IAE Nantes
Économie & Management
Pôle Sociétés

Master 1 - Économétrie et Statistique, parcours Économétrie Appliquée

Modèles sur variables latentes

Apprentissage supervisé par une approche PLS pour la discrimination des saumons suivant leur provenance et leur mode de production sur la base de données de caractérisation chimique

Gloria Isabel PALACIO BARCO

IAE - Nantes

Année universitaire 2023-2024

Description et préparation des données

Les données sont issues de l'article de Hong et al. (2023) dans Nature Communications, qui analyse l'authenticité du saumon en examinant 521 échantillons de quatre régions (Alaska, Écosse, Norvège, Islande) via la technique ICP-MS[1], pour identifier des éléments chimiques clés. Une normalisation des données est effectuée avec la méthode min-max, rendant les comparaisons entre échantillons uniformes.

La préparation des données implique le chargement du fichier CSV, le filtrage des colonnes pour se concentrer sur les éléments chimiques et l'origine des échantillons, l'ajustement des noms des colonnes pour la clarté, et la conversion de la colonne pays en catégorie. Pour finir, les données sont normalisées avec « predict() et preProcess() » du package 'caret', préparant l'analyse statistique.

Analyse descriptive

Pour débuter, j'ai utilisé la fonction « summary » afin d'examiner les statistiques descriptives essentielles telles que la moyenne, le minimum, et le maximum, et pour détecter d'éventuelles anomalies, telles que des valeurs négatives ou d'autres incohérences. Cette analyse révèle que toutes les valeurs sont positives, ce qui confirme l'absence d'erreurs flagrantes dans les données. De plus, le décompte des observations par pays est conforme aux informations préalablement fournies, ce qui indique une bonne intégrité des données. Parallèlement, j'ai procédé à une vérification pour s'assurer de l'absence de valeurs manquantes, confirmant ainsi que le dataset est complet et prêt pour les analyses suivantes.

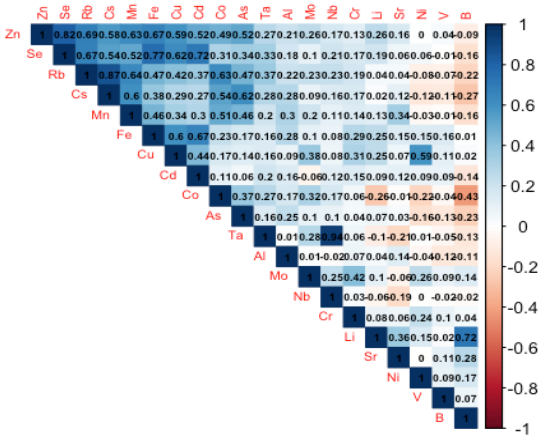
L'analyse a également permis de mettre en évidence la présence potentielle de valeurs extrêmes, en particulier en examinant les valeurs minimales et maximales qui se distinguent nettement du reste des données. Ces écarts significatifs peuvent refléter une variabilité naturelle au sein de l'échantillon analysé. Par exemple, dans le cas de l'aluminium (Al), nous observons que la valeur maximale est de 45,092.6, tandis que la médiane reste bien plus modeste, à 1,447.5. Cette constatation suggère que, bien que la majorité des valeurs se regroupent autour d'un point central relativement bas, quelques échantillons présentent des concentrations d'aluminium exceptionnellement élevées. Cela pourrait indiquer des différences spécifiques liées à l'origine géographique, aux conditions environnementales, ou à d'autres facteurs influençant la composition des échantillons.
La matrice de corrélation met en évidence des corrélations élevées entre certains éléments, notamment entre Zn et Se, ainsi qu'entre Ta et Nb, avec des niveaux de corrélation approchant 1. Ces fortes corrélations suggèrent une multicolinéarité significative entre ces variables, ce qui peut compliquer l'analyse statistique et la précision des prédictions .
Pour cette analyse, nous allons utiliser les méthodes Partial Least Squares (PLS), et plus spécifiquement la PLS Discriminant Analysis (PLS-DA), en raison de leur capacité à gérer la multicolinéarité. Ces méthodes traitent efficacement les variables corrélées en les regroupant en un nombre réduit de composantes principales. Cette démarche simplifie le modèle tout en préservant les informations essentielles nécessaires à la prédiction. Cette

[1] ICP-MS : Technique de spectrométrie de masse à plasma couplé inductif. Identification des éléments : Li, B, Al, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, As, Se, Rb, Sr, Nb, Mo, Cd, et Cs

approche s'avère particulièrement pertinente pour notre étude, notamment parce que l'analyse discriminante linéaire (LDA) impose une distribution normale des variables indépendantes, une condition qui n'est pas forcément respectée dans notre cas, comme le montre le script.

Figure 1 Tableau de contingence pays et corrélation

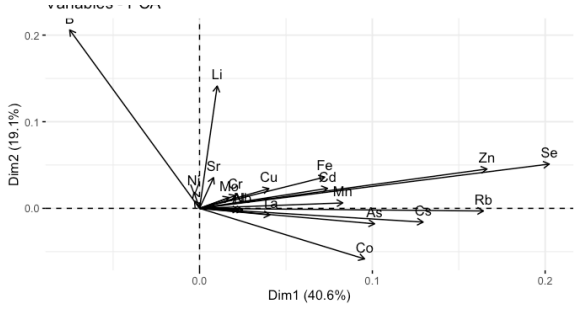
Pays	Nombre	Mode de production
Alaskan	99	sauvage (wild caught)
Iceland-F	55	élevage (farmed)
Iceland-W	90	sauvage (wild caught)
Norway	100	élevage (farmed)
Scotland	177	élevage (farmed)



Analyse préliminaire ACP

L'ACP nous permet d'explorer et de visualiser la structure générale des données en réduisant la dimensionnalité, ce qui aide à identifier les regroupements naturels et révèle les principales sources de variation :

Figure 2 Projection ACP dim 1 et 2



Nous observons que la première dimension de l'ACP, qui représente 40,6 % de la variance totale, est principalement influencée par le Rb, le Zn et le Se, ce qui confirme leur forte corrélation positive identifiée auparavant. La seconde dimension met en évidence l'importance du Li et du B, confirmée par des cosinus carrés supérieurs à 0,7. Quant à la troisième dimension, elle se distingue par une contribution significative de certains éléments comme le Cd, avec une valeur de 0,36, indiquant que cette dimension capte une part importante de la variance associée à cet élément. Ces axes principaux révèlent des liens potentiels entre les éléments mesurés, suggérant l'influence de facteurs communs sur leur présence dans les échantillons de saumon, qui pourraient indiquer des conditions environnementales spécifiques ou des caractéristiques biologiques partagées par les échantillons analysés.

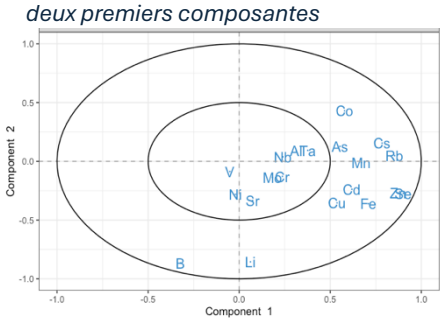
Création des jeux de données d'apprentissage et de test

J'ai réparti les données en ensembles d'apprentissage et de test en utilisant la fonction createDataPartition(...) du package caret, attribuant 80 % des données à l'ensemble d'apprentissage et les 20 % restants à l'ensemble de test. Pour assurer la reproductibilité des résultats à l'avenir, j'ai sauvegardé les indices de l'ensemble d'entraînement dans un fichier Rdata. Cela permet de garantir que les mêmes données seront utilisées pour l'entraînement dans toute réplication de l'étude. Les ensembles d'apprentissage et de test, que j'ai nommés X.app et X.test, comprennent respectivement 418 et 108 observations réparties sur le s 20 variables concernées.

Analyse Discriminante par les Moindres Carrés Partiels (PLS-DA)

Pour cette analyse, j'emploie le package mixOmics pour appliquer la PLS-DA. Je commence par l'appliquer sur l'échantillon d'apprentissage, nommé X.app, en utilisant initialement un modèle PLS-DA comportant dix composants. Cela me permet d'évaluer à la fois les performances du modèle et de déterminer le nombre optimal de composants pour le modèle final. Ensuite, je projette les échantillons dans le sous-espace formé par les deux premiers composants pour une analyse plus poussée.

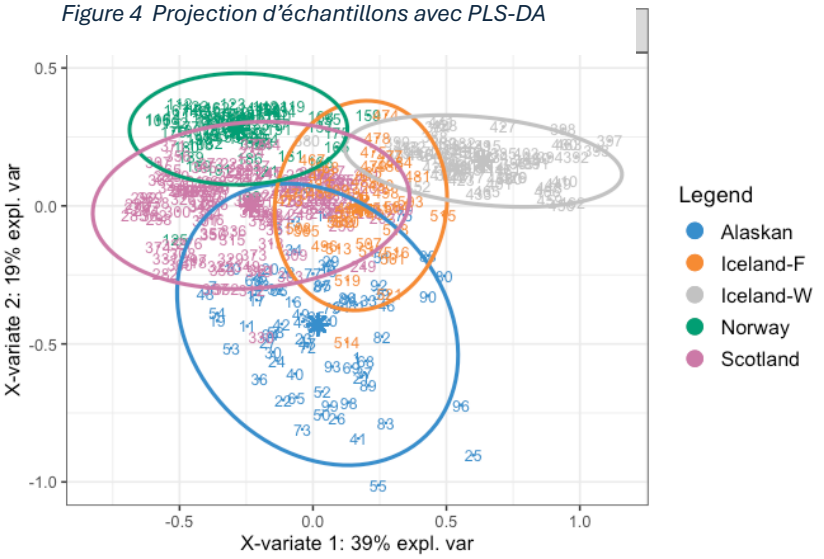
Figure 3 Cercle de corrélation associé aux deux premiers composantes



D'après la figure 3, les variables placées vers les extrémités droite et gauche du cercle (par exemple, "Co", "Cs", "Rb" à droite et "Li", "B" à gauche) sont fortement corrélées avec la Composante 1.

Plus une variable est proche du cercle, meilleure est sa représentation. Par exemple, "Co", "As", "Cs", "Rb" sont bien représentés par ces deux composantes.

Figure 4 Projection d'échantillons avec PLS-DA



La première composante latente explique 39% de la variance, tandis que l'axe vertical (la deuxième composante) 19% de la variance.
Nous pouvons observer une séparation des 5 origines différentes des saumons. Des ellipses de confiance pour chaque classe sont tracées pour mettre en évidence la force de la discrimination (niveau de confiance fixé à 95% par défaut).

Bien que les groupes aient des régions distinctes, l'existence de chevauchements (entre les échantillons d'Iceland-F, Iceland-W, et Norway par exemple), indique qu'il pourrait y avoir de la place pour améliorer la capacité discriminante du modèle, nous allons donc affiner le modèle PLS-DA existant.

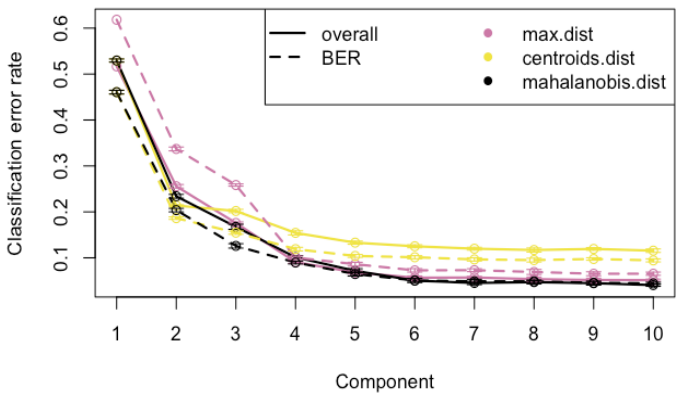
Optimisation des Hyperparamètres

Pour optimiser les hyperparamètres et évaluer les performances du modèle PLS-DA, nous allons utiliser la fonction “perf”, avec une méthode de validation croisée “Mfold” (au lieu de la méthode Leave-One-Out qui est un cas particulier de validation croisée avec k=n).
J'ai choisi une validation croisée à 10 plis, répétée 10 fois. Cette approche assure une estimation de la performance du modèle à la fois stable et fiable. La validation croisée est effectuée de manière aléatoire à chaque répétition, ce qui contribue à la robustesse de notre estimation du taux d'erreur de classification.

Nous allons également fixer le générateur de nombres aléatoires en utilisant set.seed(2024), ce qui garantit la reproductibilité des résultats. En divisant les données en 10 plis distincts, le modèle est entraîné sur 9 plis et testé sur le 10ème, processus qui est ensuite répété pour chacun des plis.
L'argument “progressBar = FALSE” est utilisé pour supprimer l'affichage de la barre de progression pendant l'exécution de la validation croisée.
De plus, avec “auc = TRUE”, nous calculons l'AUC, fournissant une mesure complémentaire de la performance du modèle. L'AUC permet de résumer la capacité du modèle à classer correctement les instances sur toute la gamme des seuils de classification.

Finalement, pour ce qui est du choix de la mesure de distance utilisée pour évaluer la performance, nous retenons l'option par défaut ("all"), ce qui nous permettra d'examiner et de comparer les résultats obtenus avec différentes mesures de distance.

Figure 5 Taux d'erreur de classification PLS-DA



Ligne continue (overall) : Représente l'erreur de classification globale moyenne sur tous les plis de la validation croisée pour chaque nombre de composants.

Ligne en pointillés (BER) : Montre le Balanced Error Rate, qui est l'erreur de classification équilibrée entre les classes. Le BER est utile lorsque les classes ne sont pas équitablement représentées dans les données.

Étant donné que le nombre d'observations dans notre jeu des données par pays sont déséquilibrées (la classe "Scotland" a nettement plus d'observations que les autres classes, notamment "Iceland-F" qui en a le moins), l'erreur overall pourrait masquer les performances du modèle sur les classes moins représentées. (BER) est donc une métrique plus appropriée, car il traite chaque classe également, indépendamment du nombre d'observations.
A partir du graphique des performances, nous observons que le taux d'erreur global et le taux d'erreur équilibré (BER) sont assez similaires et diminuent significativement lorsque le nombre de composants augmente de 1 à 2, puis continue de diminuer plus graduellement jusqu'à environ 5 composants, après quoi la diminution devient plus lente ou stagne. Cette tendance indique qu'un nombre plus important de composants peut améliorer la capacité du modèle à classer correctement les observations, jusqu'à un certain point où des composants supplémentaires n'apportent plus d'amélioration significative.
Plusieurs mesures de distance sont illustrées : la distance maximale (max.dist), la distance entre centroïdes (centroids.dist) et la distance de Mahalanobis (mahalanobis.dist).
La fonction “perf” génère le nombre optimal de composants pour nous conseiller les suivants :

	max.dist	**centroids.dist**	**mahalanobis.dist**
overall	8	7	8
BER	8	7	8

Selon les critères d'erreur et les différentes mesures de distance le nombre de composantes devrait être entre 7 et 8, mais l'idéal serait de choisir un nombre de composants qui minimise l'erreur de classification tout en évitant la complexité inutile, un choix de modèle parcimonieux (plus simple) est préférable. Dans notre graphique, ce point pourrait être autour de 4 ou 5 composants. En examinant les taux d'erreur, nous obtenons :

BER

	max.dist	**centroids.dist**	**mahalanobis.dist**
comp1	0.61827074	0.46062331	0.46062331
comp2	0.33700665	0.18642058	0.20387214
comp3	0.25841958	0.15474353	0.12595615
comp4	0.10049410	0.11892794	0.08976536
comp5	0.08613562	0.10403116	0.06469313

Pour toutes les méthodes de distance, le BER diminue à mesure que le nombre de composants augmente, en particulier après le passage de 1 à 4 composants.

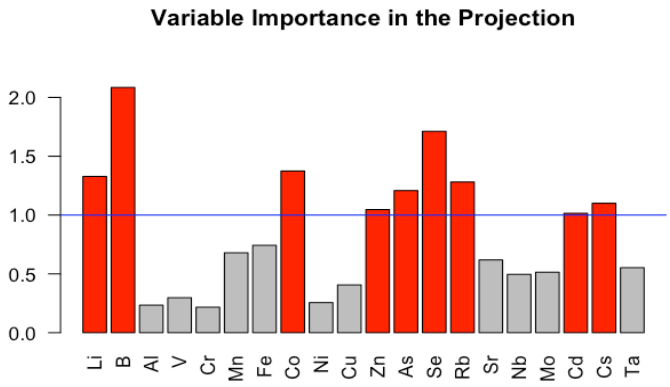
Concernant la distance nous retiendrons Mahalanobis qui présente une meilleure performance par rapport aux autres mesures de distance, elle offre l'avantage supplémentaire de prendre en considération la forte corrélation entre certaines de nos variables. En s'appuyant sur la matrice de covariance pour ajuster les distances, cette méthode capture de manière plus précise la variabilité et les interrelations des données.

Comme mentionné précédemment, nous cherchons un équilibre entre la performance de classification et la simplicité du modèle, en évitant potentiellement le surajustement qui pourrait survenir avec un grand nombre de composantes ; par conséquent, nous opterons pour 4 composantes.

Importance des variables (VIP)

Les scores VIP quantifient l'importance de chaque variable dans notre modèle en termes de contribution à la variance expliquée et à la capacité de discrimination du modèle. Un score plus élevé indique une variable plus importante (supérieur à 1 est souvent utilisé comme un seuil pour identifier les variables les plus influentes).

Figure 6 VIP



Nous pouvons observer que les prédicteurs les plus pertinents sont le lithium (Li), le bore (B), le cobalt (Co), le zinc (Zn), l'arsenic (As), le sélénium (Se), le rubidium (Rb), le cadmium (Cd), et le césium (Cs). Ces éléments chimiques sont pertinents pour déterminer l'origine du saumon, car leurs concentrations peuvent varier significativement avec les caractéristiques environnementales propres à chaque habitat aquatique. Les scores VIP sont précieux car ils synthétisent l'importance de ces variables à travers l'ensemble des composants du modèle, reflétant leur impact total sur la performance du modèle.

Figure 7 AUC

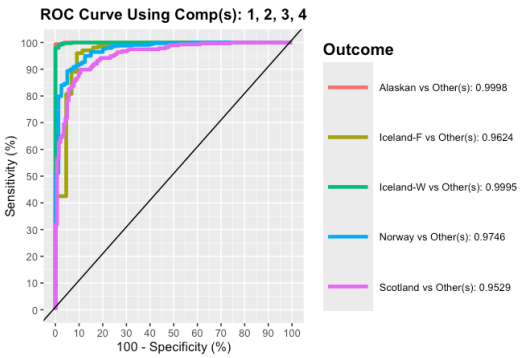


Figure 7 : Nous avons des AUC élevées pour tous les groupes suggèrent que le modèle est capable de classer avec succès les échantillons en fonction de leur origine géographique.

"Norway vs Other(s)" et "Scotland vs Other(s)" ont des AUC légèrement plus faibles que les autres groupes, mais ces valeurs (0.9746 et 0.9529 respectivement) sont toujours considérées comme indiquant une très bonne performance de classification.

Modèle PLS-DA final et conclusion

Pour le modèle final, nous conservons 4 composantes et pour la prédiction nous utiliserons la distance de mahalanobis, nous appliquons sur nos données de test (X.test) et les résultats dans la matrice de confusion sont :

Figure 8 Matrice de confusion

	Alaskan	Iceland-F	Iceland-W	Norway	Scotland
Alaskan	19	0	0	0	0
Iceland-F	0	10	0	1	0
Iceland-W	0	0	18	0	0
Norway	1	3	0	15	1
Scotland	0	1	0	2	32

Le modèle affiche une performance notable en matière de prédiction, parvenant à classer correctement tous les échantillons pour l'Alaska et l'Islande W, tandis que pour l'IslandeF, 10 échantillons sur 11 ont été correctement identifiés. Concernant la Norvège, le modèle a correctement reconnu 15 échantillons sur 20, avec quelques confusions réparties entre les classes Alaska, Islande ferme et Écosse. Enfin, pour l'Écosse, 32 échantillons sur 35 ont été exactement classés, avec de minimes erreurs incluant un échantillon attribué par erreur à l'Islande ferme et deux à la Norvège.

L'exactitude du modèle est obtenue en sommant les prédictions correctes, c'est-à-dire les valeurs sur la diagonale de la matrice de confusion, puis en divisant cette somme par le total des échantillons testés. Avec un résultat de 0.9126214, cela indique que le modèle PLS-DA a une précision de 91.26% sur l'ensemble de test, ce qui témoigne de sa bonne performance. Par conséquent, le taux d'erreur s'établit à 8.74%, reflétant la proportion d'échantillons qui ont été incorrectement classifiés.

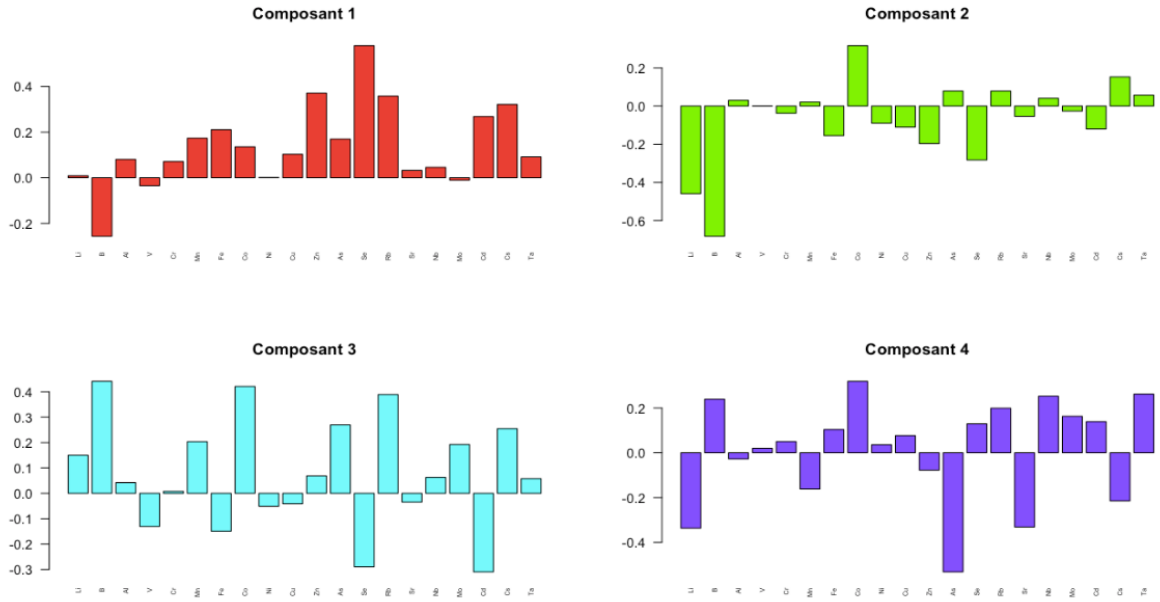
Autres métriques de performance : Pour évaluer l'exactitude et la couverture des prédictions du modèle nous calculons la précision et le rappel; la première reflète la proportion de prédictions correctes pour une classe spécifique, et le second mesure la proportion de cas réels correctement identifiés pour cette même classe.

	Alaskan	Iceland-F	Iceland-W	Norway	Scotland
precision	0.9500000	0.7142857	1.0000000	0.8333333	0.9696970
recall	1.0000000	0.9090909	1.0000000	0.7500000	0.9142857
F1	0.9743590	0.8000000	1.0000000	0.7894737	0.9411765

Pour Alaskan, la précision et le rappel atteignent respectivement 95% et 100%, reflétant une classification sans erreurs des saumons Alaskan, mais sa précision globale pour

Alaskan est légèrement diminuée. Pour Iceland-F, malgré un rappel élevé à 90.91%, la précision baisse à 71.43% à cause d'erreurs mineures affectant la classification. Les saumons Iceland-W se distinguent par une précision et un rappel parfaits de 100%. La classe Norway montre une précision de 83.33% face à un rappel plus modeste de 75%, due à des confusions avec d'autres classes. Scotland affiche une forte précision de 96.97% et un rappel de 91.43%, avec quelques erreurs marginales.

Les scores F1 élevés pour Alaskan et Iceland-W soulignent la bonne performance du modèle pour ces catégories. Scotland, avec un F1 de 0.941, révèle aussi une performance robuste, quoique légèrement inférieure. Iceland-F et Norway, avec des F1 de 0.800 et 0.789, indiquent une bonne performance générale, mais suggèrent un potentiel d'amélioration.
Le modèle PLS-DA affiche une bonne performance dans la classification des origines du saumon, bien que certaines catégories comme Iceland-F et Norway présentent des performances modérées en raison de chevauchements de caractéristiques ou de variabilités internes qui compliquent la distinction.



Au sein de notre modèle, le sélénium (Se) se démarque comme le contributeur principal du premier composant, soulignant son rôle prépondérant dans la distinction des origines du saumon. Le deuxième composant est marqué par une influence négative significative du bore (B), tandis que le cobalt (Co) et le rubidium (Rb) ressortent fortement dans le troisième composant. L'arsenic (As) caractérise le quatrième composant par une forte contribution négative. Ces éléments, en particulier le sélénium, le cobalt et l'arsenic, qui manifestent des contributions importantes à travers divers composants, jouent un rôle clé dans la détermination précise de l'origine géographique du saumon, reflétant les variations environnementales à chaque habitat.