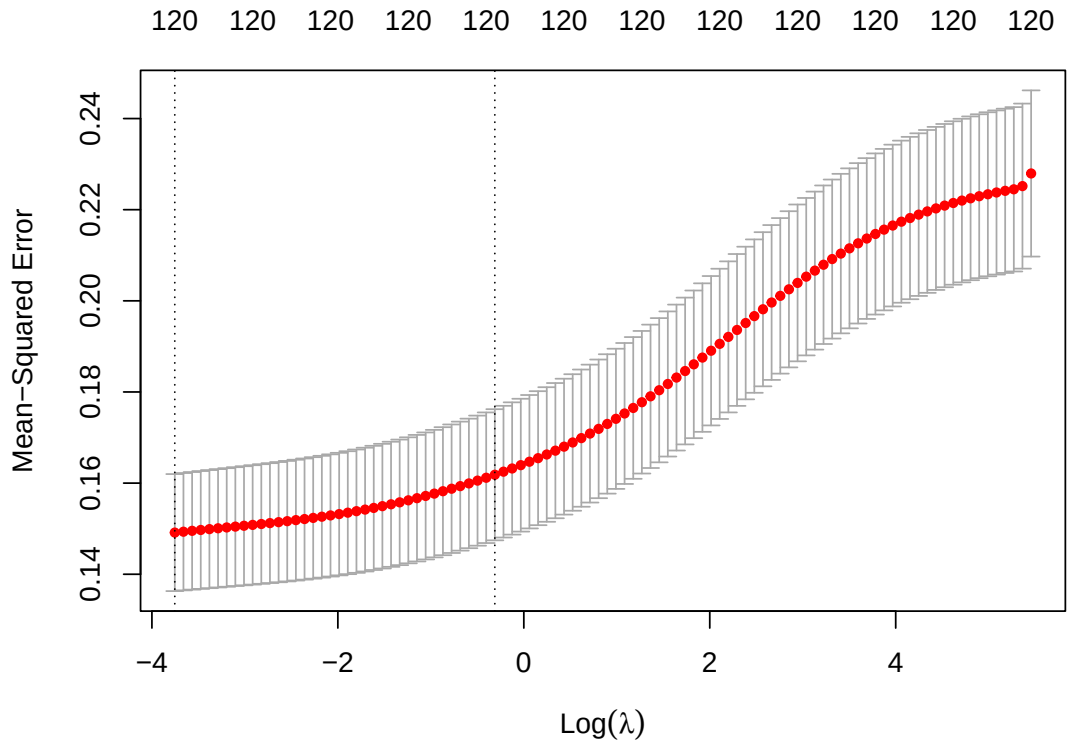


FIGURE 18 – Lambda par validation croisée

Le modèle réalisé avec ce λ_{opt} présente un R^2 égal à 0.4010496 et un MSE égal à 0.1491432. Ce modèle explique donc moins de variance que PLS.

Regardons l'évolution des coefficients de notre nouveau modèle :

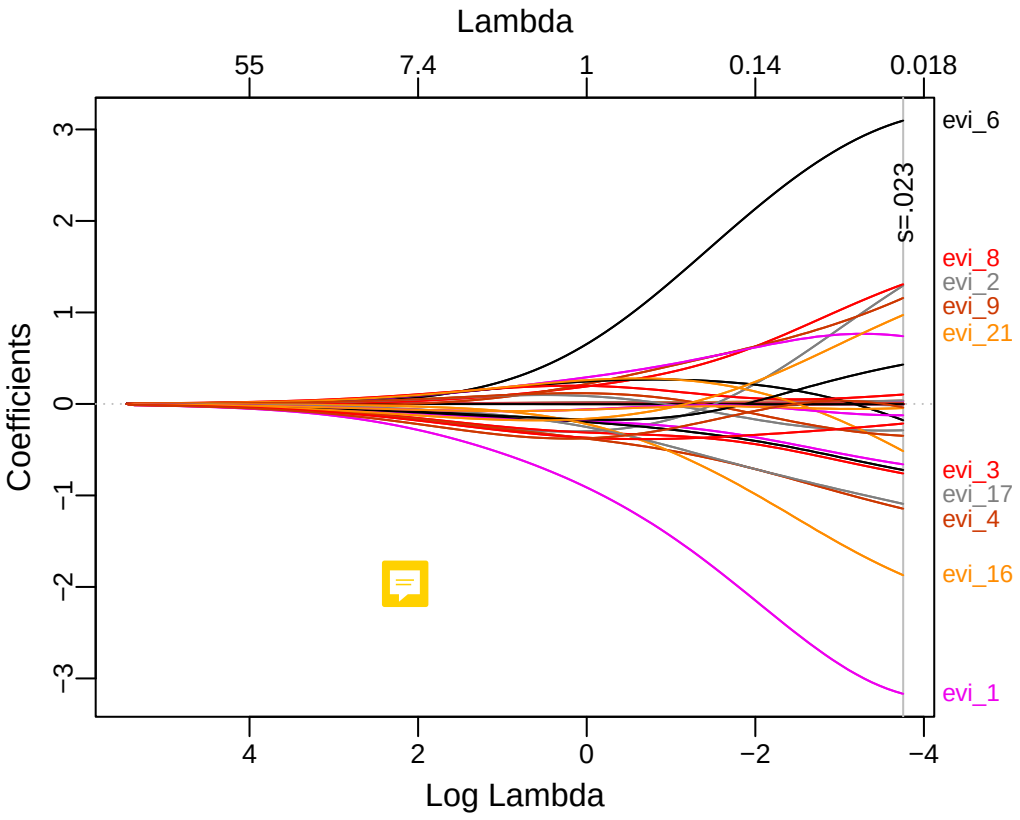


FIGURE 19 – Evolution des coeffcients en fonction de lambda

Beaucoup de coefficents sont proches de 0 et d'autres semblent être plus essentiels. Les coefficients qui ressortent le plus sont ceux liés à la photosynthèse, ce sont eux qui sont le plus contributifs à la régression.

La régression Ridge permet de contourner les problèmes de colinéarité même en présence d'un nombre important de variables explicatives ou prédictives.
Cependant, la faiblesse de ce modèle est qu'il ne permet pas de réaliser une sélection de variables ce qui peut conduire parfois à des erreurs d'interprétation.

Regardons quelques coefficients :

Variables	Valeur coefficient	Variables	Valeurs coefficient
latitude_geo	0.0345	lat_Geo1	0.0233
longitude_geo	-0.00011	lon_Geo3	-0.000715
pluvio1_geo	-0.0000538	pluvio1_Geo5	-0.00102
pluvio5_geo	-0.000429	pluvio4_Geo1	0.000846
evi_1	-3.15	evi_4	-1.14
evi_2	1.30	evi_9	1.16

Intercept : 4.9
Les coefficients obtenus par la régéression Ridge sont assez éloignés de ceux obtenus par PLS.

6.2 Régression Lasso

Dans le cas de la régression LASSO, on cherche à minimiser :

$$||y - X\beta||_W^2$$

en β sous la contrainte :

$$\sum_{j=1}^{p} |\beta_j| < \gamma$$

ce qui revient à minimiser :

$$||y - X\beta||_W^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

où $\lambda > 0$ depend de γ.

De la même manière que pour la régression Ridge, nous avons réaliser la validation croisée à 10 ensembles pour déterminer le paramètre de régularisation optimal. Nous avons alors trouvé $\lambda_{opt} = 0.0003169643$.

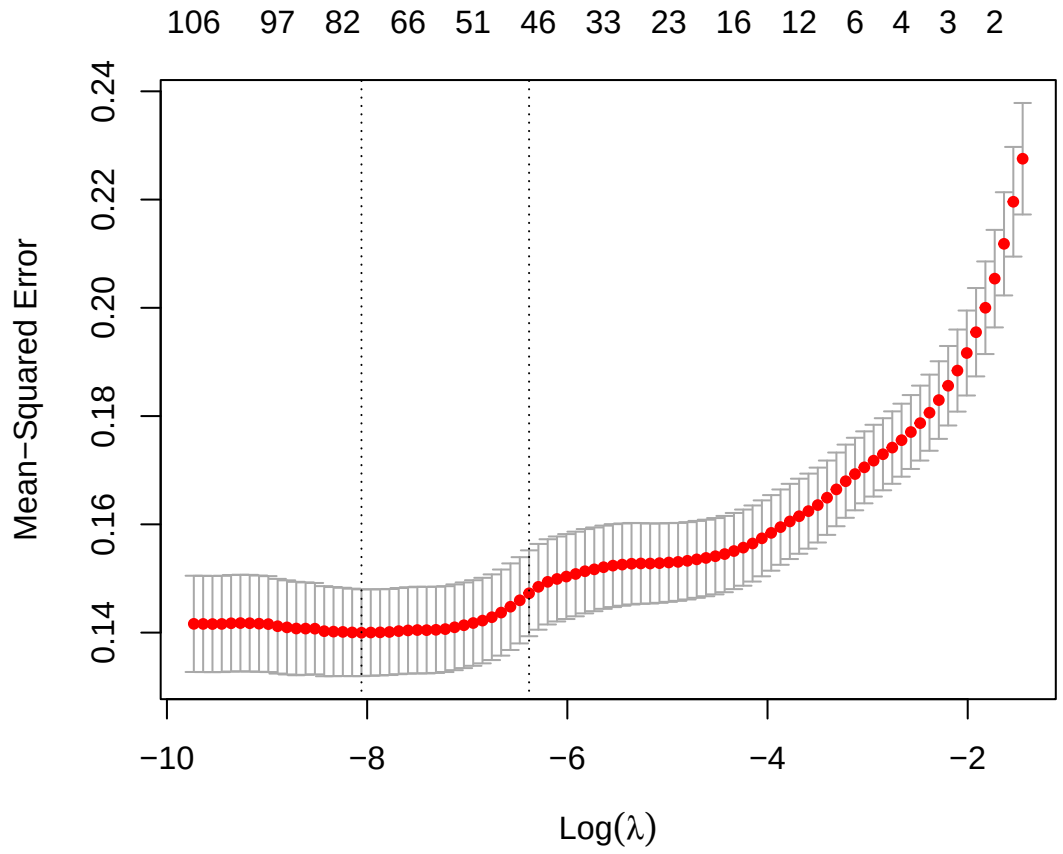


FIGURE 20 – Lambda par validation croisée

L'erreur de prédiction semble être minimale pour modèle avec entre 46 et 82 variables.

Le modèle réalisé avec ce λ_{opt} nous donne un R^2 égal à 0.4939611 (ce modèle explique donc presque la moitié de la variance présente dans les données) et un MSE égal à 0.1399783.

Regardons l'évolution des coefficients de notre nouveau modèle :

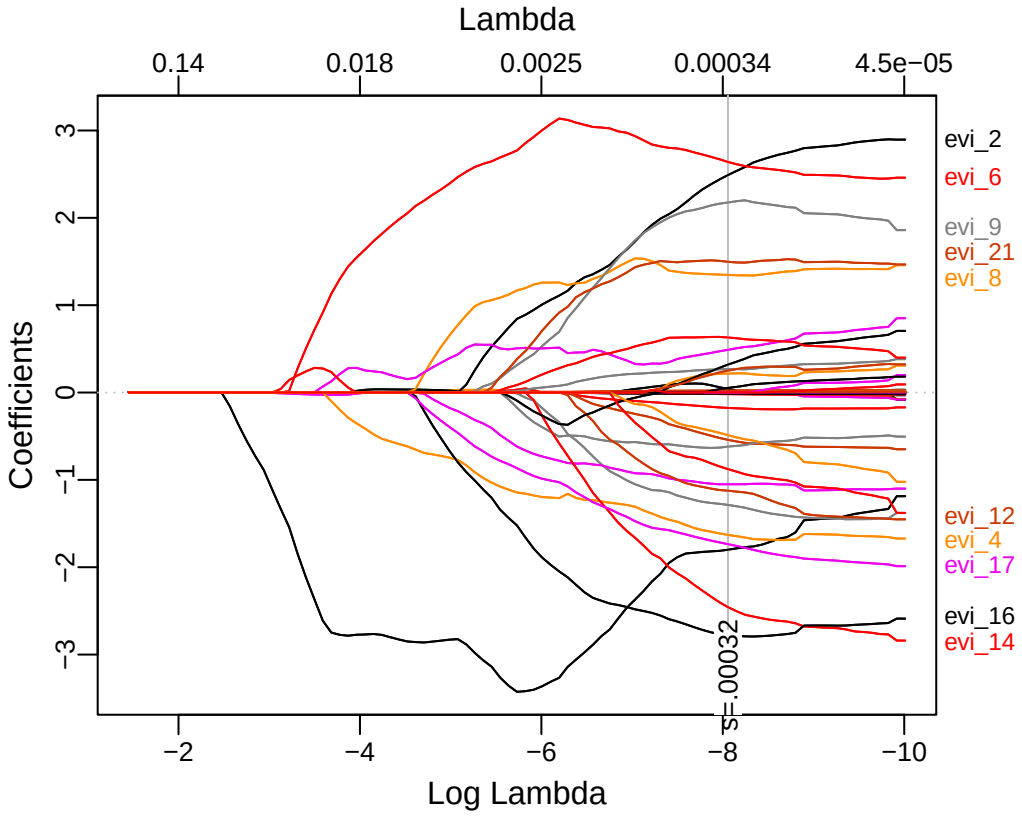


FIGURE 21 – Evolution des coefficients en fonction de lambda

De même que pour la régression Ridge, on retrouve les variables en lien avec la photosynthèse comme variables les plus contributives à la régression.

La régression Lasso permet de supprimer des variables explicatives en mettant leur poids à zéro et donc de faire de la sélection de variables.

Regardons quelques coefficients obtenus par cette dernière régression :

Variables	Valeur coefficient	Variables	Valeurs coefficient
latitude_geo	11.28	lat_Geo1	0.177
longitude_geo	-13.62	lon_Geo3	-0.058
pluvio1_geo	0.00208	pluvio1_Geo5	-0.0133
pluvio5_geo	-0.00728	pluvio4_Geo1	0.0116
evi_1	-1.88	evi_4	-1.66
evi_2	2.68	evi_9	2.372

Intercept : 1.12

Les coefficients sont de même signe que ceux obtenus par la régression Ridge, mais sont plus grand en valeur absolue. Cela s'explique par le fait que la régression Lasso met certains coefficients à zéro, ceux restant auront alors forcément plus de poids.

7 Conclusion

7.1 Comparaison des modèles

Comparons les différents prédicteurs obtenus :
Regardons le pourcentage de variance expliquée par chaque modèle :

Méthodes	R^2
Première Régression sur composantes principales	0.2127
Seconde Régression sur composantes principales	0.2073
Régression PLS	0.45
Régression Ridge	0.40
Régression Lasso	0.49

D'après ce tableau récapitulatif, le modèle qui expliquerai le plus grand pourcentage de variance serait obtenu par la régression Lasso.

Comparons maintenant les erreurs de prédictions :

Méthodes	MSE
Première Régression sur composantes principales	53.606
Seconde Régression sur composantes principales	53.97
Régression PLS	0.36
Régression Ridge	0.15
Régression Lasso	0.14

L'erreur de prédiction la plus faible est réalisée par le modèle obtenu par la régression Lasso.

Pour conclure, le modèle que nous retiendrons dans notre cas serait donc le modèle de régression LASSO.

7.2 Avantages et inconvénients

Chaque méthode présente des avantages et des inconvénients qu'il faut connaître et prendre en compte dans notre choix de modèle lorsque nous souhaitons réaliser des régressions.

Régression sur composantes principales :
Un des avantages de la régression PCR est sa facilité d'interprésation de part le fait que l'on peut avoir une représentation graphique. Néanmoins, elle n'est pas toujours très performante car les composantes sont choisies en fonction de la variance maximale de X et non en fonction de leur lien avec la variable à expliquer.

Régression PLS :
La régression PLS est adaptée aux données présentant des multicolinéarités, cependant il est parfois difficile de trouver le bon nombre de composantes à retenir.

Régression Ridge :
La régression Ridge est conseillée lorsque les variables sont fortement corrélées. Elle permet de réduire la complexité d'un modèle mais elle ne réduit pas le nombre de variables explicatives.

Régression Lasso :
Le gros avantages de la régression Lasso est qu'elle permet de faire de la sélection de variables. Il faut toutefois faire attention car lorsque les variables sont très corrélées, elle n'en sélectionne qu'une et masque donc l'influence des autres.
De plus, cette méthode n'est pas très adaptée lorsque le nombre de prédicteurs est très supérieur au nombre d'individus (ce n'était pas notre cas mais c'est important de le souligner).

8 Annexe

8.1 Tables

surface	density
5	35
15	9.6
17.5	11.09
20.5	21.02
10.5	11.81
20	24.6
5	28.8
5	14.8
6	19
5.5	2.36

TABLE 9 – Extrait des 10 premières lignes du tableau de Y

lat_Geo1	lat_Geo2	lat_Geo3	lat_Geo4	lat_Geo5	lon_Geo1	lon_Geo2	lon_Geo3	lon_Geo4	lon_Geo5
0	0	1.5	0	0	0	0	16.01	0	0
0	0	0	0	2.63	0	0	0	0	17.02
0	0	0	0	1.43	0	0	0	0	16.46
0	0	0	3.86	0	0	0	0	15.29	0
0	0	0	0	2.78	0	0	0	0	17.54
0	0	3.64	0	0	0	0	16.46	0	0
0	0	3.45	0	0	0	0	17.88	0	0
0	0	0.64	0	0	0	0	15.63	0	0
0	3.34	0	0	0	0	18.4	0	0	0
0	0	0.56	0	0	0	0	15.37	0	0

TABLE 10 – Extrait du produit d'interactions entre les variables géographiques quantitatives et les indicatrices de geology

Variables \Méthodes	RCP par thèmes	RCP sur 6 CP
lat_geo_quant	0.261918871	0.2794
lon_geo_quant	0.181598078	0.2202
altitude_geo_quant	0.050516037	0.0056
pluvio_yr_geo_quant	-0.005761298	-0.1182
pluvio_1_geo_quant	-0.241250774	-0.2568
pluvio_2_geo_quant	-0.243065744	-0.2276
pluvio_3_geo_quant	-0.171271658	-0.1748
pluvio_4_geo_quant	-0.234653476	-0.2977
pluvio_5_geo_quant	-0.076285905	-0.1132
pluvio_6_geo_quant	0.229867877	0.2315
pluvio_7_geo_quant	0.301563440	0.3123
pluvio_8_geo_quant	0.305016922	0.314
pluvio_9_geo_quant	0.118831922	0.0768
pluvio_10_geo_quant	-0.029499528	-0.1031
pluvio_11_geo_quant	-0.319578527	-0.361
pluvio_12_geo_quant	-0.242916509	-0.25
lat_Geo1_inter	0.040278521	0.039
lat_Geo2_inter	0.004605553	0.0239
lat_Geo3_inter	0.140607147	0.1153
lat_Geo4_inter	-0.032324266	-0.0071
lat_Geo5_inter	0.086561179	0.092
lon_Geo1_inter	0.013812303	0.0114
lon_Geo2_inter	-0.008524078	0.0086
lon_Geo3_inter	0.045359131	0.0339
lon_Geo4_inter	-0.035671280	-0.0143
lon_Geo5_inter	0.004982918	-0.0022
altitude_Geo1_inter	0.014182877	0.01
altitude_Geo2_inter	-0.010907950	0.0058
altitude_Geo3_inter	0.061118281	0.0428
altitude_Geo4_inter	-0.044580225	-0.0474
altitude_Geo5_inter	0.009261927	0.0046
pluvio_yr_Geo1_inter	0.013308881	0.0106
pluvio_yr_Geo2_inter	-0.011305809	0.0053
pluvio_yr_Geo3_inter	0.035525038	0.0238
pluvio_yr_Geo4_inter	-0.041422788	-0.0328
pluvio_yr_Geo5_inter	-0.001674699	-0.0098
pluvio_1_Geo1_inter	-0.005517798	-0.0089
pluvio_1_Geo2_inter	-0.025205867	-0.0118
pluvio_1_Geo3_inter	-0.112170908	-0.0986
pluvio_1_Geo4_inter	-0.058769615	-0.0825
pluvio_1_Geo5_inter	-0.040353775	-0.0537
pluvio_2_Geo1_inter	0.004691773	0.0025
pluvio_2_Geo2_inter	-0.018898955	-0.0037
pluvio_2_Geo3_inter	-0.052872095	-0.0464
pluvio_2_Geo4_inter	-0.041634105	-0.0292
pluvio_2_Geo5_inter	-0.031725113	-0.042
pluvio_3_Geo1_inter	0.014190794	0.0114
pluvio_3_Geo2_inter	-0.015436671	0,0003
pluvio_3_Geo3_inter	0.018476766	0.0104
pluvio_3_Geo4_inter	-0.039470004	-0.0238
pluvio_3_Geo5_inter	-0.021433896	-0.0337
pluvio_4_Geo1_inter	0.012516505	0.0104
pluvio_4_Geo2_inter	-0.012884241	0.003
pluvio_4_Geo3_inter	0.004306334	-0.0021
pluvio_4_Geo4_inter	-0.049274000	-0.0565
pluvio_4_Geo5_inter	-0.016959772	-0.0262
pluvio_5_Geo1_inter	0.017041288	0.0145
pluvio_5_Geo2_inter	-0.016328490	-0,0006
pluvio_5_Geo3_inter	0.031254217	0.0194

pluvio_5_Geo4_inter	-0.045105944	-0.0439
pluvio_5_Geo5_inter	-0.008986440	-0.0173
pluvio_6_Geo1_inter	0.016900323	0.0138
pluvio_6_Geo2_inter	-0.000506403	0.018
pluvio_6_Geo3_inter	0.083111700	0.0644
pluvio_6_Geo4_inter	-0.037516537	-0.0222
pluvio_6_Geo5_inter	0.022077052	0.0189
pluvio_7_Geo1_inter	0.030890892	0.0286
pluvio_7_Geo2_inter	-0.006992500	0.0107
pluvio_7_Geo3_inter	0.109456172	0.0875
pluvio_7_Geo4_inter	-0.030351902	-0.0027
pluvio_7_Geo5_inter	0.034441418	0.0279
pluvio_8_Geo1_inter	0.029011036	0.0266
pluvio_8_Geo2_inter	-0.005092900	0.0125
pluvio_8_Geo3_inter	0.110485504	0.087
pluvio_8_Geo4_inter	-0.030382834	-0.0027
pluvio_8_Geo5_inter	0.046038883	0.0443
pluvio_9_Geo1_inter	0.015269224	0.0124
pluvio_9_Geo2_inter	-0.011145187	0.0055
pluvio_9_Geo3_inter	0.052126841	0.0367
pluvio_9_Geo4_inter	-0.041137590	-0.0332
pluvio_9_Geo5_inter	0.005816191	-0,0008
pluvio_10_Geo1_inter	0.012082954	0.0097
pluvio_10_Geo2_inter	-0.014949170	0,0009
pluvio_10_Geo3_inter	0.036478807	0.0221
pluvio_10_Geo4_inter	-0.045809281	-0.0457
pluvio_10_Geo5_inter	-0.001852768	-0.0088
pluvio_11_Geo1_inter	0.001518433	-0.0014
pluvio_11_Geo2_inter	-0.019780092	-0.0049
pluvio_11_Geo3_inter	-0.055802745	-0.0532
pluvio_11_Geo4_inter	-0.054777426	-0.071
pluvio_11_Geo5_inter	-0.046894764	-0.0595
pluvio_12_Geo1_inter	-0.006900048	-0.011
pluvio_12_Geo2_inter	-0.030245497	-0.0181
pluvio_12_Geo3_inter	-0.124601935	-0.1097
pluvio_12_Geo4_inter	-0.049445438	-0.0528
pluvio_12_Geo5_inter	-0.058880451	-0.0748
Geology 1	0.013272230	0.0109
Geology 2	-0.011755099	0.0048
Geology 3	0.034521197	0.0238
Geology 4	-0.040285272	-0.0273
Geology 5	-0.001563678	-0.0091
evi_1	-0.202338894	-0.1968
evi_2	-0.237434469	-0.2065
evi_3	-0.258842732	-0.2067
evi_4	-0.237864760	-0.1875
evi_5	-0.186721432	-0.1165
evi_6	-0.114061278	-0.0468
evi_7	-0.021849849	-0.0178
evi_8	0.062135643	0.113
evi_9	0.071639501	0.0635
evi_10	0.187212438	0.1878
evi_11	0.217813433	0.1867
evi_12	0.242937685	0.2183

evi_13	0.243549498	0.2391
evi_14	0.194024412	0.2171
evi_15	0.147131649	0.1573
evi_16	0.133299457	0.1154
evi_17	0.058205741	0.033
evi_18	0.018315561	-0.0159
evi_19	-0.017663028	-0.0248
evi_20	-0.019227930	-0.0271
evi_21	-0.101620017	-0.1074
evi_22	-0.177359054	-0.1932
evi_23	-0.192380400	-0.2208

TABLE 11 – Table des coefficients de X pour chaque méthode

Interprétons les coefficients de chaque colonne séparément :

- RCP par thèmes :

Les variables qui ont des coefficients positifs (marquées en vert) augmentent la densité du peuplement arboré. Par exemple, une augmentation d'une unité dans la variable "lat_geo_quant" est associée à une augmentation de la densité du peuplement arboré, tout comme "lon_geo_quant" et "altitude_geo_quant".

Les variables avec des coefficients négatifs (le reste) ont un effet négatif sur la densité du peuplement arboré. Par exemple, une augmentation d'une unité dans la variable `"pluvio_yr_geo_quant"` ou `"pluvio_1_geo_quant"` est associée à une diminution de la densité du peuplement arboré.

- RCP sur 6 CP :

Les variables avec des coefficients positifs (marquées en vert) augmentent la densité du peuplement arboré. Par exemple, une augmentation d'une unité dans `"lat_geo_quant"` ou `"lon_geo_quant"` a un impact positif sur la densité du peuplement arboré.

Les variables avec des coefficients négatifs diminuent la densité du peuplement arboré. Par exemple, une augmentation d'une unité dans "pluvio_yr_geo_quant" ou "pluvio_1_geo_quant" est associée à une diminution de la densité du peuplement arboré.

8.2 Code

```r
#chargement des librairies necessaires
library(Factoshiny)
library(FactoMineR)
library(corrplot)
library(ade4)
library(dplyr)
library(ggplot2)

# Question 1 :

# lecture des donnees:
data_genus <- read.table("Datagenus.csv", sep = ";", header = TRUE)
#on retire la variable "forest"
data_genus = data_genus %>% select(-forest)
data_genus = data_genus[-nrow(data_genus),]

#a) Calcul de Y
data_genus = data_genus %>% mutate(density = rowSums(select(., starts_with("gen"))) / surface)

#b)
# creer les indicatrices de la variable geology
data_genus$geology = as.factor(data_genus$geology)
df = data.frame(data_genus$geology)

df_geo = acm.disjonctif(df)
for(i in 1:ncol(df_geo)){
  names(df_geo)[i] = paste("Geology", i)
}

# Calcul du produit d'interactions entre les variables geo quantitatives et les indicatrices
    de la variable geology
var_geo_quant <- data_genus %>% select(lat, lon, altitude, starts_with("pluvio"))

# Calcul du produit d'interaction entre geographie et geology
prod_inter <- data.frame(matrix(vector(), nrow = nrow(df_geo)))
geo_colnames <- colnames(var_geo_quant)

for (i in 1:length(geo_colnames)) {
  for (j in 1:5) {
    x <- var_geo_quant[, i] * df_geo[, j]
    name_col <- sprintf("%s_Geo%d_inter", geo_colnames[i], j)
    prod_inter[name_col] <- x
  }
}

#On renomme les colonnes pour du dataframe des variables geographiques quantitatives
var_geo_quant <- var_geo_quant %>%
  rename_all(~ paste0(., "_geo_quant"))

#On reconstruit la matrice X
EVI = data_genus %>% select(., starts_with("evi"))
X = cbind(var_geo_quant ,prod_inter, data_genus$geology, EVI)
X = sqrt(n/n-1) * scale(as.numeric(X))

# Inventaire des multicolinearites

X_numeric <- apply(X[, -which(names(X) == "data_genus$geology")], 2, as.numeric)
corrvars <-round(cor(X_numeric, use = "pairwise.complete.obs"),2)
corrplot(corrvars,method = "circle", is.corr = TRUE)

# Question 2

# a) ACP globale
#Graphique de la distribution de l'inertie
X_acp = cbind(var_geo_quant ,prod_inter, df_geo, EVI)
res.PCA<-PCA(X_acp,graph=FALSE)
fviz_eig(res.PCA, addlabels = TRUE, ylim=c(0,60), barfill="#006666", barcolor = "#006666",
    main="Distribution de l'inertie")

# affichage des valeurs propres
res.PCA$eig

#diagramme en batons du cos2 de chaque variable :
fviz_cos2(res.PCA, choice = "var", axes = c(1,2), col=c("#248f8f"), top=60)
```

```r

# Photographie du plan(1,2)
plot.PCA(res.PCA,choix='var',habillage = 'cos2',select='cos2  0',unselect=0,title="Graphe des variables de l'ACP")

#b) Modelisation de la densite Y

# Regression sur 8 CP
res.pca = PCA(X_acp, graph = FALSE, ncp = 8)
X_reg <- as.matrix(res.pca$ind$coord[, 1:8])
Y = data_genus$density
f1 = lm(Y~X_reg)
summary(f1)

# Graphique de y en fonction de y_hat
# On cree un dataframe pour les valeurs observees et predites
data <- data.frame(
  Y = Y,
  Y_hat = predict(f1)
)

ggplot(data, aes(x = Y, y = Y_hat)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "red") + # Ajoute une ligne de regression (y = x)
  labs(x = "Valeurs Observees (Y)", y = "Valeurs Predites") +
  ggtitle("Graphique de la regression avec 8 CP ") +
  theme_minimal()

#Calcul du mse
mean((residuals(f1))^2)

# Regression sur 6 CP

res.pca2 = PCA(X_acp, graph = FALSE, ncp = 6)
X_reg2 <- as.matrix(res.pca2$ind$coord[, 1:6])
Y = data_genus$density
f2 = lm(Y~X_reg2)
summary(f2)

# Graphique de y en fonction de y_hat

data2 <- data.frame(
  Y = Y,
  Y_hat = predict(f2)
)

ggplot(data2, aes(x = Y, y = Y_hat)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "red") +
  labs(x = "Valeurs Observees (Y)", y = "Valeurs Predites") +
  ggtitle("Graphique de la regression avec 6 CP ") +
  theme_minimal()

# Verification de la validite du test de student

#Linearite
#residus studentises en fonction de Y_chapeau
res = rstudent(f1) #residus studentises
# On cree un dataframe pour les residus et les valeurs predites
data_std <- data.frame(
  Y_hat = predict(f1),
  Residuals = res
)

# Graphique de dispersion des residus en fonction des valeurs predites
ggplot(data_std, aes(x = Y_hat, y = Residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(x = "Valeurs Predites", y = "Residus Studentises") +
  ggtitle("Graphique des residus studentises en fonction des valeurs predites") +
  theme_minimal()

# Normalite
# QQ-plot
```

```r
  plot(f1, 2)
#homoscedasticite
library(ggplot2)
res = rstudent(f1)
#  dataframe pour les carres des residus studentises et les valeurs predites
residuals_data <- data.frame(
  Y_hat = predict(f1),
  stud_res = res^2
)

# Graphique du carre des residus studentises en fonction des valeurs predites
ggplot(residuals_data, aes(x = Y_hat, y = stud_res)) +
  geom_point() +
  geom_smooth(method = "loess", se = FALSE, color = "red") +
  labs(x = "Valeurs Predites", y = "Carre des Residus Studentises") +
  ggtitle("Graphique du Carre des Residus Studentises") +
  theme_minimal()

# c) Coefficients de la variable originelle
phi = as.matrix(res.pca$var$coord)
lambda = res.pca$eig[1:8,1]
U = matrix(0,nrow=124, ncol=8)
for (i in c(1:8)) {
U[,i] = phi[,i] * 1/sqrt(lambda[i])
}
U = U[,-c(7,8)] # On ne garde que les 6 u_k qui nous interessent
coeff <- U %*% f2$coefficients[-1]
rownames(coeff) <- colnames(X_acp)

# d) Correction log

res.pca = PCA(X_acp, graph = FALSE, ncp = 8)
X_reg <- as.matrix(res.pca$ind$coord[, 1:8])
Ylog = log1p(data_genus$density)
f3 = lm(Ylog~X_reg)
summary(f3)

#Linearite amelioree?
#residus studentises en fonction de Y_chapeau
res3 = rstudent(f3) #residus studentises
data_std <- data.frame(
  Y_hat = predict(f3),
  Residuals = res3
)

# Graphique de dispersion des residus en fonction des valeurs predites
ggplot(data_std, aes(x = Y_hat, y = Residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(x = "Valeurs Predites", y = "Residus Studentises") +
  ggtitle("Graphique des residus studentises en fonction des valeurs predites") +
  theme_minimal()

# Regression sur 7CP
X_7 <- as.matrix(res.pca$ind$coord[, -7])
f2_bis <- lm(Ylog ~ X_7)
summary(f2_bis)

# Graphique de y en fonction de y_hat
data <- data.frame(Ylog = Ylog,
                   Fitted = f2_bis$fitted.values)
ggplot(data, aes(x = Ylog, y = Fitted)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "red") +
  labs(x = "Valeurs Observees log(Y+1)", y = "Valeurs Predites") +
  ggtitle("Graphique de Regression avec 7 CP") +
  theme_minimal()

# 3) Seconde regression sur CP

# ACP EVI
# Distribution de l'inertie
pca_evi<-PCA(EVI,graph=FALSE)
fviz_eig(pca_evi, addlabels = TRUE, ylim=c(0,60), barfill="#006666", barcolor = "#006666",
    main="Distribution de l'inertie")
```

```r

# Photographie du plan (1,2)
plot.PCA(pca_evi,choix='var',habillage = 'cos2',title="Graphe des variables de l'ACP")

# ACP GEOGRAPHIE
GEO = X_acp %>% select(-starts_with("evi"))

# valeurs propres
pca_geo<-PCA(GEO,graph=FALSE)
pca_geo$eig

# Photographie du plan (1,2)
plot.PCA(pca_geo,choix='var',habillage = 'cos2',title="Graphe des variables de l'ACP")

# a) Modelisation de Y

acp_evi = PCA(EVI, ncp = 3, graph = FALSE)
acp_geo = PCA(GEO, ncp = 6, graph = FALSE)

# Reunion des composantes principales
union_comp = as.matrix(cbind(acp_geo$ind$coord, acp_evi$ind$coord))
colnames(union_comp) <- paste0("geo_dim", 1:9)
colnames(union_comp)[7:9] <- paste0("evi_dim", 1:3)
f_union = lm(Y~union_comp)
summary(f_union)

#calcul du mse
mse_m2 = mean(residuals(f_union)^2)

# On retire les composantes 4, 6, 7 et 9:

f_union2 <- lm(Y ~ union_comp[, -c(4, 6, 7, 9)])
summary(f_union2)

# Graphique de y en fonction de y_hat
data_union <- data.frame(
  Y = Y,
  Y_hat = f_union2$fitted.values
)
ggplot(data_union, aes(x = Y, y = Y_hat)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "red") +
  labs(x = "Valeurs Observees", y = "Valeurs Predites") +
  ggtitle("Graphique de la Regression Lineaire avec 5 CP") +
  theme_minimal()

# coefficients de la variable originelle
coeff_u <- rbind(acp_geo$svd$V[, -c(4, 6)] %*% f_union2$coefficients[-c(1, 6)],acp_evi$svd$V[,
    2] %*% as.matrix(f_union2$coefficients[6]))
rownames(coeff_u) <- colnames(X_acp)

# Correction log

Ylog = log1p(data_genus$density)
f_logcp = lm(Ylog~union_comp)
summary(f_logcp)

# On retire la composante 1,3 et 6 de Geo et  3 de EVI

f_logcp2 <- lm(Ylog ~ union_comp[, -c(1,3,6, 9)])
summary(f_logcp2)

# Question 4

## Regression PLS

# on cree un vecteur avec la variable density
density_val <- c()
density_val <- data$density
log_density <- log1p(density_val)
log_density2 <- t(log_density)

X2 <- cbind(log_density, X)
X2 = X2[-nrow(X2),]
data_pls <- data.frame(X2)

```

```
treePls=plsr(log_density~., data=data_pls,validation="CV")
treePls$validation$PRESS
barplot(treePls$validation$PRESS)
plot(treePls)
msepcv.pls <- MSEP(treePls, estimate=c("adjCV","CV")) #estime MSE
plot(msepcv.pls, lty=1, type="l", legendpos="topright", main="",xlab="Nombre de composantes",
    ylab="MSE estime")
abline(v=22,lty=2,col="blue")

summary(treePls)
# 45.11 R2

treepls1 <- plsr(data_pls[,1]~., data=data_pls, ncomp=22)
plot(treepls1, ncomp = 22, line = TRUE, main="Regression PLS")
abline(a=0,b=1,col='red', lwd=2)
cor(data_pls[,1],treepls1$fitted.values[,1,22])^2

summary(treepls1)

reg.pls=lm(data_pls[,1]~treepls1$scores)
summary(reg.pls)
info =summary(reg.pls)
R2PLS=info$r.squared

X3 <- as.data.frame(lapply(data_pls, as.numeric))

corplsTree = cor(x=as.matrix(X3[,2:121]),y=treepls1$scores)

plot(corplsTree)
text(c(1,45), corplsTree[,2]-0.02,labels=rownames(corplsTree))
abline(h=0)
abline(v=0)

li = list()
for (l in 1:22){
  x=0
  i=0
  for ( j in 1:120){
    if (abs(corplsTree[j,l])>=0.7){
      i=i+1
      x[i]=j
    }
  }
  li[[l]]=corplsTree[x,l]
}
li
# coefficients

reg.pls1=lm(data_pls[,1]~treepls1$scores)
logdensitymodelpls <- as.matrix(treepls1$coefficients[,1,])%*%as.matrix(reg.pls1$coefficients
    [2:23])

reg.pls1$coefficients
var(treepls1$scores)

Beta_chapeau_PLS = c(reg.pls1$coefficients[1], as.matrix(treepls1$coefficients[,1,])%*%as.
    matrix(reg.pls1$coefficients[2:23]))

plot(treepls1, "correlation")

### RIDGE

library(glmnet)
library(plotmo)

ridgelog = glmnet(x=X2[, 2:121] , y=X2[,1], family="gaussian",alpha=0)
plot_glmnet(ridgelog, s=cvridgelog$lambda.min)

cvridgelog = cv.glmnet(x=as.matrix(X2[,2:121]), y=as.matrix(X2[,1]), family="gaussian",
                       alpha=0, nfolds=10)
plot(cvridgelog)

print(min(cvridgelog$cvm))
print(cvridgelog$lambda.min)

cvridgelog
```

```r

model4 <- glmnet(x=as.matrix(X2[,2:121]), y=as.matrix(X2[,1]),lambda=cvridgelog$lambda.min,
    alpha=0, family="gaussian")
model4$beta
coef(model4)

MSERIDGE=cvridgelog$cvm[cvridgelog$index[1]]
R2ridge=model4$dev.ratio
beta_chapeau_ridge=coef(model4)

sorted_coefficients <- beta_chapeau_ridge[order(abs(beta_chapeau_ridge), decreasing = TRUE)]
sorted_coefficients

noms_variables <- colnames(data_pls[,2:121])
noms_variables_tries <- noms_variables[order(abs(beta_chapeau_ridge), decreasing = TRUE)]
noms_variables_tries


## LASSO

lassolog = glmnet(x=X2[, 2:121] , y=X2[,1], family="gaussian",alpha=1)
plot_glmnet(lassolog,s=cvlassolog$lambda.min)

cvlassolog = cv.glmnet(x=as.matrix(X2[,2:121]), y=as.matrix(X2[,1]), family="gaussian",
                       alpha=1, nfolds=10)
plot(cvlassolog)

print(min(cvlassolog$cvm))
print(cvlassolog$lambda.min)

cvlassolog


model5 <- glmnet(x=as.matrix(X2[,2:121]), y=as.matrix(X2[,1]),lambda=cvlassolog$lambda.min,
    alpha=0, family="gaussian")
model5$beta
coef(model5)


MSELASSO=cvlassolog$cvm[cvlassolog$index[1]]
R2lasso=model5$dev.ratio
beta_chapeau_lasso=coef(model5)

sorted_coefficients2 <- beta_chapeau_lasso[order(abs(beta_chapeau_lasso), decreasing = TRUE)]
sorted_coefficients2

noms_variables2 <- colnames(data_pls[,2:121])
noms_variables_tries2 <- noms_variables[order(abs(beta_chapeau_lasso), decreasing = TRUE)]
noms_variables_tries2
```

Références

[HRL23] Hadley Wickham, Romain François et Lionel Henry. *Introduction to dplyr*. https://cran.r-project.org/web/packages/dplyr/vignettes/dplyr.html. 3 sept. 2023.

[Lef23] Vincent Lefieux. *Réalisez des modélisations de données performantes*. OpenClassrooms. 2023. url : https://openclassrooms.com/fr/courses/4525326-realisez-des-modelisations-de-donnees-performantes/5754143-analysez-les-resultats.

[DDT] Stéphane Dray, Anne-Béatrice Dufour et Jean Thioulouse. *ade4: Analysis of Ecological Data: Exploratory and Euclidean Methods in Environmental Sciences*. Section "Qualitative Weighted Variables", pages 56–57. url : https://cran.r-project.org/web/packages/ade4/ade4.pdf.