Species Distribution Modelling

Louis Cueff, Elias Hermance, Emeline Kerreneur - Master BMC

Présentation du projet

Le but de ce projet est tout d'abord de modéliser la distribution actuelle d'une ou plusieurs espèce(s) marine(s) en fonction de différents paramètres environnementaux comme la température ou la salinité. Ce modèle sera ensuite utilisé pour prédire leur distribution future dans des conditions de changement climatique.

Notre choix se porte sur l'espèce d'algues brunes *Saccorhiza polyschides*, localisée dans les mers bordant les pays de l'Europe de l'Ouest (Manche, Mer Baltique, Mer Celtique et façade Atlantique). De ce fait, on se concentre également sur des profondeurs comprises entre 0 et 50m au vu de la zone d'habitabilité de cette espèce.

Pour la modélisation de la distribution actuelle, on dispose de données récoltées entre 1995-2004 et 2005-2012.

Après avoir présenté l'algorithme de prédiction de distribution de populations d'espèces marines développé par Lasram et al, nous nous servirons de l'étude dont il provient comme base pour notre projet. Nous étudierons dans un premier temps la capacité de prédiction de différentes méthodes de machine learning à travers deux études de validation croisée sur nos jeux de données comportant la valeur de différents paramètres environnementaux et de présence ou absence d'individus à différente position de notre zone d'étude sur deux décennies (1995-2004 et 2005-2012). Cela nous permettra de déterminer laquelle apparaît comme la plus efficace et la meilleure dans la prédiction de présence ou d'absence d'un individu à une position précise en fonction des paramètres environnementaux, et ainsi dans un second temps de prédire, en utilisant cette méthode de machine learning, l'évolution de la distribution des populations *Saccorhiza polyschides* sur les périodes 2040-2050 et 2090-2100 selon deux scénarios de réchauffement climatique (RCP-2.6 et RCP-8.5).

Partie 1 : Description de l'algorithme de prédiction de distribution locale d'une population SDM (*Lasram et al*, 2020)

Dans cette première partie, nous allons présenter l'algorithme SDM développé dans l'article **An open-source framework to model present and future marine species distributions at local scale** de Lasram et al. paru dans *Ecological Informatics* en 2020, sur lequel nous nous baserons en partie par la suite.

Cet algorithme est basé sur cinq étapes majeures :

- réaliser un pré-traitement des données de présence pour éviter des biais liés à l'échantillonnage : l'idée consiste à appliquer une grille spécifique basée sur des gradients de température et de salinité puis de sélectionner une seule présence par case de la grille
- générer le même nombre de pseudo-absences de façon aléatoire en dehors des zones de présence après avoir précédemment éliminé les conditions environnementales les plus extrêmes
- utilisation de filtres :
  - à l'échelle globale grâce aux BEMs (Bioclimatic Enveloppe Models) : L'idée est de délimiter une enveloppe bioclimatique basée sur les paramètres de température et de salinité ainsi que les données de présence et de pseudo-absences. Pour ce faire, le package biomod combine 8 techniques de modélisation (dont les régressions multiples, les arbres de régression et les analyses discriminantes) et la validation croisée à 3 blocs. Ensuite, des cartes d'aptitudes (suitability maps) sont générées par la moyenne des 24 prédictions pondérées par le CBI (Index de Boyce) et transformées en cartes binaires (présence / pseudo-absence) pour définir une zone d'habitabilité. Les meilleurs modèles sont sélectionnés pour un CBI > 0.5 pour de futures prédictions
  - à l'échelle locale grâce aux HMs (Habitat Models) : Utilisés pour supprimer les corrélations existantes entre les différentes variables, ils sont construits de façon similaires aux BEMs
- considérer les paramètres environnementaux dans un contexte 3D (température, salinité et profondeur) en combinant BEMs et HMs
- visualiser les observations et les prédictions sur des cartes

Partie 2 - Etude de l'efficacité de prédiction de la distribution de Saccorhiza polyschides par différentes méthodes de machine learning sur la même décennie (1995-2004)

Dans cette seconde partie, nous allons comparer différents algorithmes de machine learning afin de modéliser la distribution actuelle de *Saccorhiza polyschides*.

Pour ce faire, on génère tout d'abord un tableau de valeurs comportant les présences, les pseudo-absences, les différentes variables environnementales, au niveau de la zone d'étude (latitude 10 Nord à 73 Nord et longitude 5 Est à 15 Ouest).

Afin de choisir cette zone à étudier, nous avons simplement tracé un polygone au niveau de la région d'intérêt sur Obis et nous avons récupéré les coordonnées correspondant.

Puis nous avons sélectionné nos données environnementales en fonction d'elles, en choisissant de considérer les moyennes de salinité et température sur les périodes 1995-2004 et 2005-2012. Afin d'affiner nos modèles de prédiction nous avons également pris en compte, sur ces mêmes périodes, plusieurs autres variables environnementales qui influent toutes sur le développement des algues mais également des plantes en général dont la concentration en Nitrate et en Phosphate dans l'eau. Nous considérons également la quantité de dioxygène dissous dans l'eau (DO). Bien que notre modèle soit une macroalgue, c'est-à-dire un producteur primaire qui absorbe le CO2 et produit de l'O2, une variation de la concentration de DO dans l'eau pourrait avoir des effets non modestes sur son développement. Ceci est notamment visible au niveau des forêts de laminaires ou des variations de cette concentration sont visibles entre les différentes strates qui les constituent, influençant de manière significative et différente le développement des algues de chaque niveau. Même si aucune recherche actuellement n'a étudié l'influence de la désoxygénation des océans, ou hypoxie, sur le développement et la survie de cette espèce (et d'autres espèces de macroalgues), ce processus est intéressant à prendre en considération dans notre étude comme le suggère un récent rapport de l'Union internationale pour la conservation de la nature (IUCN) (Ocean deoxygenation: Everyone's problem. Causes, impacts, consequences and solutions D. Laffoley and J.M. Baxter, 2019). Des conditions d'hypoxie ou de faible concentration de DO pourraient notamment impacter les processus de gamétogenèse, d'implantation et de survie des jeunes algues après cette implantation, qui représentent des éléments clés dans l'étude de la projection de la distribution d'une espèce.

Afin de générer les pseudo-absences, nous allons modifier le tableau des présences. Le but est d'élargir la zone de présence à une petite région autour (+/- 0.25°) et de considérer l'extérieur de l'ensemble de ces carrés comme la zone de pseudo-absence.

Pour déterminer le meilleur modèle de prédiction de la distribution des algues étudiées, nous réalisons dans un premier temps une expérience de prédiction de la distribution des algues *Saccorhiza polyschides* par validation croisée sur la même décennie où nous prédisons un échantillon des données de 1995-2004 à partir du reste des données de cette période.

Dans un premier temps, nous allons d'abord construire le tableau de données.

```
#Choix de l'espèce et de la profondeur associée à l'espèce
Species <- 'Saccorhiza polyschides'
Vertical_habitat <- 'Pelagic'

# Initialisation du tableau
data <- data.frame(Species=Species,Vertical_habitat)

# Téléchargement des données sur OBIS (Ocean Biodiversity Information Sys
tem)
```

```r
occOBISL <- robis::occurrence(scientificname = data[1, "Species"])

## Retrieved 1027 records of approximately 1027 (100%)

# Remplissage du tableau : dans le cas d'une présence, charger dans le ta
bleau les coordonnées, le nom de l'espèce, la provenance de la donnée (OB
IS), la description phylogénétique de l'espèce et enfin la date

if (dim(occOBISL)[1] > 0) {
  occOBISL <- cbind(occOBISL[, c("decimalLongitude", "decimalLatitude", "
species")],
                    rep("obis", nrow(occOBISL)),
                    occOBISL[, "eventDate"],
                    occOBISL[, c("phylum", "class", "order", "family", "g
enus")])
  names(occOBISL) <- c("longitude", "latitude", "name", "prov",
                       "date", "phylum", "class", "order", "family", "gen
us")
  occOBISL$prov <- as.character(occOBISL$prov)
  occOBISL$year <- as.numeric(substr(as.character(occOBISL$date), 1, 4))
}

#Sélection des données du tableau pour la période 1995-2004

P95_04 <- subset(occOBISL, year >= 1995 & year <= 2004)
P2012 <- subset(occOBISL, year > 2004)

# Définition des données de températures, salinités, taux de nitrates, ta
ux de phosphates, taux d'oxygène d’intérêt en fonction des profondeurs,
longitude et latitude

dataMLBTemp95 <- read.csv("woa13_95A4_t00mn01v2_Temp.csv",sep = ",", head
er = TRUE)
dataMLBTemp95 <- dataMLBTemp95[,1:13]

dataMLBTemp2012 <- read.csv("temp2012.csv",sep=",",header=TRUE)
dataMLBTemp2012 <- dataMLBTemp2012[,1:13]

dataNitrates <- read.csv("Nitrates.csv",sep=",",header=TRUE)
dataNitrates <- dataNitrates[,1:13]
dataPhosphate <- read.csv("Phosphate.csv",sep=",",header=TRUE)
dataPhosphate <- dataPhosphate[,1:13]

dataOxygen <- read.csv("oxygen.csv",sep=",",header=TRUE)
dataOxygen <- dataOxygen[,1:13]
```

```r
dataMLBSal95 <- read.csv("woa13_95A4_s00mn01v2_Sal.csv",sep=",",header=TR
UE)
dataMLBSal95 <- dataMLBSal95[,1:13]

dataMLBSal2012 <- read.csv("sal2012.csv",sep=",",header=TRUE)
dataMLBSal2012 <- dataMLBSal2012[,1:13]
y=1:nrow(dataMLBTemp95)
dataMLB<-NULL

#DataMLB ne nous sert qu'à faire un graphique Salinité/Température, afin
d'obtenir le plus de données possible et éviter les erreurs

for (i in y){
  dataMLB$MoyenneTemp[i]=rowMeans(dataMLBTemp95[i,3:13],na.rm=TRUE)
  dataMLB$MoyenneSal[i]=rowMeans(dataMLBSal95[i,3:13],na.rm=TRUE)
}

dataMLB=as.data.frame(dataMLB)
dataMLB=na.omit(dataMLB)

dataMLBTemp95 <- subset(dataMLBTemp95, LATITUDE <= 73
                        & LATITUDE >= 10 & LONGITUDE <= 5 & LONGITUDE >=
-15)
dataMLBSal95 <- subset(dataMLBSal95, LATITUDE <= 73
                       & LATITUDE >= 10 & LONGITUDE <= 5 & LONGITUDE >= -
15)
dataMLBTemp2012 <- subset(dataMLBTemp2012, LATITUDE <= 73
                          & LATITUDE >= 10 & LONGITUDE <= 5 & LONGITUDE >
= -15)
dataMLBSal2012 <- subset(dataMLBSal2012, LATITUDE <= 73
                         & LATITUDE >= 10 & LONGITUDE <= 5 & LONGITUDE >=
-15)
dataNitrates <- subset(dataNitrates, LATITUDE <= 73
                       & LATITUDE >= 10 & LONGITUDE <= 5 & LONGITUDE >= -
15)
dataPhosphate <- subset(dataPhosphate, LATITUDE <= 73
                        & LATITUDE >= 10 & LONGITUDE <= 5 & LONGITUDE >=
-15)
dataOxygen <- subset(dataOxygen, LATITUDE <= 73
                     & LATITUDE >= 10 & LONGITUDE <= 5 & LONGITUDE >= -15
)

# Définition des présences en fonction de la zone définie

EspecePresence95_04 <- P95_04[,c(1,2)]
```

```r
colnames(EspecePresence95_04) <- c("Longitude","Latitude")
EspecePresence95_04 <- subset(EspecePresence95_04, Latitude <= 73 &
                              Latitude >= 10 & Longitude <= 5 & Longitude >
= -15)

EspecePresence2012 <- P2012[,c(1,2)]
colnames(EspecePresence2012) <- c("Longitude","Latitude")
EspecePresence2012 <- subset(EspecePresence2012, Latitude <= 73 & Latitud
e >= 10
                             & Longitude <= 5 & Longitude >= -15)
###### Appliquer la même restriction de zone (au cas où on ait des "hors
zone")

# Définition des pseudo-absences

#On recrée un tableau de données, dans lequel on viendra piocher les donn
ées de présence et de pseudo absence

donnees=dataMLBTemp95
donnees2012=dataMLBTemp2012

# Définition du nombre de température (nombre de lignes)
x <- 1:nrow(donnees)

# Définition de count : permet de déterminer le nombre final de présence
count <- 0

# Initialisation de la colonne présence / absence
donnees$PresAbs <- "A"

# Initialisation d'une variance à 1 dans le but de délimiter une zone de
présence
variance <- 0.25

#Pour toutes les coordonnées, associer une absence tant qu'on est en deho
rs de la zone de présence. La consigne  j!= nrow(EspecePresence95_04)+1
permet de terminer la boucle lorsque la valeur de j correspond à la derni
ère ligne du tableau des présences. Quand on se situe dans la zone de pré
sence, on associe une présence.

for (i in x){
  j = 1
  k <- EspecePresence95_04[j,2]+variance
  l <- EspecePresence95_04[j,2]-variance
  m <- EspecePresence95_04[j,1]+variance
  n <- EspecePresence95_04[j,1]-variance
```

```
    while (((k < dataMLBTemp95[i,1]) | (dataMLBTemp95[i,1] < l)) &
           ((m < dataMLBTemp95[i,2]) | (dataMLBTemp95[i,2] < n)) &
           j != nrow(EspecePresence95_04)+1){
      j=j+1
      k <- EspecePresence95_04[j,2]+variance
      l <- EspecePresence95_04[j,2]-variance
      m <- EspecePresence95_04[j,1]+variance
      n <- EspecePresence95_04[j,2]-variance
      donnees$PresAbs[i] = "A"
    }
  if (j==nrow(EspecePresence95_04)+1){
    donnees$PresAbs[i] = "A"
  }
  else{
    donnees$PresAbs[i] = "P"
    count=count+1
  }
}

# Définition d'un tableau afin de retrouver les données facilement sans
répéter la boucle à chaque fois
donneesbackup=donnees

#Affichage du nombre de données présence sur le nombre total de données :
on pourra faire changer la donnée "variance" afin de modifier la zone dél
imitant la présence
cat(count ,"/", nrow(donnees))

## 2189 / 7383

#Pour toutes les coordonnées, ajouter les moyennes de salinité et tempéra
ture sur les profondeurs de 0 à 50m

for (i in x){
  donnees$MoyenneTemp[i]=rowMeans(dataMLBTemp95[i,3:13],na.rm=TRUE)
  donnees$MoyenneSal[i]=rowMeans(dataMLBSal95[i,3:13],na.rm=TRUE)
}

f=1:nrow(donnees2012)
for (i in f){
  donnees2012$MoyenneTemp[i]=rowMeans(dataMLBTemp2012[i,3:13],na.rm=TRUE)
  donnees2012$MoyenneSal[i]=rowMeans(dataMLBSal2012[i,3:13],na.rm=TRUE)
}

#On ajoute dans un nouveau tableau les données de phosphate, nitrates, ox
ygène puisqu'elles ne varient pas entre 1995 et 2012
```

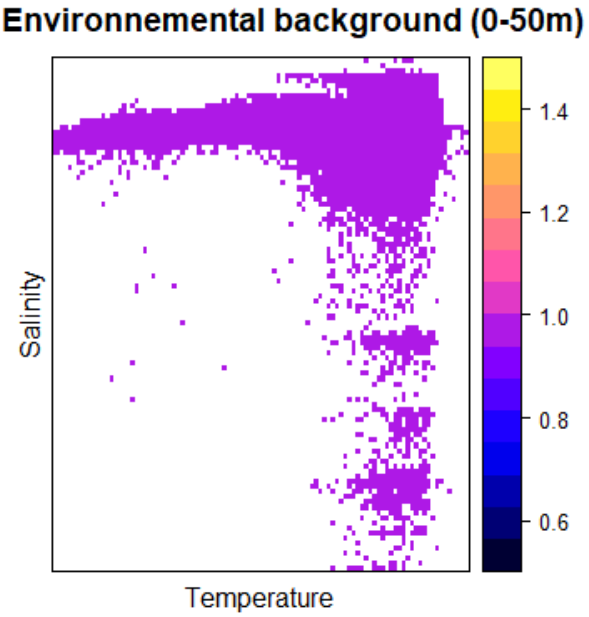
```r
NitPhos=dataNitrates
w=1:nrow(dataNitrates)

for (i in w){
NitPhos$MoyennePhosph[i]=rowMeans(dataPhosphate[i,3:13],na.rm=TRUE)
NitPhos$MoyenneNitrates[i]=rowMeans(dataNitrates[i,3:13],na.rm=TRUE)
NitPhos$MoyenneO2[i]=rowMeans(dataOxygen[i,3:13],na.rm=TRUE)
}

NitPhos=as.data.frame(NitPhos)

#On retire les lignes pour lesquelles on obtient des NA ainsi que les col
onnes qui ne nous intéressent plus pour ne garder que la valeur moyenne

NitPhos=na.omit(NitPhos)
NitPhos=NitPhos[,-(3:13)]

#Enlever les valeurs de température et de salinité de 0 à 50m pour ne gar
der que les valeurs moyennes puis enlever les NA
donnees=donnees[,-c(3:13)]
donnees=na.omit(donnees)

donnees2012=donnees2012[,-c(3:13)]
donnees2012=na.omit(donnees2012)

#On attribue pour les deux tableaux de données les valeurs de phosphate,
nitrates, oxygène en fonction de la longitude/latitude, puisque les valeu
rs n'ont pas été calculées selon les mêmes longitudes/latitudes

a=1:nrow(donnees)
var=0.5
donnees$MoyenneNit="Pas encore attribué"
donnees$MoyennePhosph="Pas encore attribué"
donnees$MoyenneO2="Pas encore attribué"

for (i in a){
  var=0.15
  b=1
  k <- NitPhos[b,2]+var
  l <- NitPhos[b,2]-var
  m <- NitPhos[b,1]+var
  n <- NitPhos[b,1]-var
  while (donnees$MoyenneNit[i]=="Pas encore attribué" &
         donnees$MoyennePhosph[i]=="Pas encore attribué" &
         donnees$MoyenneO2[i]=="Pas encore attribué")  {
    while (((m < donnees[i,1]) | (donnees[i,1] < n)) &
```

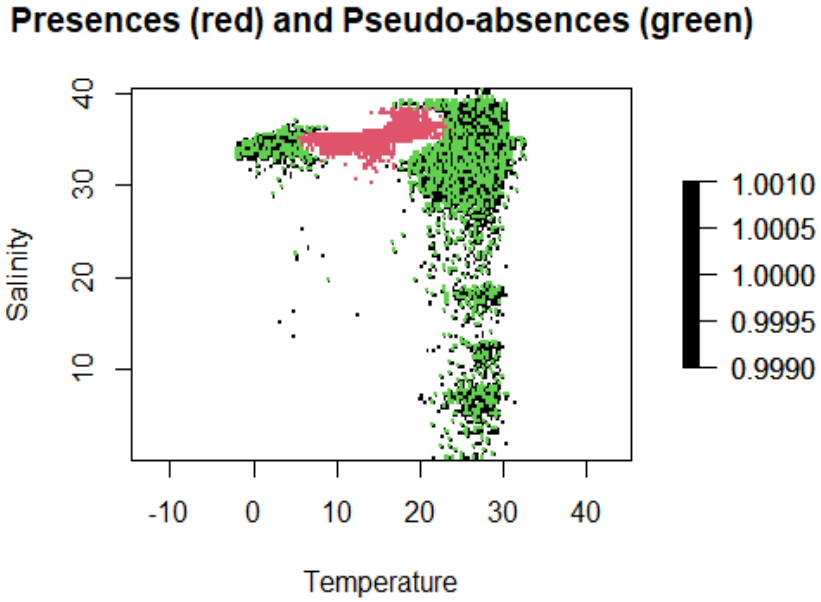
```
          ((k < donnees[i,2]) | (donnees[i,2] < l)) &
          b != nrow(NitPhos)+1){
      b=b+1
      k <- NitPhos[b,2]+var
      l <- NitPhos[b,2]-var
      m <- NitPhos[b,1]+var
      n <- NitPhos[b,1]-var
    }
    if (b==nrow(NitPhos)+1){
      donnees$MoyenneNit[i]="Pas encore attribué"
      donnees$MoyennePhosph[i]="Pas encore attribué"
      donnees$MoyenneO2[i]=="Pas encore attribué"
      var=var+0.15
      b=1
      k <- NitPhos[b,2]+var
      l <- NitPhos[b,2]-var
      m <- NitPhos[b,1]+var
      n <- NitPhos[b,1]-var
    }
  else{
    donnees$MoyenneNit[i]=round(NitPhos$MoyenneNitrates[b],digits=3)
    donnees$MoyennePhosph[i]=round(NitPhos$MoyennePhosph[b],digits=3)
    donnees$MoyenneO2[i]=round(NitPhos$MoyenneO2[b],digits=3)
    }
  }
}

b=1:nrow(donnees2012)
var=0.5
donnees2012$MoyenneNit="Pas encore attribué"
donnees2012$MoyennePhosph="Pas encore attribué"
donnees2012$MoyenneO2="Pas encore attribué"

for (i in b){
  var=0.15
  b=1
  k <- NitPhos[b,2]+var
  l <- NitPhos[b,2]-var
  m <- NitPhos[b,1]+var
  n <- NitPhos[b,1]-var
  while (donnees2012$MoyenneNit[i]=="Pas encore attribué" &
         donnees2012$MoyennePhosph[i]=="Pas encore attribué" &
         donnees2012$MoyenneO2[i]=="Pas encore attribué") {
    while (((m < donnees2012[i,1]) | (donnees2012[i,1] < n)) &
           ((k < donnees2012[i,2]) | (donnees2012[i,2] < l)) &
           b != nrow(NitPhos)+1){
```

```
      b=b+1
      k <- NitPhos[b,2]+var
      l <- NitPhos[b,2]-var
      m <- NitPhos[b,1]+var
      n <- NitPhos[b,1]-var
    }
    if (b==nrow(NitPhos)+1){
      donnees2012$MoyenneNit[i]="Pas encore attribué"
      donnees2012$MoyennePhosph[i]="Pas encore attribué"
      donnees2012$MoyenneO2[i]=="Pas encore attribué"
      var=var+0.15
      b=1
      k <- NitPhos[b,2]+var
      l <- NitPhos[b,2]-var
      m <- NitPhos[b,1]+var
      n <- NitPhos[b,1]-var
    }
    else{
      donnees2012$MoyenneNit[i]=round(NitPhos$MoyenneNitrates[b],digits=3
)
      donnees2012$MoyennePhosph[i]=round(NitPhos$MoyennePhosph[b],digits=
3)
      donnees2012$MoyenneO2[i]=round(NitPhos$MoyenneO2[b],digits=3)
    }
  }
}
```

Visualisation graphique des zones de présences et de pseudo-absences générées aléatoirement en dehors de la zone de présence en fonction des paramètres environnementaux. La méthode appliquée est celle de l'article de *Lasram et al*.

Les pseudo-absences (cellules noires) sont sélectionnées aléatoirement en dehors de la zone de présence (cellules roses) par le programme. Les présences à des positions où les valeurs de température et de salinité sont extrêmes (relativement aux valeurs moyennes) ne sont pas considérées et représentées par des cellules roses.





Nous allons modéliser nos données à partir de différents algorithmes supervisés de machine learning (explicité par la suite), attester et comparer la qualité de prédiction de nos données afin de choisir le meilleur modèle pour nos prédictions de distribution des individus durant la deuxième moitié du 21ème siècle.

Pour ce faire, nous allons dans un premier temps séparer nos données en deux jeux : un jeu de test et un jeu d'apprentissage.

```
#Pour ce faire, définir l'effectif du jeu de test à 2000 puis générer aut
ant de tirages aléatoires sans remise

nbr=2000
#Effectif jeu de test

choice=sample(1:nrow(donnees),size=nbr,replace=F)

#On ne conserve que les données sans les longitudes et les latitudes pour
ne pas les prendre en compte

donneesSLSL<- donnees[,-c(1,2)]
donneesSLSL2012<-donnees2012[,-c(1:2)]

# Définition des jeux de test et d'apprentissage à partir de ces tirages
donnees.train=donnees[-choice,c(3,4,5,6,7,8)]
donnees.test=donnees[choice,c(3,4,5,6,7,8)]
```

1)Algorithme des k plus proches voisins (knn)

Nous allons ensuite modéliser nos données selon l'algorithme des k plus proches voisins (knn) pour k allant de 1 à 9 voisins.

L'algorithme de Knn ou des K-plus-proches voisins est une méthode d'apprentissage supervisée permettant d'attribuer une classe à chaque point du jeu de données, dans ce cas présence ou absence. Pour classifier un point (coordonnée), il s'agit de déterminer les distances qui le séparent de chaque point du jeu de donnée puis de lui attribuer la classe la plus représentée par ses k plus proches voisins. Le choix du nombre de voisins k optimal se fait par validation croisée.

```
# Création d'un objet cl contenant les facteurs de classification pour le
jeu de données test : présence ou absence
cl=donnees[-choice,3]
#cl

#Pour simplifier l'utilisation de KNN nous décidons d'utiliser une boucle
BoucleKNN=NULL
```

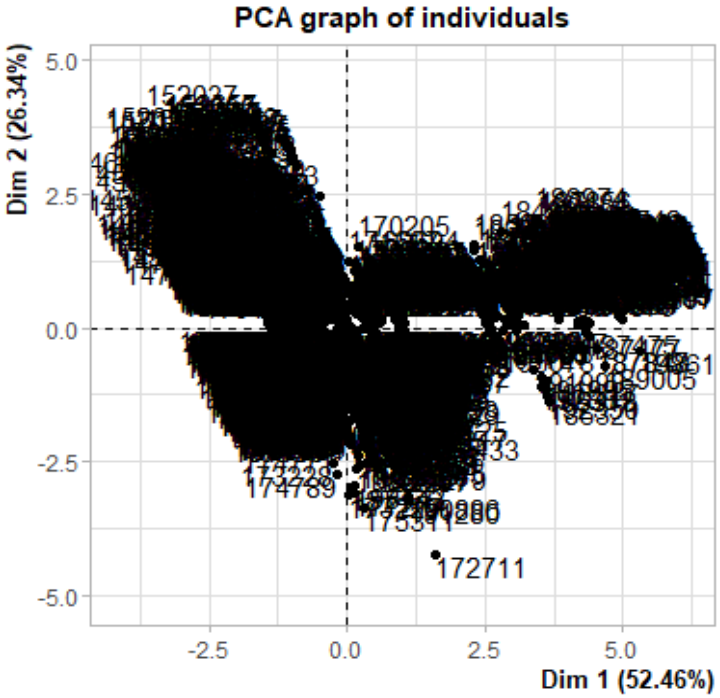
```
for (i in 1:9){
  n=i
  k=knn(donnees.train[,-1], donnees.test[,-1],cl,k=n,l=0,prob=FALSE,use.a
ll=TRUE)
  # Construction d'un tableau de classification
  tableau=table(k,donnees[choice,3])
  # Calcul du pourcentage de bonnes prédictions pour chaque voisin
  BoucleKNN$ratio[i] = ((tableau[1,1]+tableau[2,2])/nbr)*100
}

#Détermination du meilleur nombre de voisins et du pourcentage de bonnes
#prédictions associé
BoucleKNN=as.data.frame(BoucleKNN)
BoucleKNN=t(BoucleKNN)
Voisins=which.max(BoucleKNN)
cat("Le meilleur nombre de voisins est de ",Voisins,". ")

## Le meilleur nombre de voisins est de  7 .

Ratio=max(BoucleKNN)

cat("Le pourcentage de bonne prédiction du jeu train sur test est de ",Ra
tio,"%.")

## Le pourcentage de bonne prédiction du jeu train sur test est de  69.2
%.

# Définition d'un tableau de conclusion pour déterminer le meilleur des m
odèles
Conclusion=matrix(data          =          c(1,1,1),          ncol=1,nrow=3)
colnames(Conclusion)="TauxReussite"
Conclusion[1,1]=Ratio
Conclusion=as.data.frame(Conclusion)
```

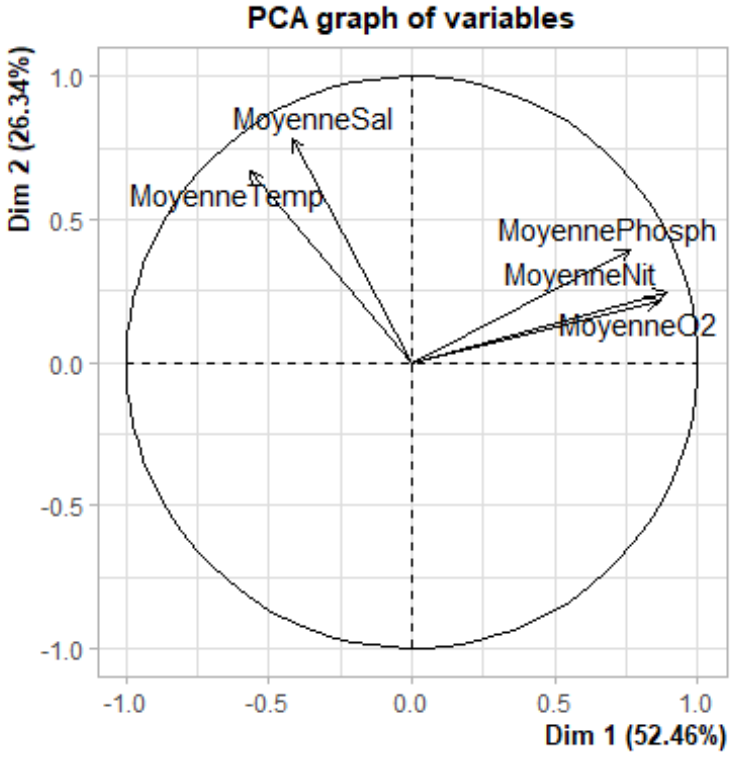
2) Analyses en composantes principales et modèle linéaire généralisé

L'analyse en composante principale est basée sur le calcul de la matrice de covariance pour permettre la réduction de la dimension tout en limitant le nombre d'informations perdues lors du passage d'un espace initial à grande dimension à un espace final à petite dimension. On associe à cette technique un modèle linéaire généralisé permettant la construction d'un modèle de régression linéaire reliant la variable présence / absence via l'estimation du maximum de vraisemblance des autres paramètres du modèle par la méthode des moindres carrés repondérés itérativement.

```
# Redéfinir les jeux pour pouvoir les modifier

donneesSLSL.train<- donnees.train
```

```r
donneesSLSL.test<- donnees.test[,-1]

donneesSLSL$MoyenneNit=as.numeric(donneesSLSL$MoyenneNit)
donneesSLSL$MoyenneO2=as.numeric(donneesSLSL$MoyenneO2)
donneesSLSL$MoyennePhosph=as.numeric(donneesSLSL$MoyennePhosph)
donneesSLSL$PresAbs=as.character(donneesSLSL$PresAbs)
donneesSLSL$PresAbs=as.character(donneesSLSL$PresAbs)

donneesSLSL.train$MoyenneNit=as.numeric(donneesSLSL.train$MoyenneNit)
donneesSLSL.train$MoyenneO2=as.numeric(donneesSLSL.train$MoyenneO2)
donneesSLSL.train$MoyennePhosph=as.numeric(donneesSLSL.train$MoyennePhosp
h)
donneesSLSL.train$PresAbs=as.character(donneesSLSL.train$PresAbs)

donneesSLSL.test$MoyenneNit=as.numeric(donneesSLSL.test$MoyenneNit)
donneesSLSL.test$MoyenneO2=as.numeric(donneesSLSL.test$MoyenneO2)
donneesSLSL.test$MoyennePhosph=as.numeric(donneesSLSL.test$MoyennePhosph)

# Réalisation de l'ACP sur les données train et test
res.pca.train=PCA(donneesSLSL.train[,-1], scale.unit = TRUE, ncp= 5, grap
h =TRUE)
```

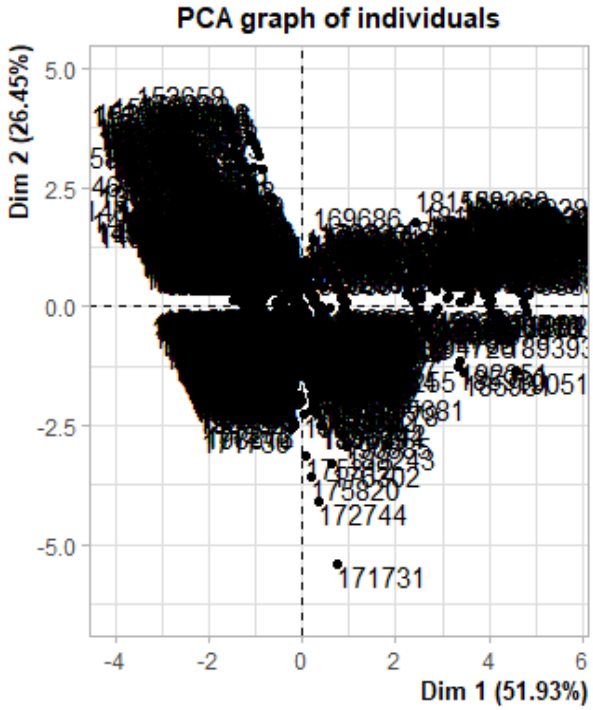


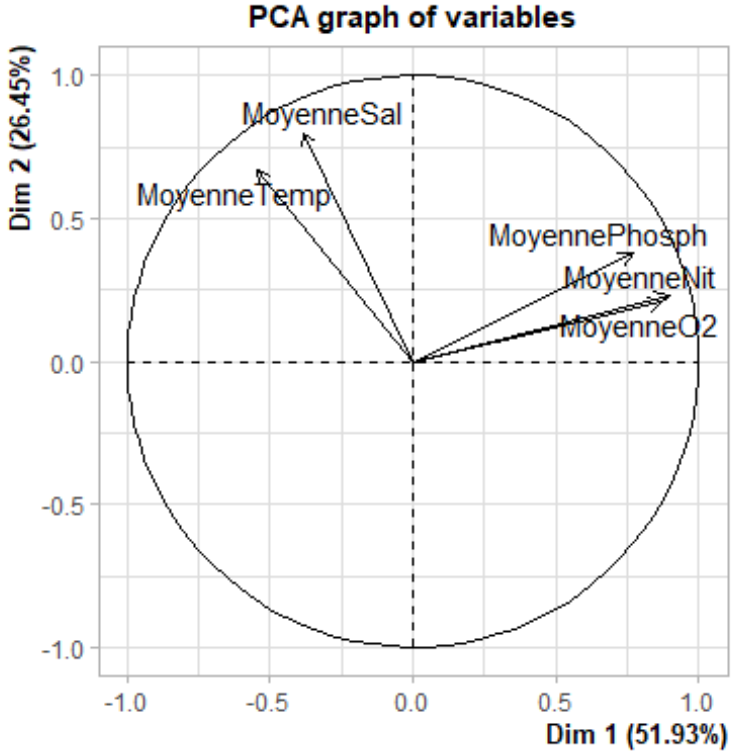


Les taux de phosphate, nitrate et dioxygène dissous dans l'eau semblent être des paramètres positivement corrélés entre eux tout comme la température et la salinité.

Cependant on remarque que ces deux dernières ne sont pas corrélées aux autres au vu de l'angle droit présent entre les flèches associées à ces deux groupes de variables. De plus on remarque que toutes les variables sont bien représentées par l'ACP au vu de la longueur importante de leur flèche.

```
x11()
res.pca.test=PCA(donneesSLSL.test, scale.unit = TRUE, ncp= 5, graph =TRUE
)
```






```
# Détermination de la distance des individus par rapport aux composantes
TrainPCA=as.data.frame(get_pca_ind(res.pca.train)$coord)
```

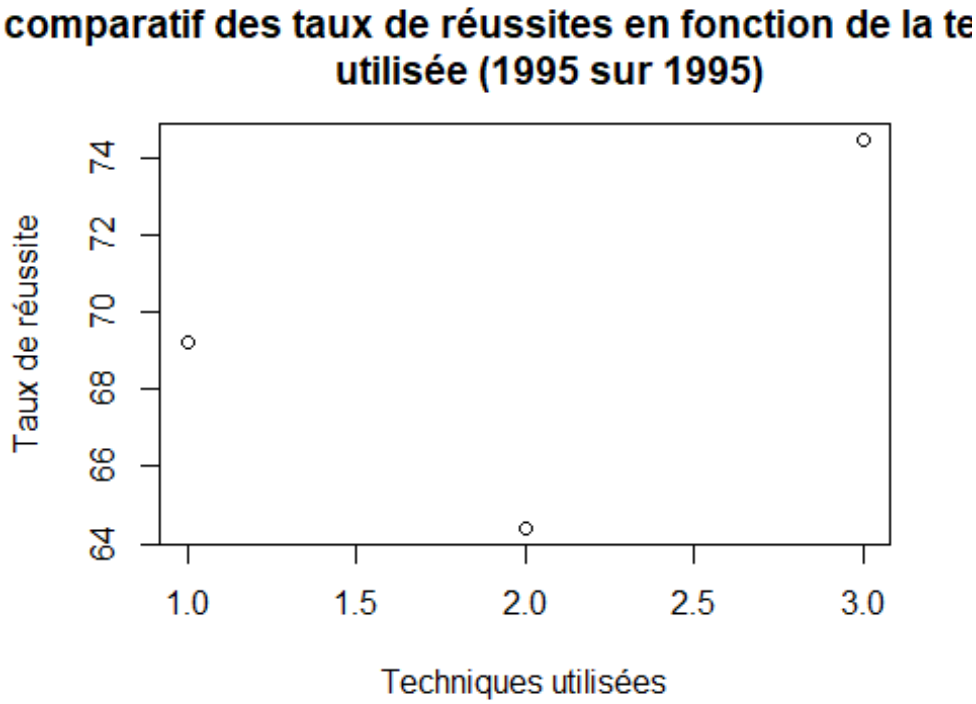
```r
TestPCA=as.data.frame(get_pca_ind(res.pca.test)$coord)

# Ajout des données présence / abscence
mod.train= cbind(donneesSLSL.train[1],TrainPCA[1:5])

mod.train$PresAbs=as.factor(mod.train$PresAbs)
mod.train$PresAbs=as.numeric(mod.train$PresAbs)-1

#Réalisation d'un modèle linéaire généralisé
TrainglmACP=glm(PresAbs ~ .,data=mod.train)
Test=as.data.frame(predict.glm(TrainglmACP, newdata=TestPCA,type="respons
e"))

# Prédiction des présences et des absences
t=1:nrow(TestPCA)
for (i in t){
  if (Test[i,1] > 0.5){
    TestPCA$PredictGLM[i]="P"
  }
  else{
    TestPCA$PredictGLM[i]="A"
  }
}

# Ajout de la colonne présence/absence
TestPCA$PresAbs=donnees.test[,1]

# Calcul du nombre de bonnes prédictions
p=1:nrow(TestPCA)
countPCAGLMtest=0
for (i in p) {
  a=TestPCA$PredictGLM[i]
  b=TestPCA$PresAbs[i]
  if (a==b){
    countPCAGLMtest=countPCAGLMtest+1
  }
  else{
    countPCAGLMtest=(((countPCAGLMtest+1)-1)/1)
  }
}

# Calcul du pourcentage de bonnes prédictions
ratioPCAGLMtest= (countPCAGLMtest/nrow(TestPCA))*100

cat("Le ratio prédiction pour PCAGLM est de ",ratioPCAGLMtest,"%")
```

```
## Le ratio prédiction pour PCAGLM est de  64.4 %

Conclusion[2,1]=ratioPCAGLMtest
```

3) Forêts aléatoires

La construction de forêts aléatoires nécessite un ensemble d'arbres de décision considérant des variables et des individus aléatoirement sélectionnés dans le jeu de données. Un arbre de décision est un algorithme supervisé visant à répartir les points dans des classes homogènes (présence / absence) à partir de variables discriminantes (température moyenne / salinité moyenne). Chaque nœud interne représente un test, chaque branche représente un résultat possible de ce test et chaque extrémité représente une classe possible. L'algorithme de forêts aléatoires permet la construction par entraînement d'un arbre de décision complet classant les individus dans des groupes (par le vote de chacun des arbres de décision).

```
#Création d'un objet fit qui regroupe les résultats du random forest
présence/absence des données95 sans longitude/latitude

donneesSLSL$PresAbs = factor(donneesSLSL$PresAbs)
fit = randomForest(PresAbs ~ ., donneesSLSL,mtry = 2)
print(fit)

##
## Call:
##  randomForest(formula = PresAbs ~ ., data = donneesSLSL, mtry = 2)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 25.52%
## Confusion matrix:
##      A    P class.error
## A 3156  604   0.1606383
## P  908 1257   0.4193995

Err=(fit$confusion)
Errorrate=((Err[1,1] + Err[1,2])* Err[1,3] + (Err[2,1]+Err[2,2])*Err[2,3]
)/(Err[1,1]+ Err[1,2]+Err[2,1]+Err[2,2])
Conclusion[3,1]=(1-Errorrate)*100
Conclusion=as.matrix(Conclusion)
rownames(Conclusion)=c("KNN","RF","GLMACP")
plot(Conclusion, xlab="Techniques utilisées",ylab="Taux de réussite",
     main="Plot comparatif des taux de réussites en fonction de la techni
que utilisée (1995 sur 1995)")
```



Pour ce qui est de la validation croisée, sur les trois techniques testées nous observons que la technique des random forest obtient un meilleur taux de réussite que les deux autres. Nous pensons donc prendre cette technique afin d'effectuer les prédictions sur 2050-2100. Afin de pouvoir confirmer ce choix, nous décidons d'effectuer une prédiction test sur un effectif de 2012 basée sur les effectifs de 1995, encore une fois en comparant ces trois techniques.

Partie 3 : Etude de l'efficacité de prédiction de la distribution de *Saccorhiza polyschides* par différentes méthodes de machine learning par validation croisée "temporelle"

Dans cette troisième partie, nous cherchons à sélectionner le meilleur des modèles parmi les plus intéressants obtenus dans la partie précédente.

Pour le déterminer, nous réalisons une expérience de prédiction de la distribution des algues *Saccorhiza polyschides* par validation croisée "temporelle" où nous prédisons les données de 2005-2012 à partir des données de 1995-2004.

1) Algorithme des k plus proches voisins (knn)

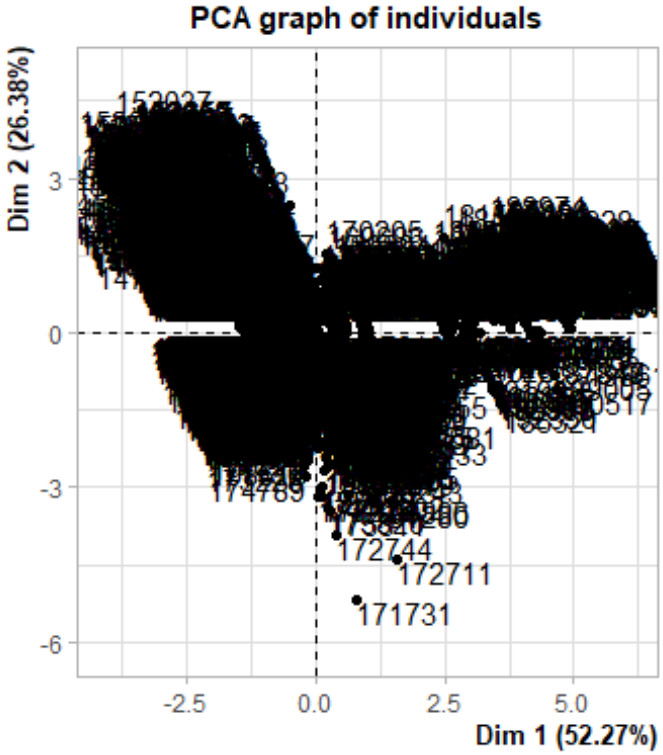
```
#Ajout d'une colonne présence/absence dans le tableau de données 2012 pou
r déterminer si la prédiction est juste
```

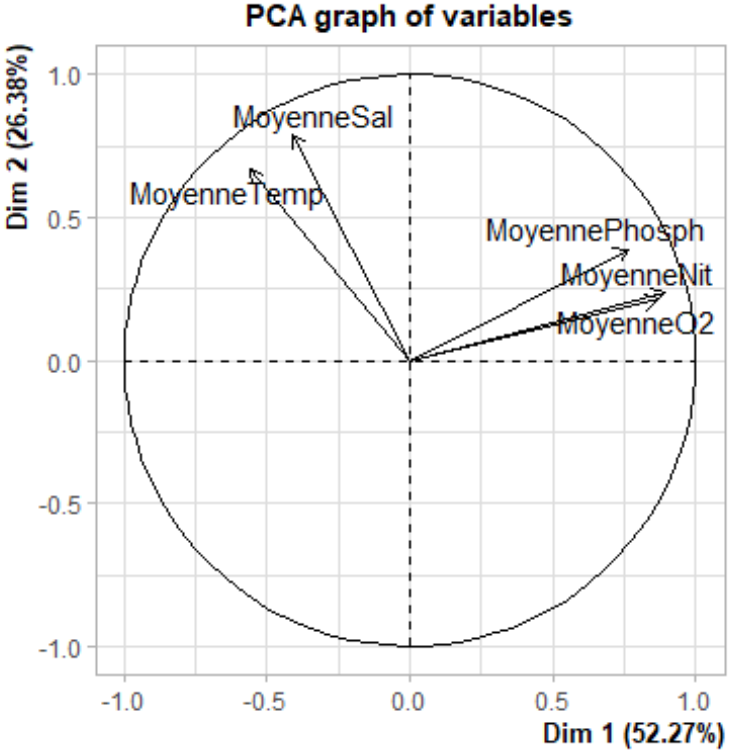
```r
x <- 1:nrow(donnees2012)
y <- 1:nrow(EspecePresence2012)
count <- 0
donnees2012$PresAbs <- "A"

variance <- 0.25

for (i in x){
  j = 1
  k <- EspecePresence2012[j,2]+variance
  l <- EspecePresence2012[j,2]-variance
  m <- EspecePresence2012[j,1]+variance
  n <- EspecePresence2012[j,1]-variance
  while (((k < dataMLBTemp2012[i,1]) | (dataMLBTemp2012[i,1] < l)) &
         ((m < dataMLBTemp2012[i,2]) | (dataMLBTemp2012[i,2] < n)) &
         j != nrow(EspecePresence2012)+1){
    j=j+1
    k <- EspecePresence2012[j,2]+variance
    l <- EspecePresence2012[j,2]-variance
    m <- EspecePresence2012[j,1]+variance
    n <- EspecePresence2012[j,1]-variance
    donnees2012$PresAbs[i] = "A"
  }
  if (j==nrow(EspecePresence2012)+1){
    donnees2012$PresAbs[i] = "A"
  }
  else{
    donnees2012$PresAbs[i] = "P"
    count=count+1
  }
}

#Cette colonne sera utilisée pour tous les tests de prédictions

#Test de prédiction avec Knn
#Sélection des paramètres en ne gardant que les colonnes oxygène/longitud
e/latitude/phosphates/nitrates

donneesSLSL2012KNN=donneesSLSL2012[,1:5]
donneesSLSLKNN=donneesSLSL[,2:6]
cl=donnees[,3]

#Prédiction de la présence/absence des algues à l'aide du Knn
KnnpredicttestSLSL <- knn(donneesSLSLKNN,donneesSLSL2012KNN,cl,k=Voisins,
l=0,
                          prob=FALSE,use.all=FALSE)
```

```
donneesSLSL2012$PredictKNN=KnnpredicttestSLSL

#Compte du nombre de prédictions réussies
p=1:nrow(donneesSLSL2012)
countSLSLKNN=0

for (i in p) {
  a=donneesSLSL2012$PredictKNN[i]
  b=donnees2012$PresAbs[i]
  if (a==b){
    countSLSLKNN=countSLSLKNN+1
  }
  else{
    countSLSLKNN=countSLSLKNN
  }
}

ratioSLSLKNN= (countSLSLKNN/nrow(donnees2012))*100

cat("Le ratio prédiction SLSL pour KNN est de ",ratioSLSLKNN,"%")

## Le ratio prédiction SLSL pour KNN est de  37.84157 %

# Définition d'un tableau de conclusion pour déterminer le meilleur des m
odèles
Conclusion2012=matrix(data = c(1,1,1), ncol=1,nrow=3)
colnames(Conclusion2012)="TauxReussite"
Conclusion2012[1,1]=ratioSLSLKNN
Conclusion2012=as.data.frame(Conclusion2012)
```

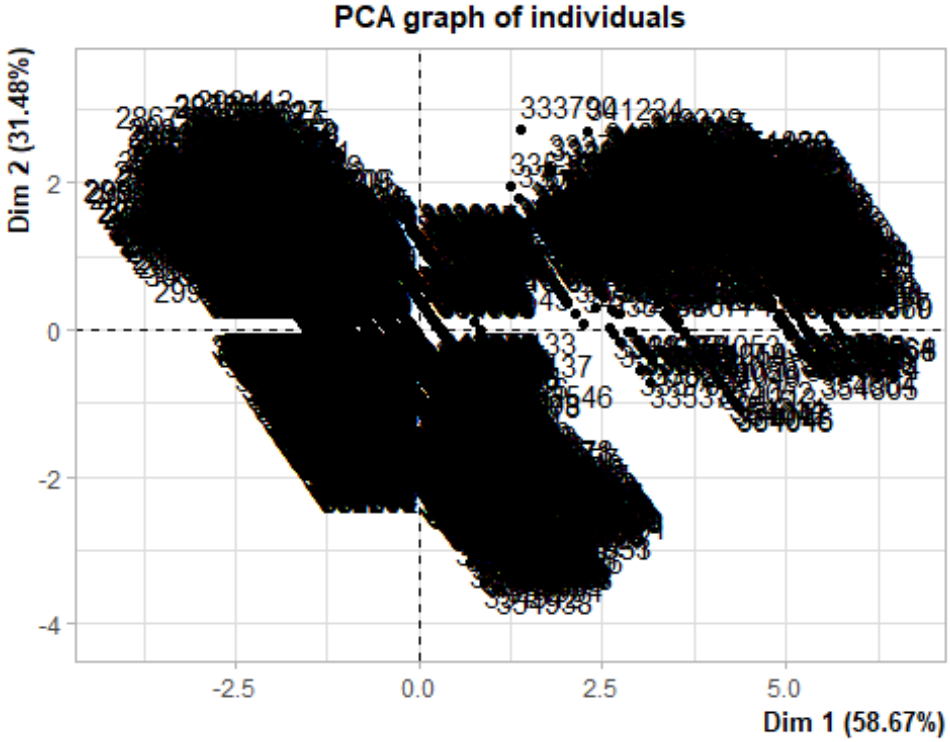
2) Analyses en composantes principales et modèle linéaire généralisé

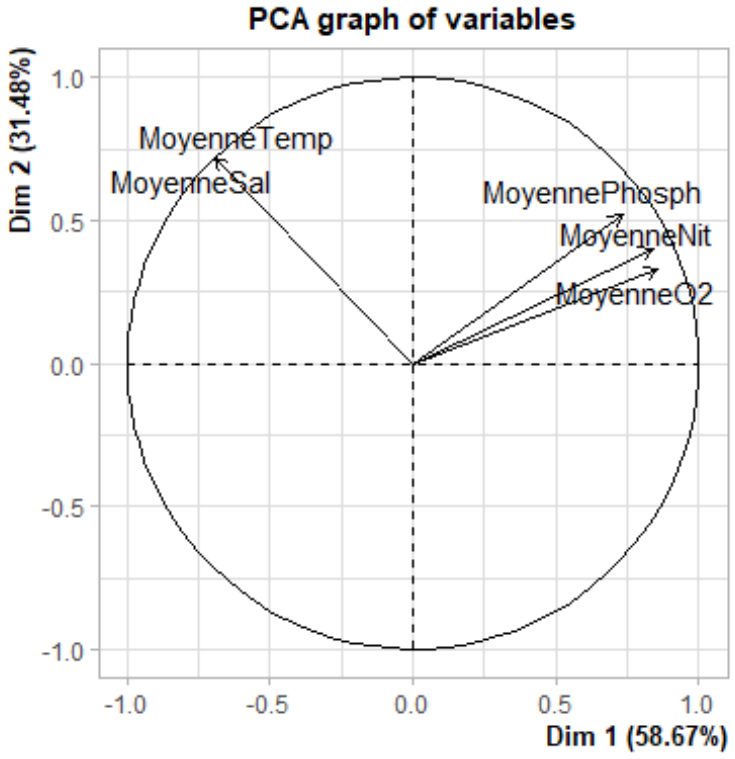
```
# Réalisation de l'ACP sur les données train (1995) puis test (2012)
res.pca.1995=PCA(donneesSLSL[,-1], scale.unit = TRUE, ncp= 5, graph =TRUE
)
```


PCA graph of individuals
Dim 2 (26.38%)
3
0
-3
-6
-2.5
0.0
2.5
5.0
Dim 1 (52.27%)
172744
172711
171731


PCA graph of variables
Dim 2 (26.38%)
1.0
0.5
0.0
-0.5
-1.0
MoyenneSal
MoyenneTemp
MoyennePhosph
MoyenneNit
MoyenneO2
-1.0
-0.5
0.0
0.5
1.0
Dim 1 (52.27%)

```r
# Détermination de la distance des individus par rapport aux composantes
PCA1995=as.data.frame(get_pca_ind(res.pca.1995)$coord)
# Ajout des données présence / abscence
mod.1995= cbind(donneesSLSL[1],PCA1995[1:5])
mod.1995$PresAbs=as.factor(mod.1995$PresAbs)
mod.1995$PresAbs=as.numeric(mod.1995$PresAbs)-1
#Réalisation d'un modèle linéaire généralisé
glmACP1995=glm(PresAbs ~ .,data=mod.1995)

donneesSLSLGLM2012=donneesSLSL2012[,1:5]
g=1:nrow(donneesSLSLGLM2012)

donneesSLSLGLM2012$MoyenneNit=as.numeric(donneesSLSLGLM2012$MoyenneNit)
donneesSLSLGLM2012$MoyenneO2=as.numeric(donneesSLSLGLM2012$MoyenneO2)
donneesSLSLGLM2012$MoyennePhosph=as.numeric(donneesSLSLGLM2012$MoyennePho
sph)
res.pca.2012=PCA(donneesSLSLGLM2012, scale.unit = TRUE, ncp= 5, graph =TR
UE)
```





On obtient ici le même résultat que pour la première ACP.

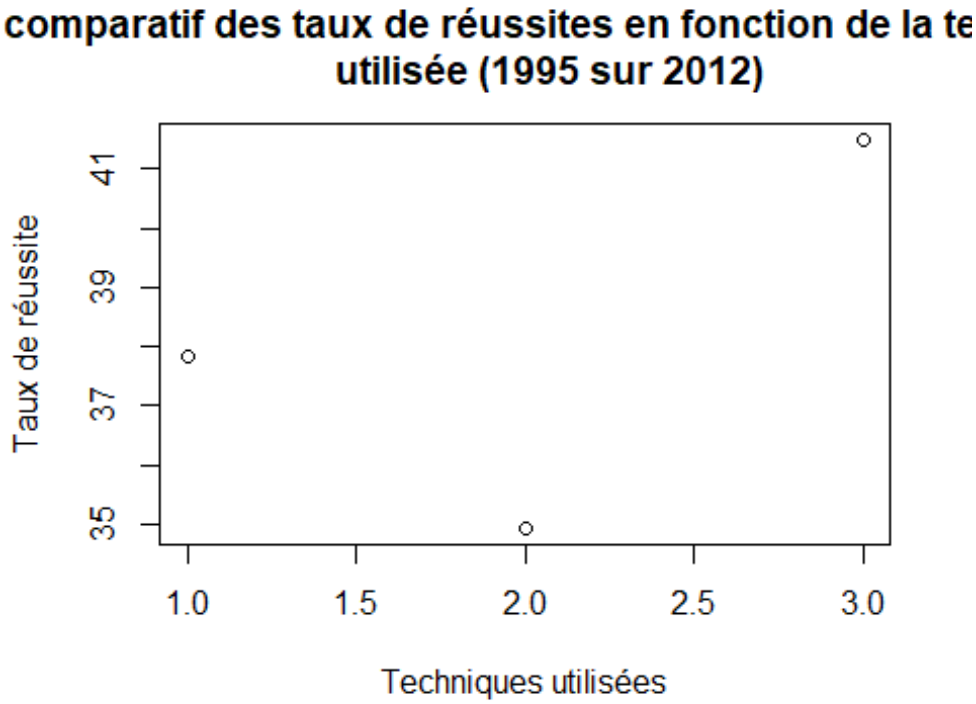
```r
PCA2012=as.data.frame(get_pca_ind(res.pca.2012)$coord)
Predict2012=as.data.frame(predict.glm
                          (glmACP1995, newdata=PCA2012,type="response"))
colnames(Predict2012)="PredictGLM"

# Prédiction des présences et des absences
t=1:nrow(PCA2012)
for (i in t){
  if (Predict2012[i,1] > 0.5){
    PCA2012$PredictGLM[i]="P"
  }
  else{
    PCA2012$PredictGLM[i]="A"
  }
}

# Ajout de la colonne présence/absence
PCA2012$PresAbs=donnees2012[,8]

# Calcul du nombre de bonnes prédictions
p=1:nrow(PCA2012)
countPCAGLM2012=0
for (i in p) {
  a=PCA2012$PredictGLM[i]
  b=PCA2012$PresAbs[i]
  if (a==b){
    countPCAGLM2012=countPCAGLM2012+1
  }
  else{
    countPCAGLM2012=(((countPCAGLM2012+1)-1)/1)
  }
}

# Calcul du pourcentage de bonnes prédictions
ratioPCAGLM2012= (countPCAGLM2012/nrow(Predict2012))*100

cat("Le ratio prédiction pour PCAGLM est de ",ratioPCAGLM2012,"%")

## Le ratio prédiction pour PCAGLM est de  34.92962 %

Conclusion2012[2,1]=ratioPCAGLM2012
```

3) Forêts aléatoires

```
#Utilisation de l'objet fit pour prédire la présence/absence des données de 2012
donneesSLSL2012$predictRF=predict(fit,donneesSLSL2012)

#Utilisation des tableaux pour comparer les prédictions et les faits
p=1:nrow(donneesSLSL2012)
countSLSLRF=0

for (i in p) {
  a=donneesSLSL2012$predictRF[i]
  b=donnees2012$PresAbs[i]
  if (a==b){
    countSLSLRF=countSLSLRF+1
  }
  else{
    countSLSLRF=countSLSLRF
  }
}

ratioSLSLRF= (countSLSLRF/nrow(donneesSLSL2012))*100

cat("Le ratio prédiction SLSL pour RF est de ",ratioSLSLRF,"%")

## Le ratio prédiction SLSL pour RF est de  41.49876 %

Conclusion2012[3,1]=ratioSLSLRF
```

Représentation graphique du taux de réussite de prédiction de chaque technique

```
Conclusion2012=as.matrix(Conclusion2012)
rownames(Conclusion2012)=c("KNN","GLMACP","RF")
plot(Conclusion2012, xlab="Techniques utilisées",ylab="Taux de réussite", main="Plot comparatif des taux de réussites en fonction de la technique utilisée (1995 sur 2012)")
```



On observe que le taux de réussite de prédiction de la présence des individus de *Saccorhiza polyschides* dans la zone étudiée est le plus important pour l'algorithme Random forest.

Ce résultat nous amène à choisir cette méthode de machine learning pour modéliser la distribution future de cette espèce sur la deuxième moitié du 21è siècle.

Partie 4 - Prédiction de la distribution des populations d'algues *Saccorhiza polyschides* au niveau de la Manche sur les périodes 2040-2050 et 2090-2100

Dans cette dernière partie, nous allons modéliser la distribution des algues *Saccorhiza polyschides* sur les périodes 2040-2050 et 2090-2100 en utilisant la méthode des Random Forest, définie dans la partie précédente comme ayant la meilleure capacité de prédiction de la distribution des individus (à la fois pour la validation croisée sur 1995 et la prédiction sur 2012). Cette modélisation se fera suivant deux scénarios de réchauffement climatique établis par le Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC): un scénario optimiste, de réduction de gaz à effet de serre (RCP-2.6) et un scénario pessimiste, de forte émission de gaz à effet de serre (RCP-8.5).

En raison de notre incapacité à obtenir les données de Température associées à chaque position (latitude/longitude) étudiée et pour les deux scénarios à partir du site du

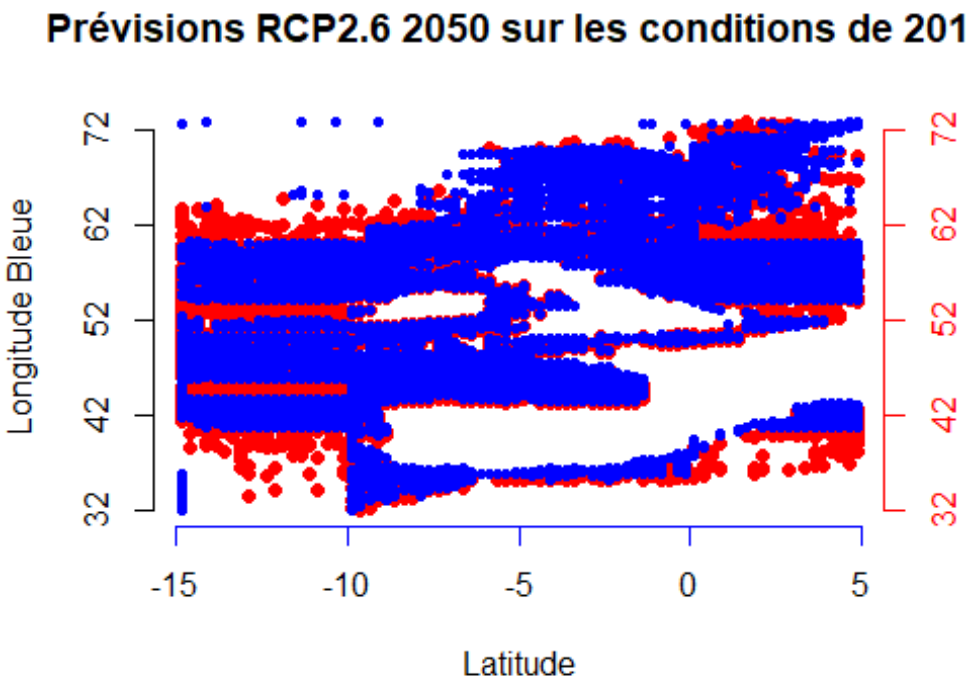
World Data Center for Climate et de part les données incomplètes fournies par Frida Ben Rais Lasram et al. dans leur étude, qui ne recouvrent pas la totalité de notre zone d'étude, nous avons choisi de modéliser les changements associés aux variables étudiées manuellement. Pour cela, nous nous sommes basés sur les données fournies par le IPCC Special Report on Ocean and Cryosphere in a Changing Climate (SROCC) publié en septembre 2019 https://www.ipcc.ch/site/assets/uploads/sites/3/2019/12/SROCC_FullReport_FINAL.pdf). Pour la période 2040-2050, les températures pour la période 2005-2012 en toutes positions ont été augmentées de 0.64°C pour le scénario RCP-2.6 et de 0.95°C pour le scénario RCP-8.5. De même, la quantité de O2 dissoute dans l'eau est diminuée de 0.9% pour le premier scénario et de 1.4% pour le deuxième. De la même manière, pour la période 2090-2100 les températures pour la période 2005-2012 à chaque position ont été augmentées de 0.73°C pour le scénario RCP-2.6 et de 2.58°C pour le scénario RCP-8.5. De même, la quantité de O2 dissoute dans l'eau est diminuée de 0.6% pour le premier scénario et de 3.9% pour le deuxième. Les paramètres de salinité ainsi que les paramètres de taux de nitrates et de phosphates dans l'eau sont considérés constants en raison de l'indisponibilité des données d'évolution de ces variables sur les périodes d'intérêt.

Bien que suivre cette stratégie de modélisation nous permettra de représenter au mieux pour les moyens que nous avons à notre disposition la future distribution de *Saccorhiza polyschides* selon ces deux scénarios, l'utilisation de ces valeurs globales ajoutées à chaque position ne nous permettra pas de représenter les effets différentiels du réchauffement climatique à entre des positions précises de notre zone d'étude (augmentation de la température plus ou moins importante etc…). De plus, ces valeurs sont associées aux changements à la surface de l'eau. Cependant, comme notre modèle vit en eau peu profonde et que pour cette étude nous avons moyenné les valeurs de températures et des autres variables sur la zone de 0 à 50 mètres de profondeur, nous estimons que ce choix de stratégie est légitime et que cette approximation peut être réalisée dans notre cas.

```
#2040-2050

#RCP-2.6 : température +0.64°C et quantité d'O2 dissoute dans l'eau -0.9%
donnees2050RCP26=donnees2012[,-8]
donnees2050RCP26$MoyenneTemp=donnees2050RCP26$MoyenneTemp+0.64
donnees2050RCP26$MoyenneO2=as.numeric(donnees2050RCP26$MoyenneO2)
donnees2050RCP26$MoyenneO2=donnees2050RCP26$MoyenneO2*(99.1/100)

#RCP-8.5 : température +0.95°C et quantité d'O2 dissoute dans l'eau -1.4%
donnees2050RCP85=donnees2012[,-8]
donnees2050RCP85$MoyenneTemp=donnees2050RCP85$MoyenneTemp+0.95
donnees2050RCP85$MoyenneO2=as.numeric(donnees2050RCP85$MoyenneO2)
donnees2050RCP85$MoyenneO2=donnees2050RCP85$MoyenneO2*(98.6/100)
```

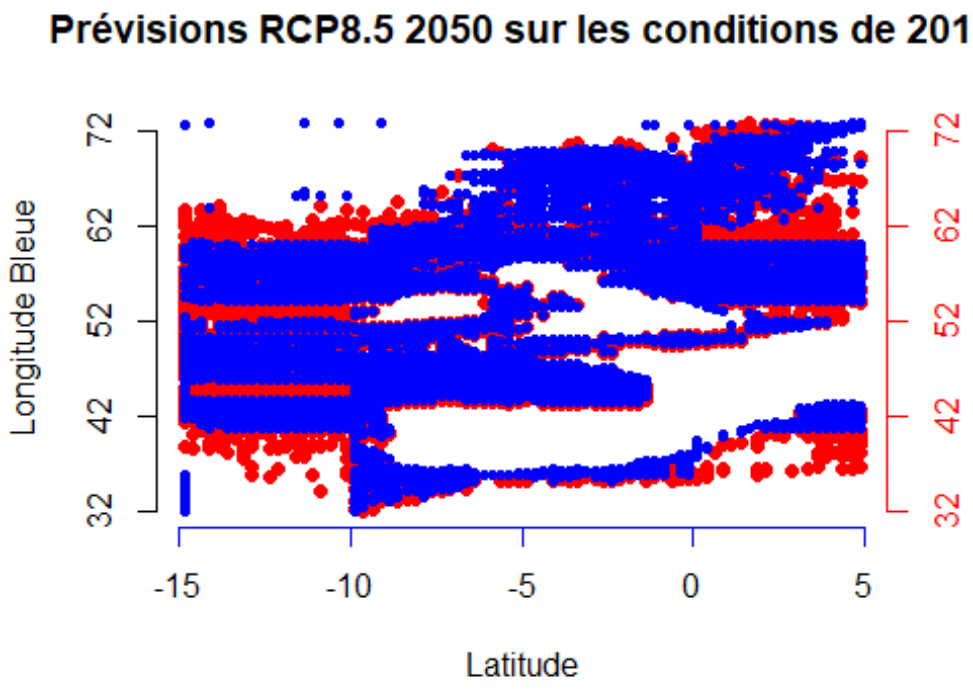
```r
#Garder les présences et leurs coordonnées associés pour la période 2005-2012
donnees2012Presabs=subset(donnees2012,PresAbs=="P")
donnees2012Pres=donnees2012Presabs[,c(1,2)]

#Préparation des données de 2005-2012 pour la prédiction random forest
donneesSLSL2012$PresAbs=donnees2012$PresAbs
dataSLSL2012=donneesSLSL2012[,-c(6,7)]
dataSLSL2012$PresAbs=as.factor(dataSLSL2012$PresAbs)

#Random forest appliqué sur les données de 2005-2012
fit2 = randomForest(PresAbs ~ ., dataSLSL2012,mtry = 2)

#Garder les coordonnées pour la période 2040-2050 et pour les deux scénarios
donneesSLSL2050RCP26=donnees2050RCP26[,-c(1,2)]
donneesSLSL2050RCP85=donnees2050RCP85[,-c(1,2)]

#Prédiction des présences
donneesSLSL2050RCP26$predictRF=predict(fit2,donneesSLSL2050RCP26)
donneesSLSL2050RCP85$predictRF=predict(fit2,donneesSLSL2050RCP85)

#Attribution des prédictions
donnees2050RCP26$PredictPresAbs=donneesSLSL2050RCP26$predictRF
donnees2050RCP85$PredictPresAbs=donneesSLSL2050RCP85$predictRF

#Garder les présences / absences ainsi que les coordonnées pour la période 2040-2050 et pour les deux scénarios
donnees2050RCP26datapred=donnees2050RCP26[,c(1,2,8)]
donnees2050RCP85datapred=donnees2050RCP85[,c(1,2,8)]

#Sélectionner que les présences
donnees2050RCP26datapred=subset(donnees2050RCP26datapred,PredictPresAbs=="P")
donnees2050RCP85datapred=subset(donnees2050RCP85datapred,PredictPresAbs=="P")

#Garder seulement les coordonnées associés à ces présences pour le plot
donnees2050RCP26Pres=donnees2050RCP26datapred[,-3]
donnees2050RCP85Pres=donnees2050RCP85datapred[,-3]

#Visualisation graphique du 1er scénario en comparaison aux données 2005-2012
plot.new()
plot(donnees2050RCP26Pres[,2],donnees2050RCP26Pres[,1],col="red",axes=F,x
```

```
lab=""
     ,ylab="",pch=16,main="Prévisions RCP2.6 2050 sur les conditions de
2012")
par(new = T)
plot(donnees2012Pres[,2],donnees2012Pres[,1],col="blue",axes=F,xlab="Lati
tude",
     ylab="Longitude Bleue",pch=20)
axis( 2 , ylim=c(-15,5),col="black",col.axis="black",at=seq(32, 72, by=10
))
axis(1, ylim=c(32,72),col="blue",col.axis="black",at=seq(-15, 5, by=5))
axis( 4 ,col="red",col.axis="red",at=seq(32, 72, by=10))
mtext("Longitude rouge",side=4,line=2.5,col="red")
```



Une réduction des émissions de gaz à effet de serre d'ici 2050 semble induire une persistance globale des populations Saccorhiza polyschides dans les zones qu'elles occupent actuellement (points bleus).

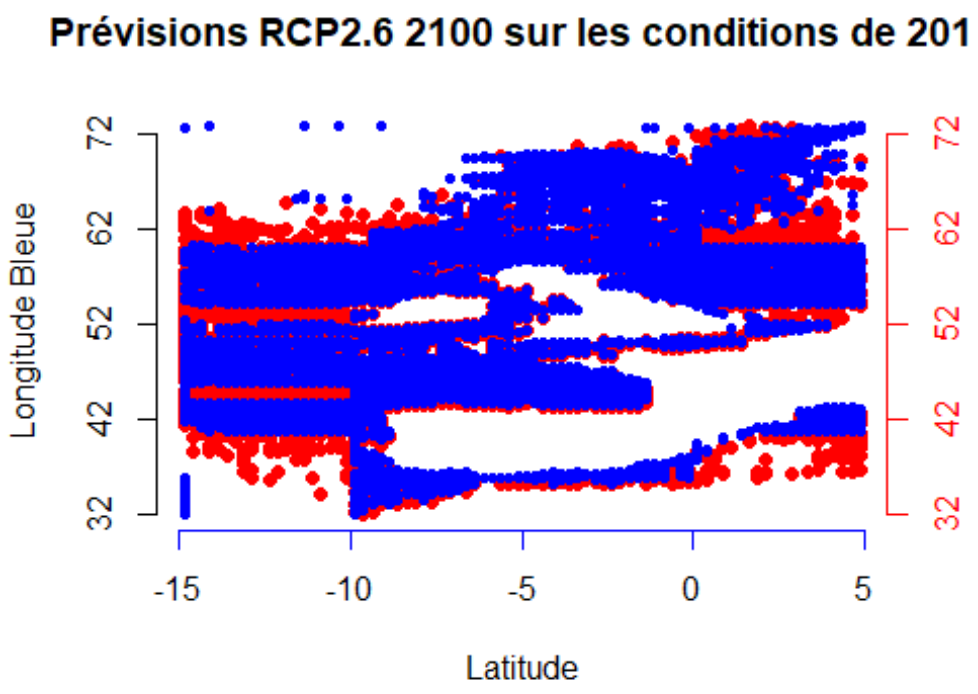
```r
#Visualisation graphique du 2nd scénario en comparaison aux données 2005-2012
plot.new()
plot(donnees2050RCP85Pres[,2],donnees2050RCP85Pres[,1],col="red",axes=F,xlab="",
     ylab="",pch=16,main="Prévisions RCP8.5 2050 sur les conditions de 2012")
par(new = T)
plot(donnees2012Pres[,2],donnees2012Pres[,1],col="blue",axes=F,xlab="Latitude",
     ylab="Longitude Bleue",pch=20)
axis( 2 , ylim=c(-15,5),col="black",col.axis="black",at=seq(32, 72, by=10))
axis(1, ylim=c(32,72),col="blue",col.axis="black",at=seq(-15, 5, by=5))
axis( 4 ,col="red",col.axis="red",at=seq(32, 72, by=10))
mtext("Longitude rouge",side=4,line=2.5,col="red")
```



Une augmentation des émissions de gaz à effet de serre d'ici 2050 ne semble également pas induire un déplacement des populations Saccorhiza polyschides des zones qu'elles occupent actuellement. Tout comme le premier scénario, un léger décalage de celles-ci vers des longitudes plus élevées est cependant observable au niveau de l'océan Atlantique et de la mer Baltique (points rouges).

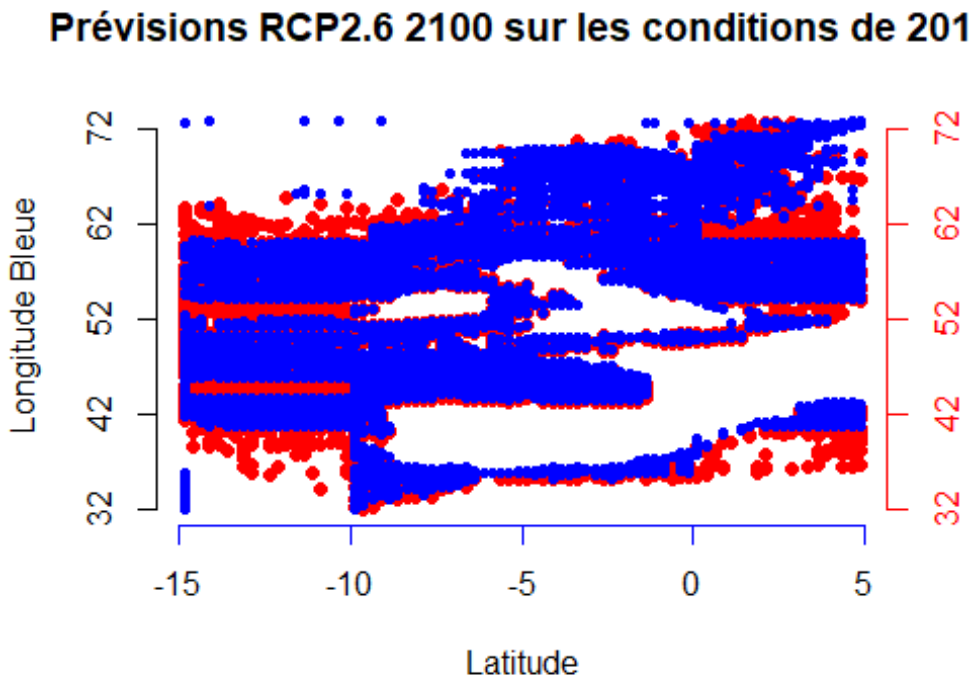
```
#2090-2100

#RCP-2.6 : température +0.73°C et quantité d'O2 dissoute dans l'eau -0.6%
donnees2100RCP26=donnees2012[,-8]
donnees2100RCP26$MoyenneTemp=donnees2100RCP26$MoyenneTemp+0.73
donnees2100RCP26$MoyenneO2=as.numeric(donnees2100RCP26$MoyenneO2)
donnees2100RCP26$MoyenneO2=donnees2100RCP26$MoyenneO2*(99.4/100)

#RCP-8.5 : température +2.58°C et quantité d'O2 dissoute dans l'eau -3.9%
donnees2100RCP85=donnees2012[,-8]
donnees2100RCP85$MoyenneTemp=donnees2100RCP85$MoyenneTemp+2.58
donnees2100RCP85$MoyenneO2=as.numeric(donnees2100RCP85$MoyenneO2)
donnees2100RCP85$MoyenneO2=donnees2100RCP85$MoyenneO2*(96.1/100)

#Garder les coordonnées pour la période 2090-2100 et pour les deux scénar
ios
donneesSLSL2100RCP26=donnees2100RCP26[,-c(1,2)]
donneesSLSL2100RCP85=donnees2100RCP85[,-c(1,2)]

#Prédiction des présences
donneesSLSL2100RCP26$predictRF=predict(fit2,donneesSLSL2100RCP26)
donneesSLSL2100RCP85$predictRF=predict(fit2,donneesSLSL2100RCP85)

#Attribution des prédictions
donnees2100RCP26$PredictPresAbs=donneesSLSL2100RCP26$predictRF
donnees2100RCP85$PredictPresAbs=donneesSLSL2100RCP85$predictRF

#Garder les présences / absences ainsi que les coordonnées pour la périod
e
#2050-2100 et pour les deux scénarios
donnees2100RCP26datapred=donnees2100RCP26[,c(1,2,8)]
donnees2100RCP85datapred=donnees2100RCP85[,c(1,2,8)]

#Sélectionner que les présences
donnees2100RCP26datapred=subset(donnees2100RCP26datapred,PredictPresAbs==
"P")
donnees2100RCP85datapred=subset(donnees2100RCP85datapred,PredictPresAbs==
"P")

#Garder seulement les coordonnées associés à ces présences pour le plot
```

```r
donnees2100RCP26Pres=donnees2100RCP26datapred[,-3]
donnees2100RCP85Pres=donnees2100RCP85datapred[,-3]

#Visualisation graphique du 1er scénario en comparaison aux données 2005-2012
plot.new()
plot(donnees2100RCP26Pres[,2],donnees2100RCP26Pres[,1],col="red",axes=F,xlab="",
     ylab="",pch=16,main="Prévisions RCP2.6 2100 sur les conditions de 2012")
par(new = T)
plot(donnees2012Pres[,2],donnees2012Pres[,1],col="blue",axes=F,xlab="Latitude",
     ylab="Longitude Bleue",pch=20)
axis( 2 , ylim=c(-15,5),col="black",col.axis="black",at=seq(32, 72, by=10))
axis(1, ylim=c(32,72),col="blue",col.axis="black",at=seq(-15, 5, by=5))
axis( 4 ,col="red",col.axis="red",at=seq(32, 72, by=10))
mtext("Longitude rouge",side=4,line=2.5,col="red")
```



Une réduction des émissions de gaz à effet de serre et un maintien à des taux faibles entre 2050 et 2100 semble induire une persistance des populations Saccorhiza polyschides dans les zones qu'elles occupent actuellement et occuperont vers 2050.

```
#Visualisation graphique du 2nd scénario en comparaison aux données 2005-2012
plot.new()
plot(donnees2100RCP26Pres[,2],donnees2100RCP26Pres[,1],col="red",axes=F,xlab="",
     ylab="",pch=16,main="Prévisions RCP2.6 2100 sur les conditions de 2012")

par(new = T)
plot(donnees2012Pres[,2],donnees2012Pres[,1],col="blue",axes=F,xlab="Latitude",
     ylab="Longitude Bleue",pch=20)
axis( 2 , ylim=c(-15,5),col="black",col.axis="black",at=seq(32, 72, by=10))
axis(1, ylim=c(32,72),col="blue",col.axis="black",at=seq(-15, 5, by=5))
axis( 4 ,col="red",col.axis="red",at=seq(32, 72, by=10))
mtext("Longitude rouge",side=4,line=2.5,col="red")
```



Une augmentation constante des émissions de gaz à effet de serre durant la deuxième moitié du 21ème siècle ne semble toujours pas induire un déplacement significatif des populations Saccorhiza polyschides des zones qu'elles occupent aujourd'hui. Un léger décalage de ces populations vers des longitudes plus élevées est cependant toujours observable au niveau de l'océan Atlantique et de la mer Baltique. De plus, une disparition des points rouges (prédiction) des longitudes plus basses est observable par rapport au même scénario pour 2050.

Conclusion

L'utilisation d'outils de machine learning nous a permis, dans cette étude, de prédire la distribution des populations d'algues brunes Saccorhiza polyschides dans l'Océan Atlantique Nord et les mers bordant les pays Scandinaves et d'Europe de l'Ouest, en fonction de deux scénarios de réchauffement climatique totalement opposés: un optimiste représentant une réduction des émissions de gaz à effet de serre et un maintien à des taux faibles durant ce siècle, et un scénario pessimiste, représentant l'inconscience de la société de ce problème environnementale et dont l'émission de ces gaz augmente constamment jusqu'en 2100. Selon ce premier scénario, nous avons pu prédire une persistance générale des populations d'algues aux positions qu'elles occupent actuellement. Cela est assez intuitif au vu de ce qu'il représente: de faibles changements des conditions environnementales qui restent alors similaires à celles actuellement. Le réchauffement climatique ne semble cependant pas avoir d'effet significatif sur cette distribution. En effet, comme nous le montrent les figures associées au scénario d'émissions constantes et élevées de ces gaz, dans cette situation aucun déplacement significatif de ces populations n'est observable. Seulement un léger déplacement vers des longitudes plus élevées (également visible pour le scénario optimiste) où l'eau est plus froide, et un dépeuplement des longitudes plus basses où l'eau est plus chaude, sont visibles. Bien que le réchauffement climatique soit mortel pour un grand nombre d'espèces, on pourrait supposer qu'ici les changements des conditions environnementales induits par celui-ci ne sont pas assez importants pour déplacer la zone d'habitabilité de cette espèce: la température de l'eau ne serait pas un facteur limitant pour sa survie car celle-ci pourrait vivre selon une gamme de température assez large. Étudier les propriétés biologiques de cette espèce d'algues brunes serait donc nécessaire par la suite afin de vérifier cette hypothèse. Il serait également intéressant d'utiliser une autre méthode de machine learning pour prédire cette distribution afin de vérifier la reproductibilité des résultats obtenus avec la méthode que nous avons choisie. D'autres paramètres pourraient également intervenir dans l'obtention de tels résultats de distribution et nous en discutons dans la dernière partie.

Discussion

Bien que cette étude permet de prédire un déplacement léger des populations de *Saccorhiza polyschides* vers des eaux plus froides en raison du réchauffement climatique, cette prédiction reste cependant incertaine car de nombreux autres paramètres évoluant en fonction du temps seraient également à prendre en considération lors de la construction du modèle de prédiction.

Premièrement, nous n'avons pas considéré les réponses physiologiques propres aux espèces face aux changements des conditions abiotiques lors du réchauffement climatique : certaines peuvent s'y adapter à travers des mécanismes génétiques et ainsi survivre même si la température ou tout autre facteur augmente ou diminue. Cependant, sachant que ces changements se passent très vite en raison de l'activité humaine (beaucoup plus vite que ceux qui se sont déjà passés dans l'histoire de la planète), cette adaptation peut apparaître comme moins probable pour celles-ci au vu de la durée nécessaire à la mise en place de ces mécanismes adaptatifs. De plus, l'action anthropique est également à prendre en considération notamment la pêche qui, bien que l'espèce étudiée ne soit pas "récoltée" et utilisée par l'Homme, présente un impact non négligeable sur les écosystèmes marins en modifiant les sols près des côtes ou en déséquilibrant le rapport naturel entre les différentes espèces qui constituent ces niches. Deuxièmement, il est important de se rappeler que dans tout écosystème, de la compétition entre différentes espèces existe pour l'accès aux ressources limitées du milieu (facteur biotique). Cette compétition dicte l'occupation et l'exclusion de certaines espèces de certaines zones du milieu. Un telle compétition existe entre *Saccorhiza polyschides* et *Laminaria digitata*, une autre espèce d'algue brune de la famille des laminaires (Valero M. et Engel C. (coord.), Arzel P., Creach A., Davoult D., Destombe C., Gevaert F., Leblanc C., Levavasseur G., Potin P., Viard F., 2008. Dynamique des champs de Laminaria digitata, ressource algale en Bretagne : impacts biotiques, abiotiques et anthropiques.). Ce facteur biotique qui influe sur la distribution des populations de chaque espèce serait donc important à considérer notamment en sachant que contrairement au modèle d'étude, *Laminaria digitata* est une algue pêchée par l'humain ce qui n'est donc pas sans conséquence au niveau du rapport de force entre les deux espèces.

Enfin, il est très probable que ces faibles déplacements observables des populations *Saccorhiza polyschides* soient dus à des jeux de données incomplets, ne représentant ainsi que très vaguement les conditions du milieu prédites pour chaque scénario RCP et pour chaque période. Prendre seulement en considération l'évolution de la température de l'eau et de la quantité de O2 dissoute dans celle-ci semble être insuffisant pour pouvoir conclure quant à un réel impact du réchauffement climatique sur la distribution des populations de *Saccorhiza polyschides*. Inclure l'évolution de la salinité, du taux de nitrates et de phosphates dissous dans l'eau serait intéressant pour améliorer le modèle. Cependant pour les deux derniers paramètres, cela nécessiterait

des études préliminaires de prédiction de leur évolution. Bien que leur intérêt biologique soit indiscutable, leur intérêt dans la construction d'un modèle de prédiction de la distribution des populations de cette espèce d'algues (parmi d'autres) apparaît donc quant à lui discutable. De plus, la distribution des populations d'une espèce pouvant suivre celle des courants marins notamment au vu de leur capacité à transporter les nutriments, inclure le gradient de température (un courant étant un déplacement horizontal d'eau dû notamment aux différences de température) le long de la zone étudiée serait également intéressant par la suite pour observer si un déplacement de ces courants dû au réchauffement climatique pourrait induire un déplacement des populations de cette espèce d'algues (ou tout autre espèce animale ou végétale marine). En annexe est trouvable une ébauche de code permettant la création de cette variable gradient de température à partir des valeurs de température de chaque position (latitude/longitude) de la zone étudiée.

Références

Lasram et al., 2020. An open-source framework to model present and future marine species distributions at local scale. *Ecological Informatics* Laffoley, D. and Baxter, J.M. 2019. *Ocean deoxygenation : everyone's problem : causes, impacts, consequences and solutions* by IUCN (International Union for Conservation of Nature) Pörtner et al., 2019. *Special Report on Ocean and Cryosphere in a Changing Climate* by IPCC (Intergovernmental Panel on Climate Change) Valero et al., 2008. Dynamique des champs de *Laminaria digitata*, ressource algale en Bretagne : Impacts biotiques, abiotiques et anthropiques.

Annexe

donnees$grad=0
colonne 1=latitude, colonne 2=longitude, colonne3=présence/absence
colonne 4= température, colonne 5=salinité, colonne 6=gradient
donneessup=donnees on crée un jeu de données sur lequel on va supprimer les positions étudiées
donneessup[, 7] = seq(1, length(t(donnees))) on crée une colonne qui va indiquer l'indice de chaque ligne

count = length(x) compteur pour la boucle while

```
while (count > 0) {
2
nbrsup = 0
lsup = c()
for (i in 1:count) {
  4
  if (donneessup[i, 2] < (donneessup[1, 2] + 0.25) & donneessup[i, 2] >
```

```
   (donneessup[1, 2] - 0.25) & donneessup[i, 1] < (donneessup[1, 1] + 0.25) &
   donneessup[i, 1] > (donneessup[1, 1] - 0.25)){
       1 nbrsup = nbrsup + 1 lsup = c(lsup, i)
    }
}
count = count - nbrsup
donnees[donneessup[c(lsup), 7], 6] = mean(donneessup[c(lsup), 4])
donneessup = donneessup[-lsup, ]
}


View(donnees)
```