ÉCOLE NATIONALE D'INGÉNIEURS DE TUNIS

Département Génie Industriel

Projet Statistique

Elaboré par :

Lakhzouri Roukaya

Encadré par :

M Anissa RABHI

2eme Année MINDS

Année universitaire : 2023/2024

Table des matières

Table des figures

Contexte et objectifs du projet

En utilisant le jeu de données Dickcissel, nous comparerons l'importance relative du climat (**clTma**), de la productivité (**NDVI**) et de la couverture du sol (**grass**) comme prédicteurs de l'abondance de Dickcissels (**abund**). Le tableau intitulé "Dickcissel" se compose de 647 individus et de 15 colonnes, mais nous n'utiliserons que 4 variables qui sont :

— **abund** : Cette colonne représente l'abondance des Dickcissels [1]. Il s'agit de la variable que nous cherchons à prédire ou à expliquer à l'aide des autres variables.

— **clTma** : L'importance relative du climat, qui est une mesure spécifique du climat.

— **NDVI** : Il s'agit de l'indice de végétation par différence normalisée, utilisé en agriculture pour évaluer la vigueur et la quantité de végétation par analyse des mesures de télédétection.

— **grass** : La couverture du sol.

Ce projet vise à comprendre comment des variables telles que le climat, la productivité de la végétation et la couverture du sol influencent cette abondance, ce qui peut fournir des informations importantes pour la conservation de ces oiseaux et de leurs habitats.

1. l'abondance d'une espèce : quantité relative au nombre d'individus d'une espèce donnée par unité de surface ou de volume par rapport au nombre total d'individus de toutes espèces confondues

Chapitre I

Analyse descriptive

Introduction :

Dans ce chapitre, nous nous intéresserons à étudier dans un premier temps les statistiques descriptives pour chacune des variables (**clTma**), (**NDVI**), (**grass**), et (**abund**), afin de déterminer les tendances générales.

Importation des données

Tout d'abord, j'ai converti le fichier Excel en un fichier CSV que j'ai nommé "dickcissel.csv". Ensuite, j'ai ouvert une nouvelle feuille Excel et cliqué sur l'onglet "Données". J'ai sélectionné "À partir d'un fichier texte/csv" et choisi le fichier "dickcissel.csv". Après avoir cliqué sur "Importer", j'ai enregistré le fichier sous le nom "data.xlsx". Enfin, j'ai exécuté ce code dans le logiciel RStudio. J'ai utilisé cette méthode pour éviter d'obtenir une table comportant une seule colonne après l'exécution. J'ai également remarqué que toutes les colonnes étaient de type chaîne de caractères dans la dataframe importée depuis Excel. Pour remédier à cela, j'ai effectué une transformation numérique sur toutes les colonnes, à l'exception de la deuxième.

```
library(readxl)
data <- read_excel("C:\\Users\\user\\Desktop\\ds projet\\data.xlsx")
#Data qui contient toutes les colonnes de "data" sauf la deuxième
donnees_sans_2eme_colonne <- data[, -2]
# vecteur contenant les noms des colonnes sauf la deuxième
colonnes <- colnames(donnees_sans_2eme_colonne)

for (i in colonnes)
```

```
{
  data[[i]] <- as.numeric(data[[i]])
}
data <- as.data.frame(data)
```

descriptions de la dataframe

Pour obtenir une vue d'ensemble rapide de la dataframe en question, nous utilisons les commandes suivantes :

— `dim(data)` :Cette fonction affiche les dimensions de la base de données.

— `str(data)` :Cette fonction affiche la structure de la dataframe, montrant le type de données de chaque colonne et les premières valeurs.

— `summary(data)` Cette fonction fournit un résumé statistique pour chaque colonne numérique de la dataframe, affichant des statistiques telles que la moyenne, le minimum, le 1er quartile, la médiane, le 3e quartile et le maximum.

```
dim(data)
str(data)
summary(data)
```

Et voici les résultats obtenus :

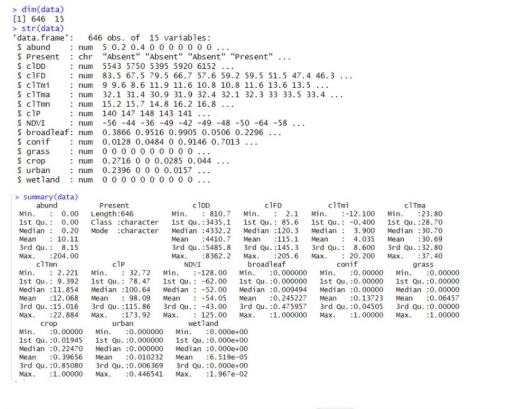

FIGURE I.1 – Description de la data

Visualisation de la dataframe

Histogrammes

Ensuite, j'ai créé un histogramme pour chacune des variables (clTma), (NDVI), (grass), et (abund) afin de visualiser la répartition les valeurs de ces variables. Le code est le suivant :

```
hist(data$abund, main="Histogramme de abund", xlab="abund")
hist(data$clTma, main="Histogramme de clTma", xlab="clTma")
hist(data$NDVI, main="Histogramme de NDVI", xlab="NDVI")
hist(data$grass, main="Histogramme de grass", xlab="grass")
```

Et voici les résultats obtenus :

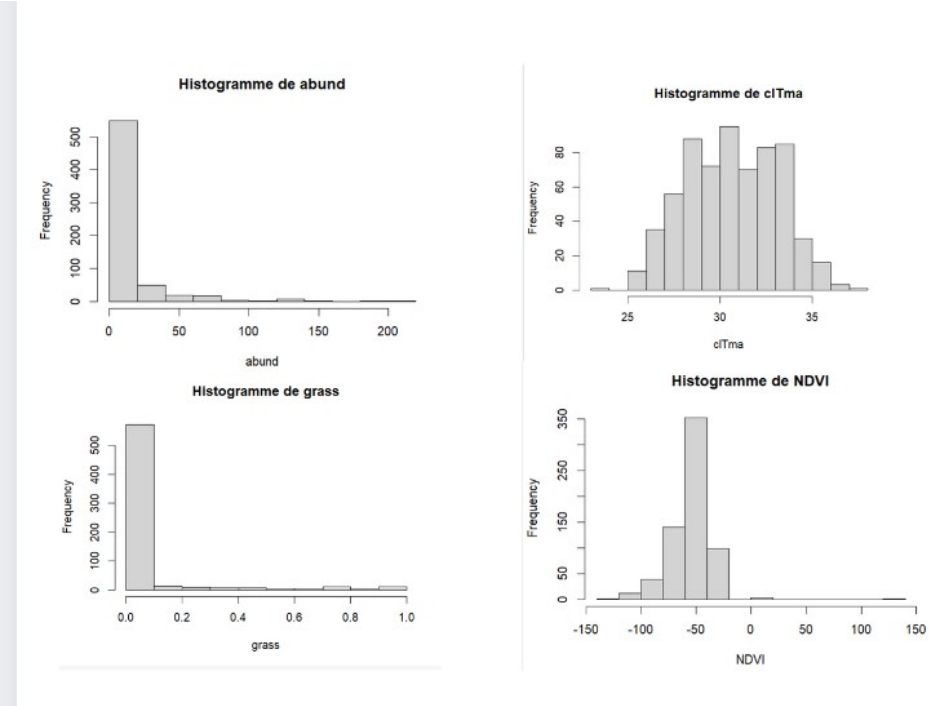

FIGURE I.2 – Histogrammes des variables

Boîte à moustaches

En appliquant la fonction "boxplot", nous créons un graphique en boîte pour chaque variable afin d'analyser la dispersion des données, notamment la médiane, les quartiles et les valeurs extrêmes.

```
boxplot(data$abund, main="Boite a moustaches de abund")
boxplot(data$clTma, main="Boite a moustaches de clTma")
boxplot(data$NDVI, main="Boite a moustaches de NDVI")
boxplot(data$grass, main="Boite a moustaches de grass")
```

Et voici les résultats obtenus :

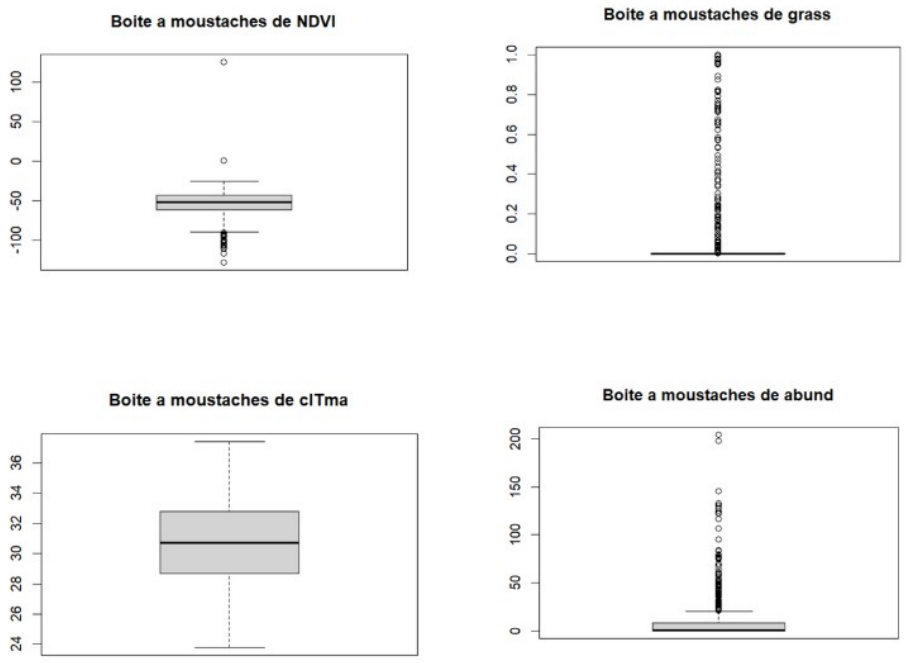

FIGURE I.3 – boxplot des variables

Interprétation :

La ligne à l'intérieur de la boîte représente la médiane, qui divise les données en deux parties. Pour les variables NDVI et CITma, la médiane est respectivement autour de -5 et 30. Les deux parties sont égales, ce qui signifie que la dispersion autour de la moyenne est presque égale pour ces deux variables. En revanche, pour les variables grass et abund, la majorité des valeurs sont concentrées du côté supérieur de la médiane, ce qui s'explique par le fait que la distribution est inclinée vers la gauche.

La relation entre les variables :

Nuages de points :

Pour étudier la relation entre les variables, j'ai choisi de tracer des nuages de points. tout d'abord j'ai utuilise la fonction paire comme indique le code si dessous :

```
# Créer un exemple de dataframe avec les variables x, y, z et w
data1 <- data.frame(
   data$abund, # Variable abund
  data$clTma, # Variable clTma
  data$NDVI, # Variable NDVI
  data$grass  # Variable grass
)

# Tracer les nuages de points entre les variables
pairs(data1)
```

Et voici les résultats obtenus :

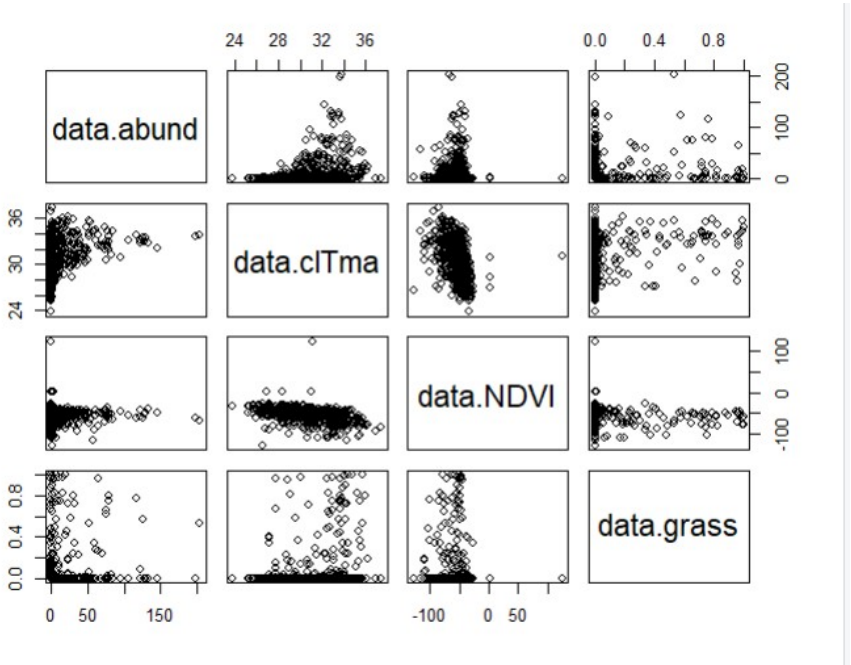

FIGURE I.4 – les nuages de points des variables

On constate le nuage des points entre (grass) et abund , ,cItima et abund , Ndvi et abund ,grass et CItma ne peut pas être ajusté à une fonction linéaire simple. Cela peut indiquer l'existence d'une relation non linéaire ou complexe entre ces variables. pour mieux identifier les reltion entre grass et NDVI ainsi que NDVI et CITma on trace le nuage de point en utilisnat ce script :

```
plot(data$NDVI, data$grass, xlab = "NDVI", ylab = "grass",
main = "Nuage de points NDVI vs grass")
plot(data$clTma, data$NDVI, xlab = "clTma", ylab = "NDVI",
main = "Nuage de points clTma vs NDVI")
```

Et voici les résultats obtenus :

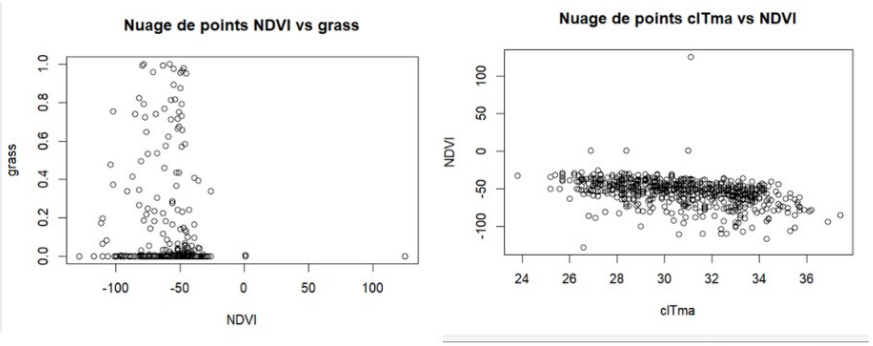

FIGURE I.5 – le nuage de points des variables grass et NDVI ainsi que NDVI et CITma

On constate que le nuage de points résultant de la comparaison entre ces variables a une apparence linéaire, ce qui pourrait indiquer une ces variables.

Matrice de corrélation

Après avoir étudié la distribution de nos données, nous avons poursuivi l'analyse en examinant la matrice de corrélation. Cette démarche vise à approfondir notre compréhension des relations entre les différentes variables en évaluant le degré de corrélation entre elles. Ceci nous permet de déterminer les variables qui sont fortement ou faiblement corrélées.
Et voici le code permettant de déterminer les corrélations entre les variables :

```
data_corr <- data[,c("abund", "grass", "NDVI", "clTma")]
cor(data_corr)
library(corrplot)
corrplot(cor(data_corr), type="upper", tl.col="black")
```

Et voici les résultats obtenus :

```
> cor(data_corr)
              abund       grass        NDVI       clTma
abund  1.00000000  0.1645269 -0.05186327  0.3193747
grass  0.16452690  1.0000000 -0.13934092  0.2748012
NDVI  -0.05186327 -0.1393409  1.00000000 -0.4195822
clTma  0.31937475  0.2748012 -0.41958221  1.0000000
```

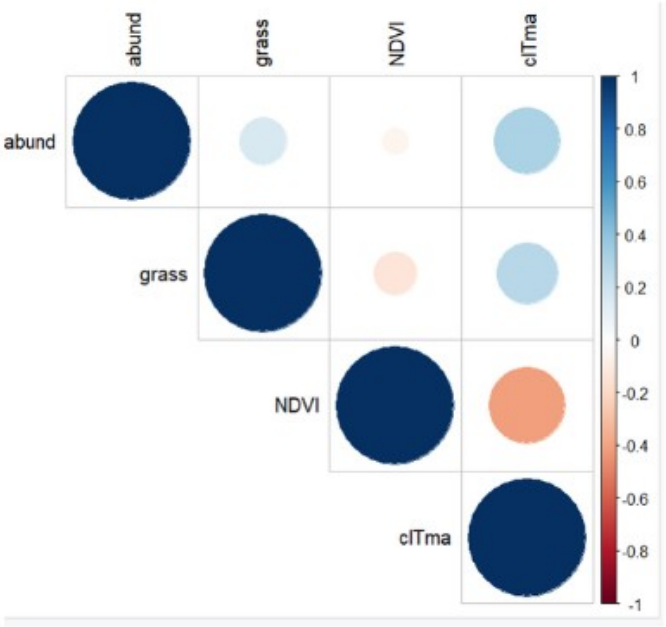

FIGURE I.6 – Représentation de la matrice de corrélation

— On remarque que la variable `abund` est bien corrélée avec la variable `CItma` (0.319) et la variable `grass` (0.16), mais la corrélation entre cette variable et `CItma` est très faible (-0.05).

— Pour la variable `grass`, on remarque qu'elle est bien corrélée avec les variables `CItma` (0.27) et `NDVI` (-0.139).

— La corrélation entre `NDVI` et `CItma` est considérée comme la meilleure valeur, atteignant -0.419.

Chapitre II

Analyse en composantes principales

Introduction

Dans ce chapitre, nous réaliserons une Analyse en Composantes Principales (ACP), une méthode statistique permettant d'analyser les correspondances entre les variables d'un jeu de données en réduisant leur dimensionnalité tout en préservant leur variabilité maximale. L'ACP nous permettra d'identifier les relations et les tendances les plus importantes entre les variables, facilitant ainsi la compréhension globale de notre ensemble de données.

Implication du code et résultats obtenus

```
# Extraire les variables importantes de notre data
data1 <- data[,c("abund", "grass", "NDVI", "clTma")]
# Réaliser une Analyse en Composantes Principales (ACP)
resultats_acp <- princomp(data1, cor = TRUE, scores = TRUE)
# Afficher un résumé des résultats de l'ACP
summary(resultats_acp)
```

Et voici les résultats obtenus :

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4
Standard deviation     1.310624 0.9800025 0.9123801 0.6995869
Proportion of Variance 0.429434 0.2401012 0.2081093 0.1223554
Cumulative Proportion  0.429434 0.6695352 0.8776446 1.0000000
```

FIGURE II.1 – un résumé des résultats de l'ACP

— La première composante principale (Comp.1) a un écart type de 1,3106 et explique 42,9% de la variance totale des données. La deuxième composante principale (Comp.2) a un écart type de 0,98 et explique 24,01% de la variance totale. La troisième composante principale a un écart type de 0,91 et explique 20,8% de la variance totale, tandis que la dernière composante principale a un écart type de 0,69 et explique 12,23% de la variance totale des données.

— En observant la proportion cumulée de la variance expliquée par chaque composante, on constate que tous les composantes principales jouent un rôle prépondérant dans la description des données.

La valeur propre de chaque composante

Maintenant, nous nous intéressons à la valeur propre de chaque composante pour mieux comprendre la proportion de la variance totale des données. Nous représentons le graphique scree plot afin de visualiser la contribution de chaque composante principale à la variance totale.

```
val.propres <- resultats_acp$sdev^2 # valeur propre = variance
print(val.propres)
plot(1:4,val.propres,type="b",ylab="Valeurs
propres",xlab="Composante",main="Scree plot")
```

Et voici les résultats obtenus :

```
> print(val.propres)
   Comp.1    Comp.2    Comp.3    Comp.4
1.7177359 0.9604049 0.8324374 0.4894218
```

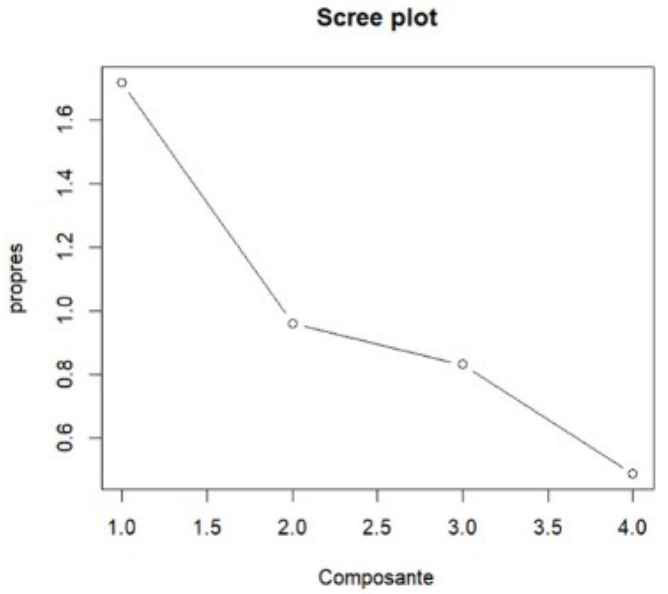

FIGURE II.2 – Scree plot

Cercle de corrélation

Nous stockons ces informations dans deux vecteurs, c1 et c2, représentant respectivement la contribution de chaque variable à la formation de la première et de la deuxième composante principale. Cela est obtenu en multipliant les chargements (loadings) de chaque variable sur cette composante par l'écart-type (standard deviation) de cette composante.

```
c1 <- resultats_acp$loadings[,1]*resultats_acp$sdev[1]
c2 <- resultats_acp$loadings[,2]*resultats_acp$sdev[2]
plot(c1,c2,xlim=c(-2,2),ylim=c(-1.5,+1.5))
abline(h=0,v=0)
text(c1,c2,labels=colnames(data1),cex=1)
symbols(0,0,circles=1,inches=F,add=T)
arrows(0,0,c1,c2,length=0.1,col = "red")
```

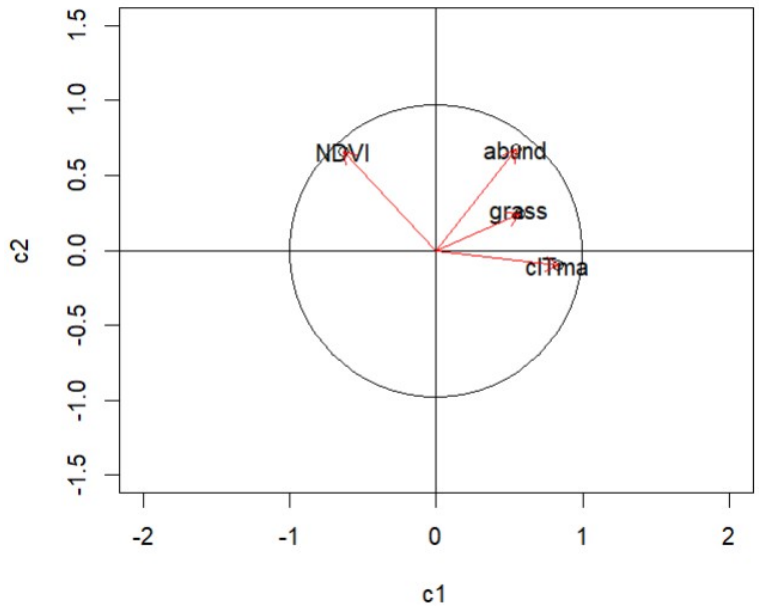

FIGURE II.3 – Cercle de corrélation

Pour le premier axe :

— La variable CITma a une corrélation positive et élevée avec le premier axe principal, car sa coordonnée est positive et proche de 1. Cela suggère que cette variable contribue fortement à la variance expliquée par cet axe.

— Les variables ‘abund‘, ‘NDVI‘ et ‘grass‘ ont une corrélation moins élevée avec le premier axe principal, car leurs coordonnées sont proches de 0.5, -0.5 et 0.5 respectivement. De plus, les variables ‘abund‘ et ‘grass‘ ont une coordonnée positive, suggérant ainsi une contribution positive à la variance expliquée par cet axe, tandis que ‘NDVI‘ a une coordonnée négative, suggérant une contribution négative à la variance expliquée par cet axe.

Pour le deuxième axe :

— Les variables ‘abund‘ et ‘NDVI‘ ont une contribution positive significative sur cet axe, de l’ordre de 0.8, ce qui signifie qu’elles sont corrélées avec cet axe et qu’elles jouent un rôle dans la variance expliquée par cet axe. Tandis que la variable ‘grass‘ a une contribution positive sur cet axe mais elle est moins importante, de l’ordre de 0.2.

— Les variables ‘CITma‘ ont des contributions très faibles sur cet axe, très proches de 0.

Chapitre III

Régression multiple

Introduction

Dans ce chapitre, nous effectuons une régression multiple de la variable abund sur les variables NDVI, grass et CITma dans le but d'étudier l'effet de chacune de ces variables sur la variable dépendante abund.

Implication du code et résultats obtenus

```
# Réaliser un modele de  Régression multiple
model=lm(abund~NDVI+grass+clTma,data=data1)
# Afficher un résumé des résultats du modèle Régression multiple
summary(model)
```

Et voici les résultats obtenus :

```
> summary(model)

Call:
lm(formula = abund ~ NDVI + grass + clTma, data = data1)

Residuals:
     Min       1Q   Median       3Q      Max
-35.327 -11.029  -4.337   2.150 180.725

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -83.60813   11.57745  -7.222 1.46e-12 ***
NDVI          0.13716    0.05486   2.500   0.0127 *
grass        10.41435    4.68962   2.221   0.0267 *
clTma         3.27299    0.40677   8.046 4.14e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.58 on 642 degrees of freedom
Multiple R-squared:  0.117,     Adjusted R-squared:  0.1128
F-statistic: 28.35 on 3 and 642 DF,  p-value: < 2.2e-16
```

FIGURE III.1 – un résumé des résultats de modèle de Regression multiples

— Les variables NDVI, grass et CITma sont significatives avec des valeurs de p-value inférieures à 0,05 (respectivement 0.0127, 0.0267 et 4.14×10^{-15}).

— Le coefficient de détermination Multiple R-squared quantifie la proportion de variance totale de la variable dépendante. Dans ce cas, les variables indépendantes expliquent environ 11.7%de la variation dans la variable dépendante, ce qui est considéré comme très faible. Par conséquent, le modèle de régression n'explique pas bien la variance des données.

— Le F-statistic évalue la pertinence globale du modèle de régression. Ici, il est de 28.35 avec une p-value inférieure à 2.2×10^{-16} ce qui indique que le modèle de régression est globalement significatif

Vérification des hypothèses de la régression

Indépendance des résidus :

```r
# Calcul des résidus du modèle
residus <- resid(model)

# Création d'un graphique de dispersion
plot(
  model$fitted.values,    # Valeurs prédites par le modèle (axe x)
  residus,                # Résidus du modèle (axe y)
  xlab = "Valeurs prédites",  # Étiquette de l'axe x
  ylab = "Résidus"             # Étiquette de l'axe y
)
```

Et voici les résultats obtenus :

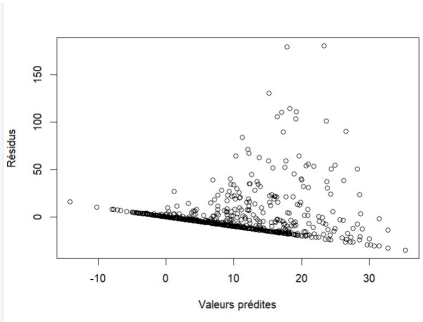

FIGURE III.2 – graphique de dispersion des résidus

Le graphique montre une structure particulière des résidus, avec une distribution linéaire. Cela indique que l'indépendance des résidus n'est pas vérifiée.

Normalité des résidus

```
# Tracer un histogramme des résidus
hist(residus, main = "Histogramme des residus")

# Tracer un graphique quantile-quantile (QQ plot)
qqnorm(residus)

# Ajouter une ligne droite au graphique quantile-quantile
qqline(residus)
```

Et voici les résultats obtenus :

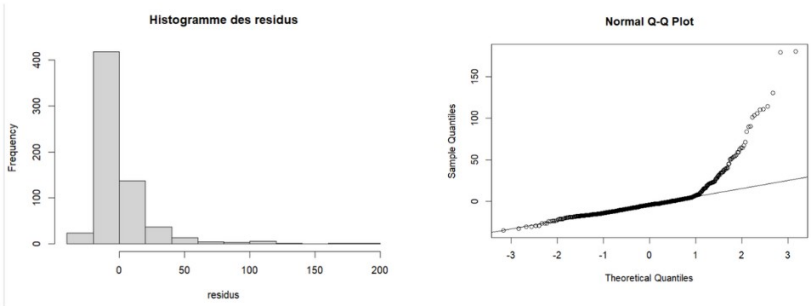

FIGURE III.3 – Verification de normalité de résidus

Les points du QQ-plot montrent une certaine proximité avec la ligne diagonale au début, mais par la suite, ils s'en éloignent. Ce comportement suggère que la distribution des résidus pourrait ne pas être normale, même si l'histogramme ne correspond pas exactement à la forme attendue pour cette distribution. Pour confirmer cette observation, nous appliquons le test de Shapiro-Wilk.

```
#appliquer le test de Shapiro-Wilk.
shapiro.test(residus)
```

```
> shapiro.test(residus)

        Shapiro-Wilk normality test

data:  residus
W = 0.66102, p-value < 2.2e-16
```

FIGURE III.4 – Résultat du test de Shapiro-Wilk

La p-value étant inférieure à 0,05, nous rejetons l'hypothèse nulle (H0 : les résidus suivent une loi normale) au profit de l'hypothèse alternative (H1 : les résidus ne suivent pas une loi normale) avec un niveau de confiance de 95%. Par conséquent, nous concluons que les résidus ne suivent pas une loi normale.

Nous concluons donc que les conditions d'application ne sont pas vérifiées.

Comparaison entre l'Analyse en Composantes Principales (ACP) et le Modèle de Régression Multiple

Après une analyse approfondie, j'ai relevé plusieurs lacunes dans notre modèle de régression linéaire. Les résidus ne présentent pas une indépendance satisfaisante, et leur distribution ne correspond pas à une distribution normale. De plus, le coefficient de détermination (R-squared) est d'environ 0.1, ce qui suggère que notre modèle ne parvient pas à expliquer une part importante de la variation observée dans les données. Ces constatations remettent en question l'adéquation du modèle linéaire pour notre ensemble de données.

En conséquence, le modèle de régression multiple ne semble pas être le choix le plus approprié pour nos données. À la lumière de ces résultats, le modèle le plus adéquat semble être l'Analyse en Composantes Principales (ACP), qui a démontré des performances supérieures à celles du modèle de régression multiple dans notre contexte spécifique.

Autres models :

Si l'on observe la figure (I.4), qui représente les nuages de points entre les variables, on constate que toutes les variables ne présentent pas de relations linéaires claires avec la variable abund. Cette observation suggère que le modèle de régression linéaire pourrait ne pas avoir bien performé, car il ne capture pas les relations complexes entre les variables. Pour tester de telles relations non linéaires, une méthode courante est d'utiliser des modèles de régression non linéaire, tels que

les modèles polynomiaux. Ces modèles peuvent mieux s'adapter aux structures non linéaires des données et fournir une meilleure compréhension des relations entre les variables. et voici l'implimentation du ce modele

modelé de polynôme de degré 2

```
# Création du modèle polynomial
model_polynomial <- lm(abund ~ poly(NDVI, 2) * poly(grass, 2)
* poly(clTma, 2), data = data1)

# Affichage des résultats du modèle
summary(model_polynomial)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-45.223  -9.613  -2.602   1.858 182.522

Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                    6.067e+00  2.618e+00   2.318 0.020778 *
poly(NDVI, 2)1                                 1.059e+02  5.898e+01   1.795 0.073116 .
poly(NDVI, 2)2                                -5.672e+02  2.661e+02  -2.132 0.033393 *
poly(grass, 2)1                               -5.662e+02  2.394e+02  -2.365 0.018351 *
poly(grass, 2)2                               -2.168e+02  9.755e+01  -2.222 0.026611 *
poly(clTma, 2)1                                4.393e+02  6.834e+01   6.428 2.59e-10 ***
poly(clTma, 2)2                               -7.285e+01  4.378e+01  -1.664 0.096648 .
poly(NDVI, 2)1:poly(grass, 2)1                -6.052e+03  4.750e+03  -1.274 0.203095
poly(NDVI, 2)2:poly(grass, 2)1                -6.254e+04  2.608e+04  -2.398 0.016790 *
poly(NDVI, 2)1:poly(grass, 2)2                -1.160e+03  2.549e+03  -0.455 0.649155
poly(NDVI, 2)2:poly(grass, 2)2                -1.982e+04  1.022e+04  -1.938 0.053028 .
poly(NDVI, 2)1:poly(clTma, 2)1                 5.487e+03  1.824e+03   3.009 0.002726 **
poly(NDVI, 2)2:poly(clTma, 2)1                 1.887e+04  6.920e+03   2.726 0.006583 **
poly(NDVI, 2)1:poly(clTma, 2)2                 1.832e+02  1.654e+03   0.111 0.911831
poly(NDVI, 2)2:poly(clTma, 2)2                -4.716e+03  3.866e+03  -1.220 0.222957
poly(grass, 2)1:poly(clTma, 2)1                1.528e+04  6.065e+03   2.520 0.011999 *
poly(grass, 2)2:poly(clTma, 2)1                4.474e+03  2.545e+03   1.758 0.079211 .
poly(grass, 2)1:poly(clTma, 2)2               -1.328e+04  3.607e+03  -3.681 0.000252 ***
poly(grass, 2)2:poly(clTma, 2)2               -1.487e+03  1.976e+03  -0.752 0.452094
poly(NDVI, 2)1:poly(grass, 2)1:poly(clTma, 2)1  9.238e+04  1.434e+05   0.644 0.519748
poly(NDVI, 2)2:poly(grass, 2)1:poly(clTma, 2)1  1.442e+06  6.709e+05   2.149 0.032003 *
poly(NDVI, 2)1:poly(grass, 2)2:poly(clTma, 2)1  2.301e+05  7.955e+04   2.893 0.003948 **
poly(NDVI, 2)2:poly(grass, 2)2:poly(clTma, 2)1  7.136e+05  2.802e+05   2.547 0.011116 *
poly(NDVI, 2)1:poly(grass, 2)1:poly(clTma, 2)2 -3.578e+05  1.533e+05  -2.334 0.019906 *
poly(NDVI, 2)2:poly(grass, 2)1:poly(clTma, 2)2 -9.941e+05  3.874e+05  -2.566 0.010522 *
poly(NDVI, 2)1:poly(grass, 2)2:poly(clTma, 2)2 -1.543e+05  8.681e+04  -1.777 0.075985 .
poly(NDVI, 2)2:poly(grass, 2)2:poly(clTma, 2)2 -3.876e+05  1.888e+05  -2.053 0.040532 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.69 on 619 degrees of freedom
Multiple R-squared:  0.2144,    Adjusted R-squared:  0.1814
F-statistic: 6.496 on 26 and 619 DF,  p-value: < 2.2e-16
```

FIGURE III.5 – Résultat du modèle polynomiale de degré 2

— Le coefficient de détermination Multiple R-squared quantifie la proportion de variance totale de la variable dépendante. Dans ce cas, les variables indépendantes expliquent environ 18.14% de la variation dans la variable dépendante, ce qui est considéré comme faible mais meilleur que le modèle de régression multiples. De plus, la valeur de R-square a augmenté jusqu'à 23%.

— Le F-statistic évalue la pertinence globale du modèle de régression. Ici, il est de 6.49 avec une p-value de l'ordre de 10^{-6}, ce qui signifie que les variables indépendantes ont une influence significative sur la variable dépendante.

modelé de polynôme de degré 4

```r
# Création du modèle polynomial
model_polynomial <- lm(abund ~ poly(NDVI, 4) * poly(grass, 4)
* poly(clTma, 4), data = data1)

# Affichage des résultats du modèle
summary(model_polynomial)
```

```
Residual standard error: 19.86 on 521 degrees of freedom
Multiple R-squared:  0.4455,    Adjusted R-squared:  0.3136
F-statistic: 3.376 on 124 and 521 DF,  p-value: < 2.2e-16

>
```

FIGURE III.6 – Résultat du modèle polynomiale de degré 4

— Le coefficient de détermination Multiple R-squared quantifie la proportion de variance totale de la variable dépendante. Dans ce cas, les variables indépendantes expliquent environ 31.36% de la variation dans la variable dépendante, ce qui est considéré comme faible mais meilleur que les modèles précédents. De plus, la valeur de R-square a augmenté jusqu'à 44.5%.

— Le F-statistic évalue la pertinence globale du modèle de régression. Ici, il est de 3.376 avec une p-value de l'ordre de 10^{-6}, ce qui signifie que les variables indépendantes ont une influence significative sur la variable dépendante.

Conclusion

Notre projet a porté sur l'étude des Dickcissels, où nous avons comparé l'impact relatif du climat (**clTma**), de la productivité (**NDVI**) et de la couverture du sol (**grass**) en tant que prédicteurs de l'abondance des Dickcissels (**abund**), à l'aide du logiciel R. Nous avons entamé notre analyse en examinant les statistiques descriptives de chaque variable pour saisir les tendances générales, en appuyant nos conclusions avec des graphiques pour une visualisation plus claire.

Ensuite, nous avons procédé à une analyse en composantes principales (ACP) afin de mettre en lumière les corrélations entre les différentes variables. Par la suite, nous avons réalisé une régression multiple, sélectionnant une variable à expliquer tout en fournissant des interprétations pour chaque méthode utilisée. Enfin, nous avons introduit deux modèles polynomiaux qui ont offert des performances supérieures à la régression multiple, en raison de la nature non linéaire des relations présentes dans les données.