Projet

Analyse des données multivariées

● Contexte biologique

Parmi les Salmonidés, certains remontent chaque année les cours d'eau, à des époques déterminées. Beaucoup de Saumons proprement dits, nés en eau douce, se rendent à la mer puis remontent les fleuves à l'époque de la ponte : ce sont les migrateurs. D'autres sont sédentaires : ceux-là vivent dans les grandes rivières ou dans les lacs aux eaux pures et transparentes ou même dans les ruisseaux limpides.

La répartition des espèces de poissons sur un cours d'eau dépend de beaucoup de facteurs, mais la pente, le gabarit et la température sont les principaux. M. Verneaux en 1973 a démontré que ces paramètres expliquent majoritairement la répartition.

● Contenu des données

Les données décrivent 274 bassins hydrographiques européens et sont réparties en 4 fichiers.

- *data_geo* contient des variables géographiques ;
- *data_climat* contient des variables climatiques ;
- *data_landcover* contient des variables d'occupation du sol ;
- *data_sp* contient les présences-absences de 10 espèces de Salmonidés.

● Démarche analytique

- ACP de data_sp, puis de data_geo, data_climat et data_landcover
- ACC : AFC du tableau des espèces et ACP normée du tableau milieu
- ACC librairie vegan
- Analyse de co-inertie
- Approche descriptive : statistique exploratoire
- Classification : clustering k-means
- Apprentissage supervisée : arbres de décisions
- Conclusion

● Exploration des données

Pour synthétiser l'information contenue dans les données, il est possible de réaliser une Analyse en Composantes Principales (ACP), par exemple sur le tableau des espèces.

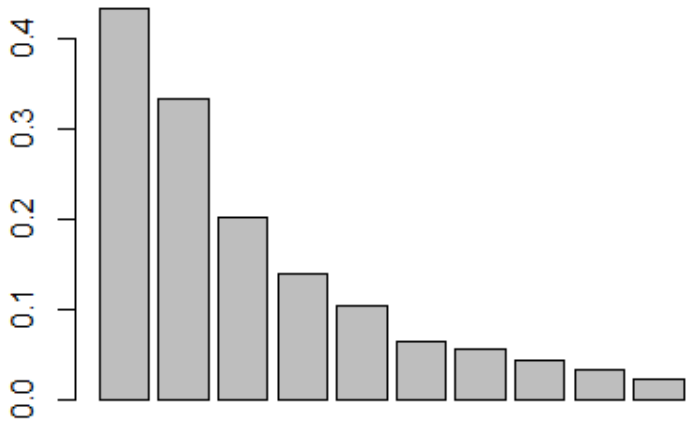


Figure 1 : Valeurs propres de acp_sp

Intéressons nous d'abord aux valeurs propres. Afin d'obtenir la proportion de variance portée par les 2 premiers axes de l'ACP, on divise la somme des valeurs propres correspondantes par la somme totale des valeurs propres. On effectue cela pour une ACP non normée et pour une ACP normée.

On obtient approximativement 53% de la variance totale expliquée par les 2 premiers axes dans le cas d'une ACP non normée. 10% de l'information est perdue pour une ACP normée. Il a donc été décidé de sélectionner l'ACP non normée afin de poursuivre l'étude.

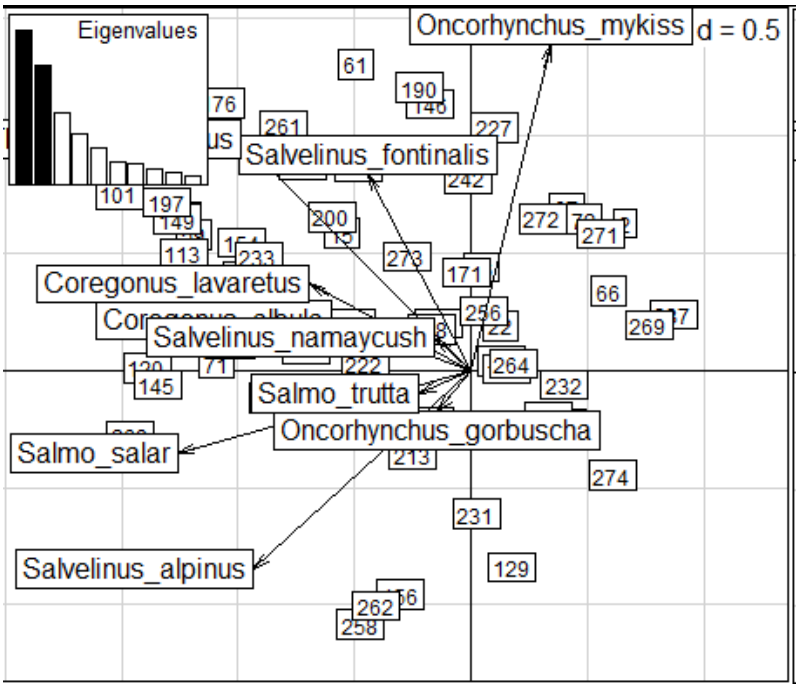


Figure 2 : Biplot de l'ACP non normée de data_sp

Chacun des axes vise à maximiser la dispersion des points. On a décidé de s'intéresser aux espèces avec une répartition particulière, ce qui signifie qu'elles sont éloignées de l'origine et donc fortement corrélées à l'un des axes. Par exemple, les deux espèces *Oncorhynchus mykiss* et *Salvelinus alpinus* sont fortement corrélées à l'axe 2.

● Problématique

Dans un premier temps, nous allons nous demander si la répartition particulière d'espèces telles que *Oncorhynchus mykiss* ou *Salvelinus alpinus* peut s'expliquer par des variables environnementales. Dans ce cas, on pourrait déterminer lesquelles.

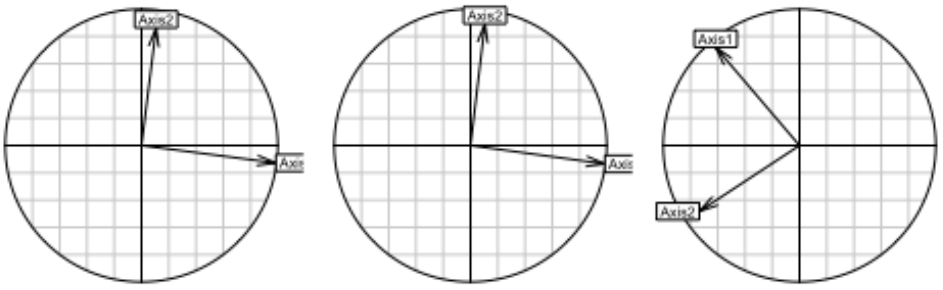
Dans un second temps, nous nous intéresserons aux différences de diversité des espèces entre les sites.

● Analyse canonique des correspondances (ACC)

L'ACC correspond à un couplage entre une Analyse Factorielle des Correspondances (AFC) sur le tableau des espèces, et une ACP normée sur le tableau milieu. Pour choisir le tableau milieu le plus adapté à l'analyse, il est possible de projeter les axes d'inertie de l'AFC, représentés par les flèches, sur ceux de l'ACC.

A partir du tableau des variables environnementales, l'ACC cherche des axes qui séparent au mieux les espèces. Ces axes devraient alors être proches de ceux de l'AFC s'ils donnent une bonne représentation du tableau des espèces.

On utilise les 3 tableaux milieu *data_geo, data_climat, data_landcover* contenant des variables géographiques, climatiques et d'occupations du sol.



data_geo *data_climat* *data_landcover*

Figure 3 : Cercles de corrélations des variables du milieu

Pour *data_geo* et *data_climat*, les axes trouvés par l'ACC sont très proches et corrélés à ceux de l'AFC, et donnent donc une bonne représentation du tableau des espèces. La suite de l'analyse devrait se baser sur l'un de ces deux tableaux.

Pour le tableau *data_climat*, il est possible de représenter graphiquement les positions des sites de bassins hydrographiques et des espèces de Salmonidés.

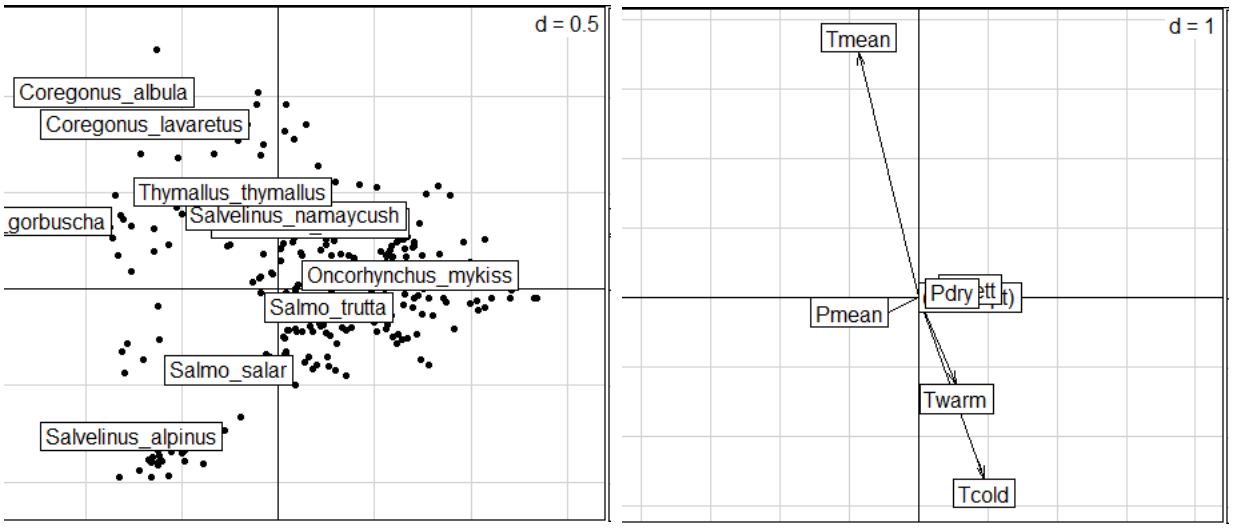


Figure 4 : (à gauche) Projection des sites ; (à droite) Projection des variables de data_climat
Localisation des sites et des espèces selon des variables environnementales

Les sites sont représentés par les points noirs, leur position est définie selon leurs caractéristiques environnementales. Les axes sont construits avec les coefficients de combinaisons linéaires des variables environnementales.

On observe que les espèces retrouvées dans de nombreux bassins hydrographiques sont majoritairement situées près de l'origine, telles que *Oncorhynchus mykiss, Salvelenus namaycush et Salmo trutta*. Éloignées de l'origine, des espèces telles que *Salvelinus alpinus* sont caractéristiques de certains bassins.

L'hypothèse de départ est que la répartition des espèces caractéristiques peut être influencée par des variables environnementales. Il est intéressant de mêler la projection des variables de data_climat à celle des sites. On peut affirmer que les variables les plus représentatives, c'est-à-dire celles contribuant le plus aux axes sont celles représentées par les flèches les plus longues. On étudie aussi les flèches les plus corrélées aux axes.

Par juxtaposition des graphiques, on étudie l'influence de x variables sur y individus. On observe une corrélation positive entre *Salvelinus alpinus* et la température des 3 mois les plus froids (*Tcold*), ainsi qu'une corrélation entre *Coregonus albula* et la température annuelle moyenne (*Tmean*).

L'hypothèse qui peut ressortir est que la répartition particulière de *Salvelinus alpinus* s'explique en partie par une préférence de cette espèce pour les températures froides (variable *Tcold*).

ACC librairie vegan

La construction d'un modèle est nécessaire pour déterminer quelles variables expliquent significativement la répartition des espèces. En fonction de nos résultats obtenus avec l'ACC précédente, nous allons tester l'effet des températures.

Twarm : température des trois mois les plus chauds (°C)

Tcold : température des trois mois les plus froids (°C)

Notre hypothèse nulle H0 du test statistique est que ces deux variables n'ont pas d'effet significatif sur la répartition des espèces.
Un test ANOVA a d'abord été effectué sur notre premier modèle :

```
Permutation test for cca under reduced model
Permutation: free
Number of permutations: 999

Model: cca(formula = data_sp ~ Twarm + Tcold, data = data_climat)
          Df    ChiSquare   F           Pr(>F)
Model     2     0.42412     53.741      0.001 ***
Residual  271   1.06934
```

p-value < 0.05, H0 est donc rejetée.
Ainsi la répartition des espèces peut s'expliquer selon ces deux variables prises ensemble.

Test ANOVA effectué ensuite sur notre second modèle :

```
Permutation test for cca under reduced model
Permutation: free
Number of permutations: 999

Model: cca(formula = data_sp ~ Pwett + Condition(Twarm + Tcold), data =
data_climat)
          Df    ChiSquare   F           Pr(>F)
Model     1     0.00727     1.8474      0.072 .
Residual  270   1.06207
```

p-value > 0.05, H0 n'est donc pas rejetée.
La variable *Pwett n*'améliore pas le modèle, ce qui signifie que *Pwett* n'a pas d'effet significatif une fois que *Twarm* et *Tcold* sont prises en compte. *Pwett* n'a pas d'effet additif supplémentaire.

Analyse de co-inertie

La donnée *data_landcover* n'étant pas retenue pour l'ACC, elle sera utilisée pour l'analyse de co-inertie. L'objectif est d'établir une structure commune entre *data_landcover* et *data_sp*.

Une ACP normée est réalisée sur le tableau *data_landcover* car les unités des variables sont différentes. Un couplage est ensuite effectué entre cette ACP et celle réalisée précédemment sur *data_sp*. Nous obtenons les résultats suivants :

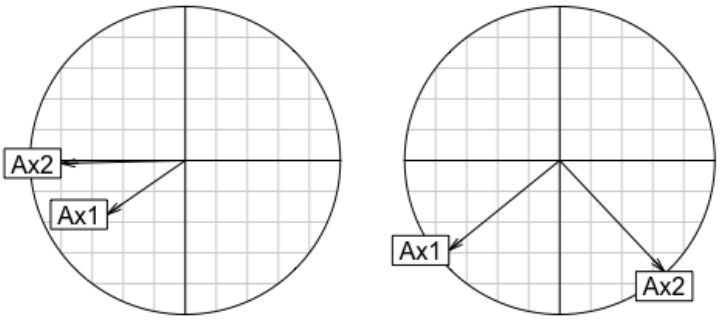


Figure 5 : (à gauche) Cercle de corrélation acp_sp axes / coinertia axes ; (à droite) Cercle de corrélation acp_landcover axes / coinertia axes

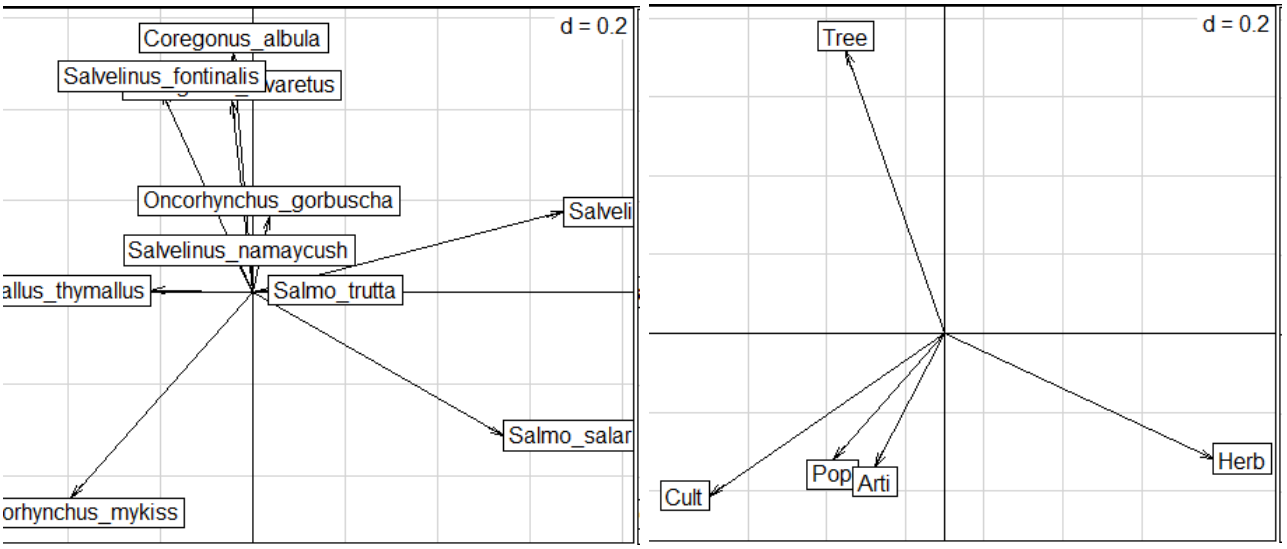


Figure 6 : (à gauche) Axes principaux (acp_sp cols) ; (à droite) Composantes principales (acp_landcover cols)

De nouveaux liens entre les variables et les espèces ont été établis à l'aide de cette analyse de co-inertie.

On observe une corrélation positive entre la densité de population humaine et *Oncorhynchus mykiss*, cela pourrait suggérer une adaptation de cette espèce à la cohabitation avec les humains.

Au contraire, on observe une corrélation négative entre la densité de population et l'espèce *Salvelinus alpinus.* Cette espèce pourrait donc privilégier les zones moins densément peuplées.

Le pourcentage de prairies est corrélé positivement à l'espèce *Salmo Salar*, tout comme le pourcentage de forêt corrélé positivement à *Salvelinus fontinalis*. Ces espèces pourraient préférer des zones à faible densité de population. On pourrait émettre l'hypothèse que cela s'expliquerait par des facteurs tels que la pollution de l'eau.

- Approche descriptive : statistique exploratoire

La première partie de l'analyse se focalisant sur la répartition des espèces en fonction des différentes variables environnementales, nous allons maintenant étudier les différences de diversité des espèces entre les sites.

Ces différences de diversité vont être étudiées selon deux critères : une forte et une faible diversité. Cette dernière contient des sites avec moins d'espèces que le premier quartile. Cela correspond à des sites correspondant à une seule espèce tel que *Oncorhynchus mykiss.*

Notre ACC réalisée précédemment le démontre car cette espèce se situait proche de l'origine.

- Classification : clustering k-means

Un clustering, apprentissage non supervisé, a été réalisé à l'aide d'une approche k-means dont le but est d'affecter arbitrairement des centres de clusters (nommés centroids), puis d'assigner chaque point de nos données au centroid qui lui est le plus proche. Ceci s'effectue jusqu'à assigner toutes les données à un cluster.

Sachant que l'on a décidé de répartir les données pour cette approche descriptive en deux groupes (selon les quartiles), faible et forte diversité, nous aurons donc deux clusters. Toutes les variables environnementales ont été prises en compte. Un clustering semble intéressant ; il s'agit de celui réalisé avec la longitude et la latitude provenant des variables géographiques.

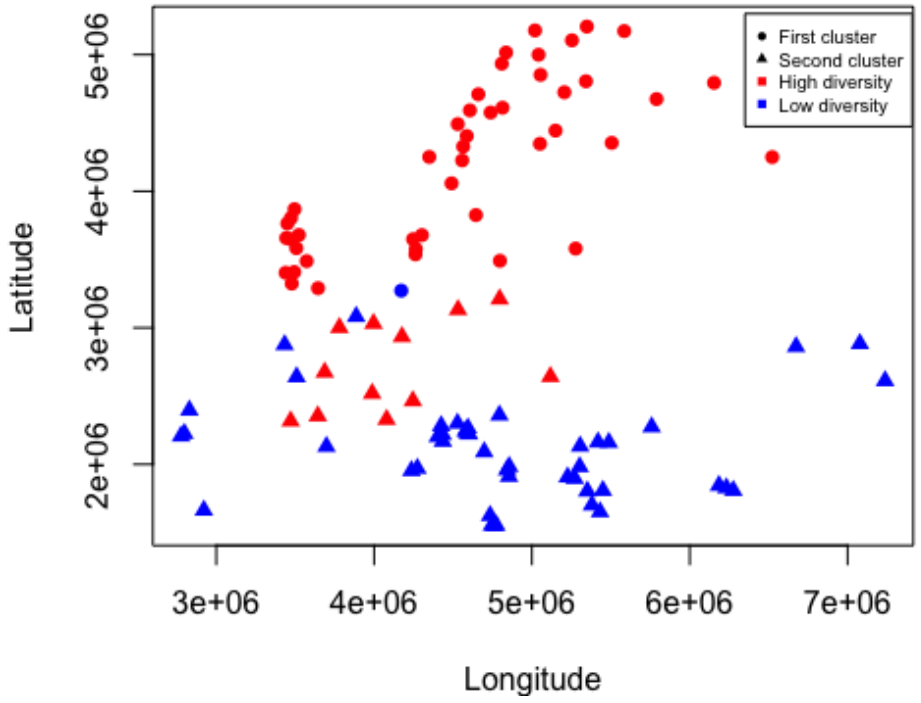


Figure 7 : Clustering k-means avec longitude, latitude et taux de diversité

L'objectif ici est d'établir une corrélation entre la diversité des sites et les deux clusters. Le premier cluster semble regrouper les sites à forte diversité tandis que le second cluster apparaît regrouper les sites à faible diversité. On peut remarquer que cette association est particulièrement importante pour le premier cluster et les sites de forte diversité.

● Apprentissage supervisée : arbres de décisions

Les arbres de décisions sont des outils d'aide à la décision, visant à produire une procédure de classification interprétable. Ils ont la capacité de sélectionner automatiquement les variables discriminantes. Concrètement, il s'agit d'une décomposition du problème de classification en une suite de tests correspondant à une partition de l'espace des données en sous-régions homogènes en termes de classe.

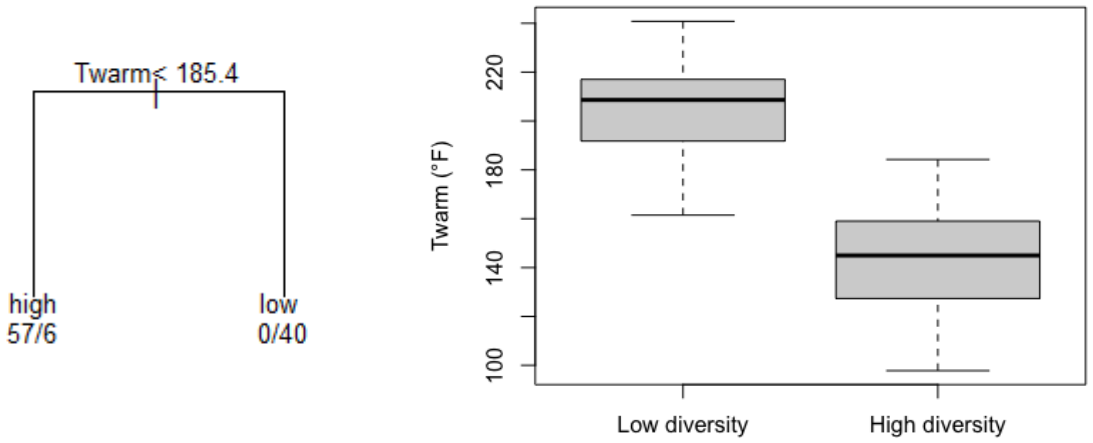


Figure 8 : (à gauche) Arbre de décision basé sur Twarm ; (à droite) Boxplot de Twarm selon la diversité

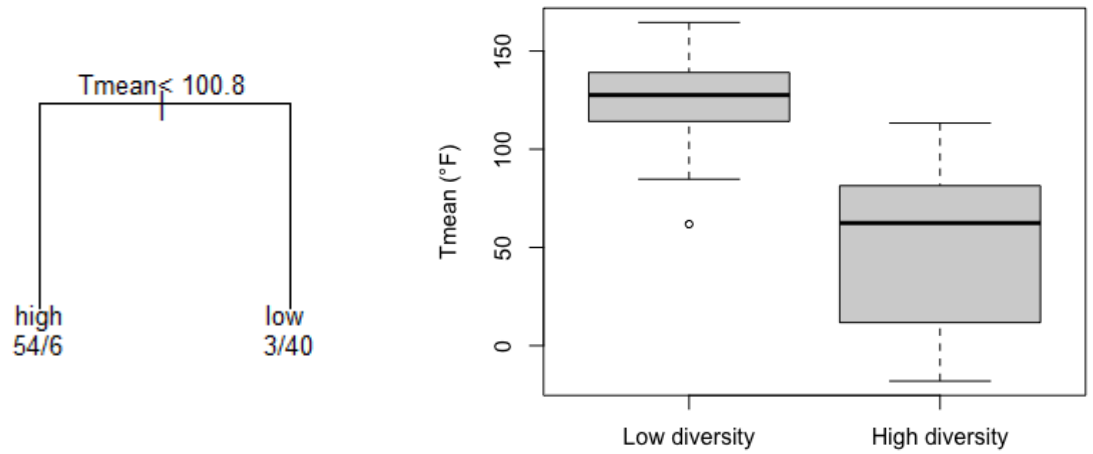


Figure 9 : (à gauche) Arbre de décision basé sur Tmean ; (à droite) Boxplot de Tmean selon la diversité

Twarm semble être la variable la plus adaptée pour discriminer les sites avec un nombre élevé ou faible d'espèces. On en déduit que les sites possédant une température des trois mois les plus chauds élevée semblent avoir une plus faible diversité. Le boxplot et l'arbre de décision ci-dessus illustrent ces propos.

Pour la température annuelle moyenne, la température de 100.8 °F est cruciale . En effet, il s'agit du seuil à partir duquel les sites seront classés avec 3 erreurs pour low (faible diversité). Cependant, 6 erreurs sont commises pour high (forte densité), ce qui signifie que 6 sites low seront classés comme high.

Il est nécessaire de compléter cette analyse par une comparaison des moyennes à l'aide d'un Test de Student afin de vérifier si la différence de température des trois mois les plus chauds entre Low diversity et High diversity est significative.
On vérifie que les 2 distributions sont normales (H0 : normalité) :

```
Shapiro-Wilk normality test

data:  data_climat_low$Twarm
W = 0.96248, p-value = 0.1432
```

p-value > 0.05, H0 n'est donc pas rejetée pour le groupe data_climat_low.

```
Shapiro-Wilk normality test

data:  data_climat_high$Twarm
W = 0.97641, p-value = 0.3278
```

p-value > 0.05, H0 n'est donc pas rejetée pour le groupe data_climat_high.

On vérifie que les variances sont homogènes (H0 : homoscédasticité) :

```
data:  data_climat_low$Twarm and data_climat_high$Twarm
F = 0.72028, num df = 45, denom df = 56, p-value = 0.257
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.4144521 1.2736844

sample estimates:
ratio of variances
    0.7202846
```

p-value > 0.05, H0 n'est donc pas rejetée.

On compare les moyennes (H0 : moyennes égales):

```
data:  data_climat_low$Twarm and data_climat_high$Twarm
t = 15.088, df = 101, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
```

```
95 percent confidence interval:
55.75732    Inf

sample estimates:
mean of x         mean of y
204.3853          141.7346
```

p-value < 0.05, H0 est rejetée. H1 étant que la température est plus élevée dans les sites à faible diversité par rapport aux sites à forte diversité quand il s'agit des trois mois les plus chauds.

D'après ces tests, les différences de températures des trois mois les plus chauds entre des sites à faible et forte diversité est significative : température significativement plus élevée dans les sites à faible diversité. Cela signifie qu'il s'agit probablement d'un facteur influant la survie et le développement des Salmonidés.

Concernant Tmean une autre analyse par une comparaison des moyennes a été effectuée avec un test de Wilcoxon car le groupe de forte diversité n'a pas une distribution normale.

```
Wilcoxon rank sum test with continuity correction

data:  data_climat_low$Tmean and data_climat_high$Tmean
W = 2539, p-value < 2.2e-16
alternative hypothesis: true location shift is greater than 0
```

p-value < 0.05, H0 est rejetée. H1 étant que la température est plus élevée dans les sites à faible diversité par rapport aux sites à forte diversité en prenant en compte la température annuelle moyenne.

Ainsi les différences de température annuelle moyenne entre des sites à faible et forte diversité est significative : température significativement plus élevée dans les sites à faible diversité.

• Conclusion

L'objectif de ce projet est une étude approfondie de données multivariées sur les Salmonidés. Différentes questions écologiques abordées ont été éclaircies par diverses méthodes d'analyses. De plus, des tests statistiques sont venus compléter ces analyses afin d'affirmer ou d'infirmer des différences entre les groupes d'intérêt.

Des variables environnementales peuvent-elles expliquer la répartition particulière d'espèces ?
Existe-t-il des différences de diversité des espèces entre les sites ?

Premièrement, nous avons tenté d'établir d'éventuels liens entre des variables environnementales et la répartition des espèces. L'étude de l'impact de variables géographiques et climatiques avec une ACC a permis de mettre en évidence un facteur semblant jouer un rôle majeur dans la répartition des espèces : il s'agit de la température. De surcroît, l'analyse de co-inertie sur les variables d'occupation du sol a révélé une corrélation négative vis-à-vis de densité de population (habitants/km2) et du pourcentage de surfaces artificielles pour une majorité d'espèces de Salmonidés. Des facteurs tels que la pollution de l'eau pourraient avoir un impact.

Deuxièmement, il a été question d'approfondir la question des différences de diversité des espèces entre les sites. Deux critères semblent avoir une influence sur cette diversité : la température annuelle moyenne et la température des trois mois les plus chauds. Ainsi la température est un facteur primordial pour les Salmonidés car elle affecte de manière différente les espèces. Elle pourrait jouer un rôle dans leur survie et leur développement.

Pour aller plus loin, il serait pertinent de ne pas se limiter à l'étude des Salmonidés mais à l'ensemble des poissons afin d'étudier les différences de diversité des espèces entre les 274 sites d'intérêt.